

LOW-SHOT LEARNING OF SUBSTRATE SPECIFICITY ON TRANSMEMBRANE TRANSPORT PROTEINS

SIMA ATAEI

THE DEPARTMENT
OF
COMPUTER SCIENCE AND SOFTWARE ENGINEERING

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
CONCORDIA UNIVERSITY
MONTRÉAL, QUÉBEC, CANADA

MARCH 2025

© SIMA ATAEI, 2025

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the document prepared

By: **Ms. Sima Ataei**

Entitled: **Low-shot learning of substrate specificity on transmembrane transport proteins**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (DOCTOR OF PHILOSOPHY)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Claudio Contardo	_____	Chair
Dr. Igor Jurisica	_____	External Examiner
Dr. Simone Brugiapaglia	_____	Examiner
Dr. Sabine Bergler	_____	Examiner
Dr. Tristan Glatard	_____	Examiner
Dr. Gregory Butler	_____	Supervisor

Approved	_____	Dr. Sabine Bergler,
		Graduate Program Director

November 22, 2024	_____	Dr. Mourad Debbabi, Dean
		Faculty of Engineering and Computer Science

Abstract

Low-shot learning of substrate specificity on transmembrane transport proteins

Sima Ataei,

Concordia University, 2025

Transmembrane transport proteins are essential for cellular processes, selectively moving substrates across membranes. Traditional wet lab methods for detecting substrate specificity, like binding and uptake assays, are costly and impractical for large-scale studies. Computational approaches, particularly machine learning (ML), could offer efficient alternatives. State-of-the-art (SOTA) models to date can predict general groups of substrates carried by transporters, such as organic ions. The SOTA model TooT-SC achieves a Matthew’s Correlation Coefficient (MCC) of 0.82 predicting 11 general substrate classes. This research presents novel computational methods for predicting transported substrates using few-shot, one-shot, and zero-shot learning techniques to handle imbalanced datasets, leveraging large language models (LLMs). These low-shot learning models enhance substrate specificity prediction.

An automatic pipeline is introduced to create machine learning (ML)-ready datasets for specific substrate groups, integrating the Chemical Entities of Biological Interest (ChEBI) and Gene Ontology (GO) databases to address the lack of annotated protein sequence data. Initial studies confirm the effectiveness of transformer-based Protein Language Models (PLMs), adapted from natural language processing (NLP), in this context.

The research focuses on three key projects: TooT-Open-ICAT (Open-world classification of Inorganic Cations and Anions Transporters) predicts inorganic ion transport using open-world classification; TooT-Triplet-SPEC (Triplet training for substrate SPECificity prediction) predicts specific substrates through metric learning; TooTranslator shifts from classification to regression to predict substrates of uncharacterized proteins.

The TooTranslator model advances the SOTA by improving predictions for fine-grained classes accurately predicting 93 specific substrates with an MCC of 0.92. Furthermore, the models show promise in predicting true labels for unseen classes.

To my grandma,

Whose right to education was taken by patriarchy, a wound she still feels at 85.

And to every woman, still fighting for the dignity of this basic human right.

Acknowledgments

I express my gratitude to my supervisor, Dr. Gregory Butler, for their guidance throughout this process. I am also grateful to the members of my committee for their insightful feedback and valuable discussions.

A special thanks to my friends and colleagues, particularly Hamed Ghazikhani, whose support and encouragement made this experience memorable. To my family, thank you for your love, patience, and belief in me—this achievement would not have been possible without you.

Table of Contents

List of Figures	xii
List of Tables	xv
1 Introduction	1
1.1 Context	1
1.2 Motivation	5
1.2.1 Transmembrane transport proteins	5
1.2.2 Computational methods on substrate specificity prediction	7
1.3 Preliminary studies	9
1.4 Problem Definition	9
1.5 Research questions	10
1.6 Challenges	11
1.7 Contributions	11
2 Background and related works	13
2.1 Introduction	13
2.2 Large Language Models: BERT	14
2.3 Protein Language Models: ProtBERT-BFD	16
2.4 Open-world classification	19
2.5 Imbalanced Dataset	21
2.5.1 Long-tailed imbalance	22
2.5.2 Learning from imbalanced data	22
2.6 Low-shot Learning	24
2.7 Metric-based methods	26
2.8 Few-shot classification using siamese networks	26
2.8.1 FaceNet	27

2.8.2	Sentence-BERT	28
2.9	Low-shot classification on protein sequences	29
2.9.1	Deep learning proteins using a Triplet-BERT network	29
2.9.2	ProTranslator	30
2.10	Substrate class prediction	32
2.11	Evaluation Metrics	33
2.12	Loss Functions	35
2.12.1	Focal loss	35
2.12.2	Triplet loss	36
2.12.3	Mean Squared Error and weighted Mean Squared Error	36
2.12.4	Mean Absolute Error	37
2.12.5	Huber Loss	37
2.13	Summary	37
3	Preliminary Study	39
3.1	Introduction	39
3.2	Datasets for Substrate Specificity	39
3.2.1	Chemical Entities of Biological Interest (ChEBI)	40
3.2.2	Gene Ontology (GO)	40
3.3	GO and ChEBI cross-reference	41
3.4	SGI Dataset Construction Pipeline	42
3.4.1	Algorithms	45
3.4.2	C2GO_C map from CHEBI leaves	47
3.4.3	C2GO_G map from GO leaves	49
3.4.4	Build C2GO map	50
3.4.5	Map Refinement	50
3.4.6	Labeling Data	54
3.4.7	Identity Reduction	55
3.5	Example Datasets	55
3.5.1	TooT-ICAT Dataset series	55
3.5.2	TooT-SPEC Dataset series	59
3.6	TooT-ICAT	62
3.6.1	Dataset description	62
3.6.2	Methodology	63
3.6.3	BERT Representation	63

3.6.4	Logistic Regression	63
3.6.5	Fine-tuned BERT language model	63
3.6.6	Cross-validation	64
3.6.7	Results and discussion	64
3.7	TooT-SPEC	71
3.7.1	Dataset description	71
3.7.2	Dataset Analysis	72
3.7.3	Methodology	73
3.7.4	Fine-tuning ProtBERT-BFD model	74
3.7.5	Cross-validation	74
3.7.6	Results and discussion	74
4	TooT-Open-ICAT: Open classification of inorganic ion transport proteins	78
4.1	Introduction	78
4.2	Problem statement: Inorganic ion prediction	79
4.3	Challenges	80
4.4	Methodology	81
4.4.1	Introduction	81
4.4.2	Network Structure	81
4.4.3	Dataset	83
4.4.4	Training	85
4.4.5	Thresholded classification layer	86
4.5	Baseline: Adding miscellaneous class	88
4.6	Results	89
4.7	Discussion	90
4.7.1	Micellaneous class effect	90
4.7.2	Imbalanced classes in open-classification	93
4.7.3	Novel class detection in open-classification	95
4.8	Conclusion	95
5	TooT-Triplet-SPEC: One-shot classification of transmembrane transport proteins	97
5.1	Introduction	97
5.2	Problem statement: specific substrate prediction	98

5.3	Methodology	98
5.3.1	Offline sampling strategy	100
5.3.2	Online sampling strategy	100
5.3.3	Training	101
5.3.4	Baseline	103
5.4	Dataset	103
5.5	Classification	104
5.6	Challenges	104
5.6.1	Memory limitaion	105
5.6.2	Computational time limitation	106
5.7	Results	107
5.7.1	Offline Sampling Results	107
5.7.2	Online Sampling Results	112
5.8	Discussion	115
5.8.1	Sampling strategy	115
5.8.2	Interpreting sample distance	118
5.8.3	Majority class effect on training	122
5.8.4	Binding sites and attention heads	126
5.9	External validation	128
5.9.1	Evaluation using ChEBI terms	128
5.9.2	Evaluation using GPT4	129
5.10	Conclusion	146
6	TooTranslator: Zero-shot classification of <i>de novo</i> transport proteins	147
6.1	Introduction	147
6.2	Problem statement: <i>de novo</i> substrate prediction	149
6.3	Methodology	149
6.3.1	Phase 1: Vectorization and projection	150
6.3.2	Network structure	153
6.3.3	Training	154
6.3.4	Phase 2: Classification	155
6.3.5	TooTranslator: Validity	156
6.4	Results	157
6.4.1	Results for Group A: seen classes	158
6.4.2	Results for Group B: unseen classes	160

6.5	Discussion	163
6.5.1	Training process positively affects unseen classes distances	163
6.5.2	Training dataset composition is critical	166
6.5.3	Regression versus Classification	167
6.5.4	Unseen class prediction analysis	168
6.6	Conclusion	175
7	Conclusion	176
7.1	Introduction	176
7.2	Contributions	177
7.3	Developed Tools	180
7.4	Addressing Research Questions	181
7.5	Research publications	182
7.6	Future work	183
A	C2GO Mapping	203
A.1	Introduction	203
B	Dataset Information	212
B.1	SPEC dataset series	212
C	TooT-SPEC	215
C.1	Introduction	215
C.2	TooT-SPEC results on hold-out testset	215
C.3	TooT-SPEC results for 3-Fold Cross-Validation	218
D	TooT-Open-ICAT	223
E	TooT-Triplet	225
E.1	Introduction	225
E.2	TooT-Triplet detailed results	225
E.2.1	Hold-out test set results	226
E.2.2	Folds results	231
E.2.3	TooT-Triplet Proton Excluded Results	236
E.3	Other method results	245
E.3.1	Baseline Results	245

E.3.2	Fine-tuned TooT-SPEC + KNN results	250
E.4	TooT-Triplet external validation	256
E.4.1	TooT-SC dataset	256
E.4.2	Organism analysis	265
E.5	TooT-Triplet distances	267
F	TooTranslator	271
F.1	Results	271
F.2	Inorganic ion transporter prototypes	296

List of Figures

1	Transport process through the using different types of transmembrane transport proteins [DTM19].	3
2	An example of the selectivity filter in transmembrane transport proteins. .	4
3	ProtBERT model architecture [EHD ⁺ 22].	18
4	Zero-data learning of new tasks [LEB08].	25
5	The learning effect using Triplet Loss function.	27
6	Example of ChEBI ontology hierarchy for inorganic ion.	41
7	Gene Ontology hierarchy for membrane transport in 3 domains.	42
8	An example of GO and ChEBI cross-reference from GO-Plus.owl file. . . .	43
9	The ChEBI hierarchy for CHEBI:15377-water	46
10	The ChEBI term for metal cation with its ancestors, and descendants. . . .	52
11	Detailed process for extracting TooT-ICAT dataset series.	56
12	The ancestor chart for GO term GO:0004129 from Gene Ontology.	58
13	Comparison of validation performance in the training of TooT-ICAT. . . .	65
14	UniProt-SPEC-100 dataset analysis.	72
15	Log-scale of sample distribution on each class for UniProt-SPEC	73
16	An example of class-specific threshold calculation in one fold	84
17	The testing process of TooT-Open-ICAT.	86
18	Thresholded network structure in TooT-Open-ICAT.	87
19	The test sample groups included in TooT-Open-ICAT analysis.	91
20	Confusion matrix for the hold-out test set results.	92
21	Heatmap of prediction values for penultimate layer for each test sample. .	94
22	Training process in TooT-Triplet with offline sampling strategy.	100
23	Training process in TooT-Triplet with online sampling strategy.	101
24	TooT-Triplet results for the hold-out test set trained on three distance metrics.	108

25	Evaluation in training TooT-Triplet using Manhattan distance.	109
26	Silhouette score comparison of TooT-Triplet (offline, Manhattan) with Base- line embedding space (frozen) for each class in UniProt-SPEC-100.	136
27	Overall silhouette score for three model groups.	137
28	(Offline) Distance distributions for negative and positive nearest neighbor.	137
29	(Offline) Distance distributions for positive nearest neighbor.	138
30	(Offline) Distance distributions for negative nearest neighbor.	138
31	Class separated distributions of distances for the negative and positive nearest neighbor in majority classes encodings extracted from TooT-Triplet (offline).	139
32	Class separated distributions of distances for the negative and positive nearest neighbor in middle classes encodings extracted from TooT-Triplet (offline).	140
33	Class separated distributions of distances for the negative and positive nearest neighbor in minority classes encodings extracted from TooT-Triplet (offline).	141
34	Binding sites of protein sequence P06703.	142
35	AlphaFold 3D structure and binding site of P06703 on position 65.	142
36	Heatmaps of 30 attention heads in TooT-Triplet for P06703 over 16 layers.	143
37	Heatmap of head 13 in TooT-Triplet for P06703 over 16 layers.	144
38	Analysis of the sequence P06703 from UniProt-KB.	144
39	Venn diagram of TooT-SC and UniProt-SPEC-100 dataset sequences.	145
40	TooTranslator model with two components: vectorization and projection.	151
41	An example ChEBI entry with definition and SMILES string.	152
42	t-SNE analysis of TooTranslator prototype using UniProt-ICAT-100 dataset.	170
43	Results of 3-fold cross-validation for TooTranslator using four loss functions.	171
44	t-SNE plot for protein and label representations.	171
45	Distribution of distances to the true label for TooTranslator representations before and after training.	172
46	Distribution of correct prediction range for unseen class samples in TS2.	174
47	Confusion matrix for hold-out test set of UniProt-SPEC-100 using TooT- Triplet model (offline, Manhattan)	227
48	TooT-SC dataset results using TooT-Triplet	244

49	Comparison of class specificity in TooT-SC dataset and UniProt-SPEC-100 dataset	256

List of Tables

1	Performance of SOTA Models on Substrate Specificity Prediction.	8
2	The intermediate results: map <i>C2GO_C</i>	66
3	The intermediate results: map <i>C2GO_G</i>	67
4	Catalytic activity GO terms eliminated from C2GO map of ICAT datasets.	68
5	Four extracted TooT-ICAT dataset series.	68
6	Size of substrate classes for each TooT-ICAT dataset.	69
7	Five-fold cross-validation results for TooT-ICAT datasets.	69
8	Hold-out test set results on TooT-ICAT datasets.	70
9	Detailed results for UniProt-ICAT-100 dataset classification.	70
10	Four extracted TooT-SPEC datasets.	71
11	TooT-SPEC results on test sets for 4 different datasets.	75
12	TooT-SPEC 3-fold cross-validation results for UniProt-SPEC-100.	75
13	MCC comparison of unclassified classes in different folds.	77
14	UniProt-ICAT-binary dataset details	85
15	Calculated thresholds for each class based on cross-validation.	88
16	Results of the ICAT transporter detector.	89
17	Results of the hold-out test set of specific substrate predictor.	89
18	Results TooT-Open-ICAT trained with cv threshold.	90
19	Hyperparameter search space for TooT-Triplet with offline sampling strategy.	103
20	LoRA configuration details.	106
21	TooT-Triplet (offline) results on hold-out test set.	107
22	TooT-Triplet results for 3-fold cross-validation.	108
23	Hold-out test results for TooT-Triplet: offline sampling, 90 epochs, Man- hattan.	109
23	Hold-out test results for TooT-Triplet: offline sampling, 90 epochs, Man- hattan.	110

23	Hold-out test results for TooT-Triplet: offline sampling, 90 epochs, Manhattan.	111
23	Hold-out test results for TooT-Triplet: offline sampling, 90 epochs, Manhattan.	112
24	TooT-Triplet (online) results on hold-out test set.	112
25	Hold-out test results for TooT-Triplet: online sampling, 10 epochs, Manhattan.	112
25	Hold-out test results for TooT-Triplet: online sampling, 10 epochs, Manhattan.	113
25	Hold-out test results for TooT-Triplet: online sampling, 10 epochs, Manhattan.	114
25	Hold-out test results for TooT-Triplet: online sampling, 10 epochs, Manhattan.	115
26	Performance comparison of four model groups for one-shot classification on the hold-out test set of UniProt-SPEC-100 dataset.	117
27	MCC comparison for unclassified classes in TooT-Triplet and Baseline 1. .	118
28	Average distance of nearest neighbor after training.	120
28	Average distance of nearest neighbor after training.	121
28	Average distance of nearest neighbor after training.	122
29	Comparison of TooT-Triplet models in presence/absence of majority class.	123
29	Comparison of TooT-Triplet models in presence/absence of majority class.	124
29	Comparison of TooT-Triplet models in presence/absence of majority class.	125
30	SC dataset proteins with matching TooT-Triplet prediction; incorrectly categorized by GPT4.	130
30	SC dataset proteins with matching TooT-Triplet prediction; incorrectly categorized by GPT4.	131
31	Proteins with incorrect SC dataset labels predicted correctly by TooT-Triplet.	132
32	Sequences of SC dataset with the incorrect prediction using TooT-Triplet. .	132
32	Sequences of SC dataset with the incorrect prediction using TooT-Triplet. .	133
33	SC dataset entries with unmatching prediction where both could apply. . .	134
33	SC dataset entries with unmatching prediction where both could apply. . .	135
34	CHEBI terms and definitions for UniProt-ICAT-100 classes.	157
37	TooTranslator Huber loss function results for the hold-out test set.	158
37	TooTranslator Huber loss function results for the hold-out test set.	159

37	TooTranslator Huber loss function results for the hold-out test set.	160
35	Results of TooTranslator trained with four loss functions on TS1.	161
36	3-Fold Cross-validation results for TooTranslator with four loss functions. .	161
38	Results of the TooTranslator with Huber loss for TS2 dataset.	161
38	Results of the TooTranslator with Huber loss for TS2 dataset.	162
38	Results of the TooTranslator with Huber loss for TS2 dataset.	163
39	Distances from actual label on TS2 with three TooTranslator models. . . .	164
39	Distances from actual label on TS2 with three TooTranslator models. . . .	165
39	Distances from actual label on TS2 with three TooTranslator models. . . .	166
40	TooTranslator predicted labels for unseen class samples from TS2	173
41	The correct substrate prediction in TS2 for each top candidates range . . .	174
42	C2GO map with ChEBI terms and their corresponding GO terms	203
42	C2GO map with ChEBI terms and their corresponding GO terms	204
42	C2GO map with ChEBI terms and their corresponding GO terms	205
42	C2GO map with ChEBI terms and their corresponding GO terms	206
42	C2GO map with ChEBI terms and their corresponding GO terms	207
42	C2GO map with ChEBI terms and their corresponding GO terms	208
42	C2GO map with ChEBI terms and their corresponding GO terms	209
42	C2GO map with ChEBI terms and their corresponding GO terms	210
42	C2GO map with ChEBI terms and their corresponding GO terms	211
43	Details of substrates' sample count in four extracted datasets	212
43	Details of substrates' sample count in four extracted datasets	213
43	Details of substrates' sample count in four extracted datasets	214
44	Detailed test results of TooT-SPEC model on UniProt-SPEC-100	215
44	Detailed test results of TooT-SPEC model on UniProt-SPEC-100	216
44	Detailed test results of TooT-SPEC model on UniProt-SPEC-100	217
45	TooT-SPEC 3-fold cross-validation results for UniProt-SPEC-100	218
45	TooT-SPEC 3-fold cross-validation results for UniProt-SPEC-100	219
45	TooT-SPEC 3-fold cross-validation results for UniProt-SPEC-100	220
45	TooT-SPEC 3-fold cross-validation results for UniProt-SPEC-100	221
45	TooT-SPEC 3-fold cross-validation results for UniProt-SPEC-100	222
46	TooT-Open-ICAT prediction for transportome of 5 model organisms	224
47	TooT-Triple (offline, Euclidean) results for the hold-out test set of UniProt-SPEC-100	228

47	TooT-Triplet (offline, Euclidean) results for the hold-out test set of UniProt-SPEC-100	229
48	TooT-Triplet (offline, Cosine) results for the hold-out test set of UniProt-SPEC-100	229
48	TooT-Triplet (offline, Cosine) results for the hold-out test set of UniProt-SPEC-100	230
48	TooT-Triplet (offline, Cosine) results for the hold-out test set of UniProt-SPEC-100	231
49	TooT-Triplet (offline, manhattan) results for Fold 1 of UniProt-SPEC-100 .	231
49	TooT-Triplet (offline, manhattan) results for Fold 1 of UniProt-SPEC-100 .	232
49	TooT-Triplet (offline, manhattan) results for Fold 1 of UniProt-SPEC-100 .	233
50	TooT-Triplet (offline, manhattan) results for Fold 2 of UniProt-SPEC-100 .	233
50	TooT-Triplet (offline, manhattan) results for Fold 2 of UniProt-SPEC-100 .	234
50	TooT-Triplet (offline, manhattan) results for Fold 2 of UniProt-SPEC-100 .	235
51	TooT-Triplet (offline, manhattan) results for Fold 3 of UniProt-SPEC-100 .	235
51	TooT-Triplet (offline, manhattan) results for Fold 3 of UniProt-SPEC-100 .	236
52	SPEC results for triplet offline without proton Fold 1	237
52	SPEC results for triplet offline without proton Fold 1	238
53	SPEC results for triplet offline without proton Fold 2	238
53	SPEC results for triplet offline without proton Fold 2	239
53	SPEC results for triplet offline without proton Fold 2	240
54	SPEC results for triplet offline without proton Fold 3	240
54	SPEC results for triplet offline without proton Fold 3	241
54	SPEC results for triplet offline without proton Fold 3	242
55	Results for UniProt-SPEC-100 hold-out test set using Baseline 1	245
55	Results for UniProt-SPEC-100 hold-out test set using Baseline 1	246
56	Results for UniProt-SPEC-100 hold-out test set using Baseline 2	246
56	Results for UniProt-SPEC-100 hold-out test set using Baseline 2	247
56	Results for UniProt-SPEC-100 hold-out test set using Baseline 2	248
57	Results for UniProt-SPEC-100 hold-out test set using Baseline 3	248
57	Results for UniProt-SPEC-100 hold-out test set using Baseline 3	249
57	Results for UniProt-SPEC-100 hold-out test set using Baseline 3	250
58	Results for UniProt-SPEC-100 hold-out test set using TooT-SPEC and KNN (Manhattan)	250

58	Results for UniProt-SPEC-100 hold-out test set using TooT-SPEC and KNN (Manhattan)	251
58	Results for UniProt-SPEC-100 hold-out test set using TooT-SPEC and KNN (Manhattan)	252
59	Results for UniProt-SPEC-100 hold-out test set using TooT-SPEC and KNN (Euclidean)	252
59	Results for UniProt-SPEC-100 hold-out test set using TooT-SPEC and KNN (Euclidean)	253
60	Results for UniProt-SPEC-100 hold-out test set using TooT-SPEC and KNN(cosine)	254
60	Results for UniProt-SPEC-100 hold-out test set using TooT-SPEC and KNN(cosine)	255
62	TooT-Triplet predictions for transportome of five model organisms	265
62	TooT-Triplet predictions for transportome of five model organisms	266
63	Inter and intra-class distances for training set after triplet training	267
63	Inter and intra-class distances for training set after triplet training	268
64	Average distance of the nearest neighbor in TooT-SPEC training set after TooT-Triplet training on Proton Excluded dataset	268
64	Average distance of the nearest neighbor in TooT-SPEC training set after TooT-Triplet training on Proton Excluded dataset	269
64	Average distance of the nearest neighbor in TooT-SPEC training set after TooT-Triplet training on Proton Excluded dataset	270
65	TooTranslator results with weighted MSE loss function for the hold-out test set of UniProt-SPEC-100	271
65	TooTranslator results with weighted MSE loss function for the hold-out test set of UniProt-SPEC-100	272
65	TooTranslator results with weighted MSE loss function for the hold-out test set of UniProt-SPEC-100	273
66	TooTranslator results with MAE loss function for the hold-out test set of UniProt-SPEC-100	273
66	TooTranslator results with MAE loss function for the hold-out test set of UniProt-SPEC-100	274
67	TooTranslator results with MSE loss function for the hold-out test set of UniProt-SPEC-100 (description and SMILES)	275

67	TooTranslator results with MSE loss function for the hold-out test set of UniProt-SPEC-100 (description and SMILES)	276
68	TooTranslator results with MSE loss function for the hold-out test set of UniProt-SPEC-100 (SMILES)	276
68	TooTranslator results with MSE loss function for the hold-out test set of UniProt-SPEC-100 (SMILES)	277
68	TooTranslator results with MSE loss function for the hold-out test set of UniProt-SPEC-100 (SMILES)	278
69	TooTranslator results with MSE loss function for the hold-out test set of UniProt-SPEC-100 (description)	278
70	10 Nearest Neighbors predicted by TooTranslator-Huber for each test sample in TS2	280
71	Results of TooTranslator model trained on UniProt-ICAT-100 with label definition	296
72	Results of TooTranslator model trained on UniProt-ICAT-100 with label SMILES string vector	296
72	Results of TooTranslator model trained on UniProt-ICAT-100 with label SMILES string vector	297
73	Results of TooTranslator model trained on UniProt-ICAT-100 with label definition and SMILES string summation vector	297
74	Results of TooTranslator model trained on UniProt-ICAT-100 with label definition and SMILES string concatenation vector	297
74	Results of TooTranslator model trained on UniProt-ICAT-100 with label definition and SMILES string concatenation vector	298
75	Results of TooTranslator model with frozen weights label definition and SMILES string concatenation vector	298

Chapter 1

Introduction

1.1 Context

Proteins are one of the most prominent biomolecules across organisms. This group of biological macromolecules is responsible for major essential processes in a cell, from catalyzing chemical reactions and regulating gene expression to forming the major constituents of the cells. Therefore any study of cells relies on understanding the involved proteins, their structure, and functionality.

Definition 1.1.1. A *protein* is a biological macromolecule consisting of a chain of amino acids linked by peptide bonds.

A protein sequence is a chain of amino acids that folds into a three-dimensional structure, which determines its functionality. To elucidate the function and structure of a protein in a wet lab, biologists typically isolate an uncharacterized protein and perform various experiments. To reveal the structure of proteins, methods such as X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM) are widely employed [Wüt89, Küh14]. Functional characterization often relies on understanding the protein's location, interactions with binding partners, and sequence similarity. Techniques such as affinity chromatography, co-immunoprecipitation, and cross-linking mass spectrometry are used to identify protein-protein interactions, while homology-based methods aid in deducing functionality by comparing sequences to known homologs [AJL⁺03, FS89]. These approaches provide complementary insights into a protein's role and mechanism.

These wet lab experiments are increasingly complemented by *in silico* approaches, where computational experiments are conducted. The term *in silico* originates from the

silicon used in computer chips and refers to experiments conducted through computational models rather than laboratory methods. These methods, enhanced by integrated databases and advances in machine learning, have become reliable tools for guiding and prioritizing wet lab experiments [EMT07].

Definition 1.1.2. *The **cell membrane** is a lipid bilayer separating the interior of the cell from the outside environment, providing a fixed environment inside the cell protected from external conditions [ABH⁺15].*

Definition 1.1.3. *A **transmembrane protein** is a protein located in the cell membrane spanning the entire lipid bilayer, with regions exposed on both the cell's interior and exterior.*

Definition 1.1.4. *A **transmembrane transport protein** is a transmembrane protein that facilitates the movement of specific molecules across the cell membrane.*

This study focuses on a specific under-studied group of proteins called transmembrane transport proteins. As explained in detail in Section 1.2, this group of proteins plays an essential part in the cell by assisting in the relocation of different substances across the membrane. Despite their vital role in any organism, the challenging process of wet lab studies on this specific group of proteins necessitates *in silico* studies on transmembrane transport protein characterization.

Definition 1.1.5. *A **binding site** is a region on the surface of a macromolecule, such as a protein (e.g. a cavity or pocket) that binds to another molecule through intermolecular forces [SSC19].*

Definition 1.1.6. *In this context, a **substrate** or ligand in biochemistry is a chemical substance binding to a macromolecule such as a protein forming a complex to serve a biological purpose.*

Definition 1.1.7. ***Substrate specificity** is the characteristic of a protein to selectively transport a specific substrate [EGZ95].*

This research explores the substrate specificity of transmembrane transport proteins. The majority of these proteins have a substrate binding site that is highly selective, binding to a specific substrate (or a group of substrates with similar chemical structures) and transporting them across the membrane. As a result, each transport protein facilitates the movement of a particular chemical substrate, or a group of substrates, across the membrane [Dia14]. The substrate specificity of a transport protein refers to the specific

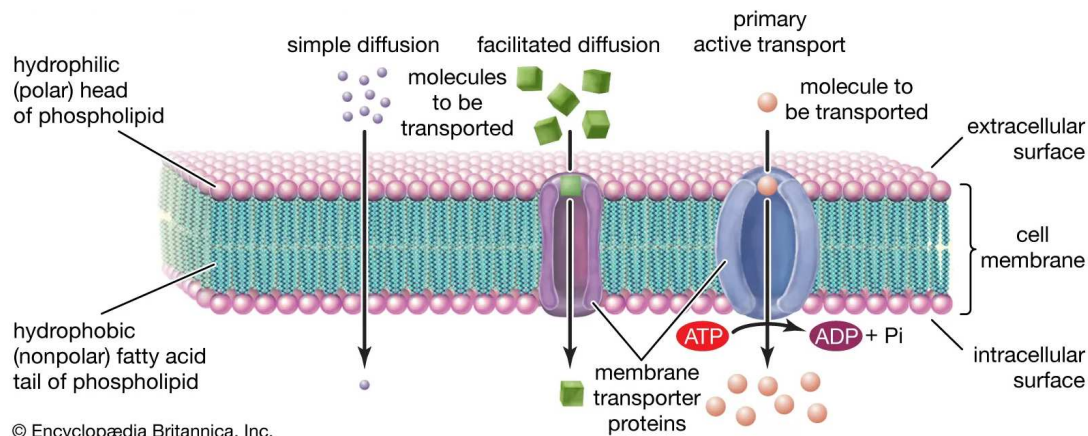


Figure 1: Transport process through the using different types of transmembrane transport proteins [DTM19].

substrate(s) it transports through the membrane [MD19]. For example, Figure 2 presents the This subject has been a matter of interest due to its importance in various biological processes e.g. drug-target interactions [RAAS11], neural signaling [PZA07], antibiotic resistance [WP03a], hormonal signaling [BS18, Gol87], cancer therapy [Brö20, ZWGC15], and more [WHSC14].

Recent *in silico* studies using machine learning approaches have effectively classified transmembrane transport proteins to the level of substrate classes, such that given transport protein the model predicts the class of the transported substrate (e.g. inorganic cations as the substrate class). The state-of-the-art research [AB22a] has successfully classified a dataset of transporters into 11 substrate classes. However, state-of-the-art is still far from solving the problem of predicting specific substrates (rather than the substrate class). First, the variety of specific substrates or substrate classes carried across the membrane is more than 11 classes. Second and more importantly, the capability of the models to predict uncharacterized proteins referred to as *de novo* in the literature is essentially non-existent.

Definition 1.1.8. A *de novo protein* is a new protein either designed from scratch or sequenced directly from biological samples without prior reference to existing protein databases or sequence information [MRIE24].

This research focuses on predicting the substrate specificity of transmembrane transport proteins and explores *de novo* sequence prediction as a more complex, yet unresolved, challenge. Following previous *in silico* studies, this research utilizes machine learning to

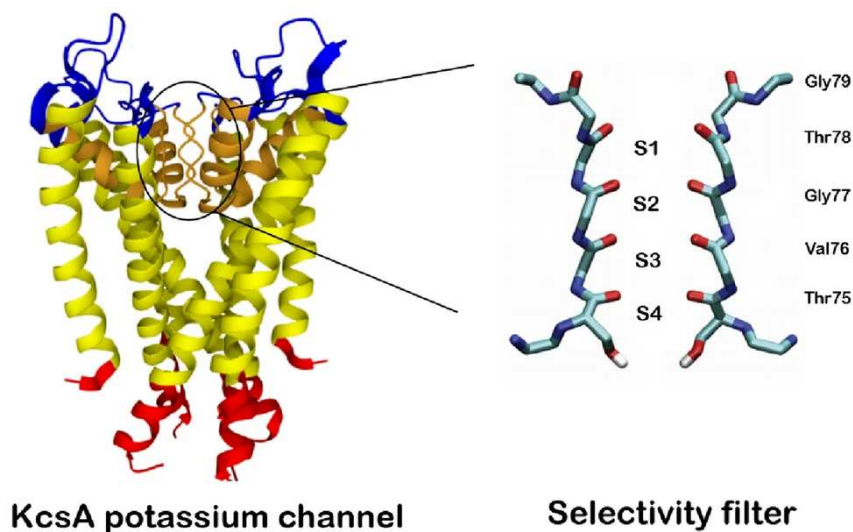


Figure 2: An example of the selectivity filter in transmembrane transport proteins. A KcsA potassium ion channel representation; PDB id: 1K4C (Left). Two P-loop monomers in the selectivity filter (Right) [SSS24].

solve the problem. One of the main challenges of this study is the lack of available curated data, leading the research toward applying low-shot learning algorithms (see Section 2.6).

This thesis is organized as follows: The remainder of this chapter discusses the research motivation, defining transmembrane transport proteins, their roles, and the limitations of studying this protein group in Section 1.2 and Section 1.4 details the research problem and establishes the research questions. Chapter 2 provides background and reviews state-of-the-art studies essential for understanding this research. Chapter 3 covers preliminary projects, including a pipeline for dataset construction and basic studies on fine-tuning protein language models. Chapter 4 expands on the open-world classification of inorganic ion transport proteins in the TooT-Open-ICAT study. Chapter 5 introduces TooT-Triplet, a distance-based model for one-shot classification of specific substrates. Chapter 6 explores substrate specificity prediction for *de novo* proteins in the TooTranslator project. The final chapter presents the thesis’s conclusions.

1.2 Motivation

1.2.1 Transmembrane transport proteins

Transmembrane transport proteins facilitate the selective passage of vital compounds ranging from small ions to macromolecules through intra- and extracellularly using either one of the processes of active transport or facilitated diffusion [LBK⁺21, Ste12]. This biological process is an important part of nutrient acquisition, development, cellular homeostasis, cell metabolism, signaling and communication, and drug uptake and efflux [YH15]. The substrates transported by these proteins can range from small ions such as sodium and potassium; sugars such as glucose; to larger compounds such as other proteins and messenger molecules and many more [LB94].

This group of proteins plays a significant role in all of the stages of any cell's life. The importance of transmembrane transport proteins is reflected in more than 60 human genetic diseases caused by the malfunctioning of transporters and channels [Dia14]. Investigating transporter substrate specificity helps develop targeted therapies. These therapies can address genetic defects caused by malfunctioning transport proteins and improve patient outcomes. As an example, consider drug uptake through the cell membrane process. The pharmacological product efficiency depends on the identification of the transporter associated with the specifically developed drug in target cells, binding to the correct transport protein, and completing the process of drug transport by the identified protein [AKEG09].

The process of genome-scale reconstruction of a metabolic network (GENRE) builds a map to represent the gene-protein-reaction (GPR) association between the genes, the proteins as gene products, to the reaction carried out by the protein [TP10]. This process covers cell metabolism, transport, regulation, and signaling. Ideally, a GENRE can assign gene-protein-reaction associations based on Enzyme Commission (EC) classification or molecular function terms of the genes on Gene Ontology. The transport reactions model the transport of substrates across membranes. The prediction of GPR association for transport proteins, in particular, the specific substrate, or substrates, is not well-developed [DP18, AB15].

The process of experimental characterization of substrate specificity given a transporter has notoriously been a challenging task. This group of proteins must be integrated into a membrane system to enable and detect their transmembrane activity. Also, the hydrophobic and metastable nature of these proteins makes the study conditions more

complex [LXHNE17]. After providing the correct setting for transmembrane transport protein to function, protein-ligand binding experiments are conducted.

The term “ligand”, widely used in the biochemistry field, in this context, is the specific substrate that a transport protein binds to and subsequently carries across the membrane. The wet lab assays for identifying the substrate specificity of a transport protein typically follow the following procedure: First, associate a candidate ligand (substrate), X, to a specific transporter, P, and then perform the experiments to determine the feasibility of the docking process as a result of intermolecular forces between the two parties [Pol10, FZC18]. During the described process, if the binding experiment of candidate substrate X and the protein P is not successful, the procedure needs to be repeated using the next substrate candidate.

These limitations leave transport proteins not sufficiently characterized in the proteomic databases. In particular, searching the following query, 16,355 transmembrane transport proteins are retrieved from the Universal Protein Knowledgebase (UniProt-KB version 2024_04, published on July 24, 2024, and accessed on September 17, 2024.) [The21b].

$\langle$ goa: (“transmembrane transporter activity [22857]”)
AND existence: “Evidence at the protein level [1]” $\rangle$

From this number only 24 protein sequences have evidence of the queried GO term for transmembrane transporter activity in an experimental level. The following query is applied to retrieve sequences with experimentally inferred GO terms.

$\langle$ goa: (“transmembrane transporter activity [22857]”,
evidence: “inferred from experiment [EXP]”) AND
existence: “evidence at the protein level [1]” $\rangle$

The scarcity of experimentally characterized transmembrane transport proteins arises from the challenges associated with substrate specificity characterization in wet lab settings. This highlights the need for computational methods, also known as *in silico* research to study this group of proteins [WBR⁺10]. Applying machine learning techniques to solve the substrate specificity problem could reduce the time and resources needed for the wet lab characterization process.

The computational approach proposed in this research follows a procedure different from that of wet lab experiments. Here, the proposed model learns the characteristics associated with each specific substrate transporter. In other words, the model is trained

on a class set of transmembrane transport proteins and predicts the specific substrate based on this training rather than experimenting with one-to-one protein-ligand binding feasibility. The training inputs are protein sequences labeled with their specific substrates, and predictions are made using various supervised computational models.

1.2.2 Computational methods on substrate specificity prediction

Various computational methods have been developed to classify transport proteins into substrate classes based on the substances they transport. As described in Section 2.10, these classification models employ traditional machine learning algorithms to assign transport proteins to substrate categories. Table 1 highlights the top three existing computational models for substrate class prediction. These methods, while effective to some extent, face limitations in computational efficiency and, more critically, focus on predicting broad substrate classes rather than specific substrates.

For example, the state-of-the-art research by Alballa et al. [AB22a] addresses this issue by employing support vector machines (SVMs) to classify protein sequences into 11 substrate classes: `inorganic cations`, `inorganic anions`, `water`, `organic anions`, `organo-oxygens`, `amino acids and derivatives`, `other organonitrogens`, `nucleotides`, `organic heterocyclics`, `miscellaneous`, and `nonselective`.

Although these models provide a valuable foundation by assigning protein sequences to their corresponding substrate classes, they fail to meet the need for substrate-specific predictions in practical applications. For instance, in drug-target interaction studies, broad classifications such as `inorganic cation` transporters are insufficient when precise information, such as identifying sodium-potassium pumps as specific targets, is essential for therapeutic development.

The main challenge in previous studies has been the limited availability of data, which has led to combining small classes into larger ones for training traditional machine learning models [AB22a, MCZ14]. This issue can be addressed using modern pre-trained models. These models rely on transfer learning and reduce the need for large datasets in classification tasks.

Definition 1.2.1. The *transportome* is the set of transmembrane transport proteins in an organism.

Table 1: Performance of SOTA Models on Substrate Specificity Prediction.

Method	# Classes	Minimum size	MCC	Issues
	Substrate	Class	Prediction	
TrSSP [MCZ14]	7	60	0.49	Slow (PSSM generation and SVM)
TranCEP [AAB20]	7	60	0.63	Slow to classify (MSA generation)
TooT-SC [AB22a]	11	24	0.82	Slow to classify (SVM ensemble)
	Specific	Substrate	Prediction	
TooT-ICAT [AB22b]	12	10	0.95	Limited coverage (inorganic ions)
TooT-SPEC	83	3	0.91	13 unpredicted classes

The first three entries are SOTA models for substrate class prediction and the last two entries are our preliminary studies on specific substrate prediction as presented in Sections 3.6 and 3.7.

Existing substrate classification models have two key limitations in addition to computational inefficiencies and their focus on general class predictions. First, their performance in transportome analysis is limited. A transportome contains protein sequences with unknown substrates, which many state-of-the-art (SOTA) models do not account for in predictions. Second, some known substrates are missing from the training sets of SOTA models due to insufficient samples or coverage. This leads to false classifications, as predictions are restricted to the classes seen during training.

Previous research handles this issue by defining a class of **Other transporters** with samples of various unknown classes. This solution confuses the model in the training process, searching for similarities between samples in one class, resulting in the lowest prediction performance in this class. For example, the state-of-the-art research [AB22a], defines this class as **Miscellaneous** which has the lowest values between 11 classes in all performance metrics, e.g. Matthews Correlation Coefficient (MCC) of 0.58 in 10-fold cross-validation. The second lowest class in the same experiment is **Other organonitrogens** with a gap distance of 0.1 in MCC.

The second deficiency of the state-of-the-art methods is on *de novo* sequence analysis. In the field of biology, the term *de novo* is used to refer to newly sequenced proteins that potentially have new functionality. The major problem in predicting the specific substrate transported by these proteins is discovering if this protein carries one of the existing substrate classes seen by the model or if it is a new substrate uncovered in the training set. The existing classification models assign a label from the trained dataset to any input protein ignoring the possibility of transporting new substrates.

1.3 Preliminary studies

The preliminary studies in this research investigate the use of transfer learning with large language models for predicting the substrate specificity of transmembrane transport proteins. These studies, detailed further in Chapter 3, initially focus on a subset of 12 transport proteins involved in transporting Inorganic Cations and Anions (ICAT) across membranes (Section 3.6). The research later expands to include 96 specific (SPEC) transporter classes in the TooT-SPEC project (Section 3.7).

As the first efforts in specific substrate prediction, these models demonstrate promising results by leveraging transfer learning in large language models. However, it is important to note that general class prediction methods are not directly comparable to specific class predictions due to differences in the defined class categories. As shown in the last two rows of Table 1, TooT-ICAT achieves a Matthews Correlation Coefficient (MCC) of 0.95 for 12 specific classes. This performance is extended to 83 specific classes in TooT-SPEC while maintaining comparable MCC.

The challenge of limited data availability per class is addressed, with the TooT-SPEC project achieving predictions for 83 out of 96 classes using as few as three samples per class. Nonetheless, these studies reveal persistent challenges in predicting minority classes within highly imbalanced datasets. Identifying these issues has played a critical role in refining research questions and guiding the design of subsequent projects.

1.4 Problem Definition

The overarching goal of this research is to develop an automated method for characterizing the specific substrate(s) transported by each transmembrane transport protein across the membrane. In other words, the objective is to create a model that, given a transmembrane transport protein sequence, predicts the specific substrate a transporter carries based on the input sequence. This model will also extend its predictive capabilities to *de novo* protein sequences that lack prior annotations. Given a set of transmembrane transport protein sequences, the model detects the specific substrate carried by the protein even if the model has never seen a transporter of that substrate.

Initially, this problem can be approached as a classification task, where transport protein sequences serve as samples and specific substrates act as labels. The classification model could be defined as multi-label since there are proteins that carry more than one

substrate or simplified to a single-label classification problem for training purposes with one main substrate. Here, we defined the dataset as single-labeled proteins to reduce the complexity of the problem, meaning that for each transporter in the dataset, there is only one label assigned. However, the problem could be expanded to multi-label as described in Section 3.4.6.

However, the insight of defining the project as a classification problem is only partially correct. A classification model correctly predicts the substrate carried by an input sequence if the model has seen the substrate class in the training process. On the other hand, it assigns any input sequence to one of the labels in the training set; ignoring the fact that the substances being transported may not be limited to the labels in our dataset. A closed world classification would limit the analysis of the transportome and the *de novo* sequences.

To build a reliable model for biologists, we need to open the model to the possibility of analyzing a newly sequenced *de novo* protein and predicting the substrate it carries regardless of the fact that the model has seen that class of substrate before or not. This paradigm is proposed to be implemented using two separate solutions of open-world classification and metric learning. As explained in Section 2.4 and Section 2.7 open-world classification does not identify the new substrate class but is able to reject the input sequence if it is not assigned to any of the classes in the dataset [SXL17]; while metric learning approaches are capable of expanding to unseen classes [MVPC12].

One of the main challenges in this study is the limited availability of curated data. Traditional machine learning models rely on large quantities of data per class to train effectively. However, this limitation has been addressed by leveraging a large pre-trained model. In this research, to encompass as many substrate classes as possible during training, we approach the problem as a few-shot, one-shot, or even zero-shot task, collectively referred to as low-shot learning (See Section 2.6). By utilizing pre-trained models, low-shot algorithms benefit from the knowledge transferred from the pre-training phase. The low-shot learning approach, based on metric learning algorithms, is detailed in Section 2.6.

1.5 Research questions

This research is focused on answering three main questions.

Question 1: “Given a transportome of a specific organism can a model predict the specific inorganic ions each protein transports?”

Question 2: *“Given a transmembrane transport protein from a long-tailed imbalanced dataset can a model predict the specific substrate it carries?”*

Question 3: *“Given a de novo transmembrane transport protein can a model predict the specific substrate it carries?”*

Each research question and the implemented solution are described in detail in Chapters 4, 5, and 6.

1.6 Challenges

The first challenge of this research is the lack of a complete dataset covering the transport proteins of any organism labeled with their specific transported substrate. The previous works cover datasets with labels from substrate classes not fine-grained to the specific substrate level. Chapter 3 introduces a pipeline to automatically build the datasets from a desired substrate group of interest based on Gene Ontology and ChEBI annotations.

The next challenge is addressing the class imbalance in the developed datasets. Due to their abundance in the cell membrane or research interests in the field of biology, some transporters are well-studied and thus have more protein samples available in UniProt, which is the source of our dataset. On the other hand, many substrate classes have few protein samples experimentally characterized in the dataset. For example, **sodium** (1+) transport proteins are a well-known group with more samples available in public datasets, while **folate** (2-) transporters are less studied and have fewer samples. This challenge is addressed by employing low-shot learning approaches, as detailed in Chapters 4 and 5.

1.7 Contributions

This research advances the characterization of transmembrane transport proteins by applying machine learning to substrate specificity prediction. It leverages large language models (LLMs) in various approaches to address key challenges in the field.

The study focuses on predicting substrate specificity in single-substrate transmembrane transport proteins, utilizing transfer learning with ProtBERT-BFD to classify substrates effectively. It extends this approach to long-tailed, imbalanced datasets and incorporates SciBERT and ChemBERTa-2 for novel class detection. By using pre-trained language models, the research demonstrates a significant reduction in the labeled data required while maintaining strong predictive performance.

The thesis further explores metric learning to improve classification for underrepresented classes in imbalanced datasets. An open-world classification approach with thresholded ProtBERT-BFD is introduced to detect unseen substrate classes. Additionally, a regression-based model is developed to map protein and substrate encodings into a shared space, enabling the identification of novel substrates.

Chapter 2

Background and related works

2.1 Introduction

This chapter establishes the foundation for understanding the key concepts behind this research. Specifically, it provides the background knowledge and formal definitions related to four major areas: protein language models, open-world classification, imbalanced datasets, and low-shot learning. Each of these topics plays a critical role in addressing the challenges posed by substrate specificity prediction in transport proteins, which often involves highly unbalanced data and unpredictable new classes.

First, Section 2.3 focuses on the evolution of protein representation methods, leading to powerful protein language models for extracting informative encodings from raw amino acid sequences. Next, Section 2.4 highlights the importance of open-world classification in biological contexts where novel substrates or protein functions may arise that were not part of the original training set. Following that, Section 2.5 provides an in-depth look at imbalanced datasets, which are widespread in real-world biological applications. In particular, we show how traditional machine learning methods struggle with the severe minority classes present in the transport protein dataset studied here.

Since multiple classes in our studied datasets have very few samples—even as low as three sequences—Section 2.6 explores low-shot learning approaches and offers a short review of prominent methods like FaceNet (an early few-shot method in computer vision) and Sentence-BERT (which brought few-shot ideas to natural language processing). We then connect these approaches specifically to protein sequence problems in Section 2.9, highlighting existing strategies for applying few-shot and one-shot learning to imbalanced protein data.

To complement these methodologies, Section 2.11 presents the evaluation metrics used in subsequent chapters, including both classification and clustering measures. Lastly, Section 2.12 provides an overview of the loss functions employed in different chapters, such as focal loss, triplet loss, and mean squared error, which are critical for addressing the challenges posed by imbalanced datasets and enhancing the learning process.

By bringing these diverse topics together, this chapter provides a clear foundation for understanding substrate specificity prediction in the chapters that follow.

2.2 Large Language Models: BERT

Definition 2.2.1. *A **Language Model (LM)** is a probabilistic model designed to predict the likelihood of sequences of words in a natural language [MRS08].*

Definition 2.2.2. ***Self-attention mechanism** is a method that calculates the relevance of each component in a sequence to constrain other components of the sequence [VSP⁺17].*

Recent advancements in Natural Language Processing (NLP) have been significantly driven by the development of Large Language Models (LLMs). Central to these models are transformer architectures [Wol20], which utilize self-attention mechanisms to effectively capture and represent the contextual relationships within textual data [VSP⁺17]. The transformer model has fundamentally transformed the field, enabling breakthroughs in various NLP applications, including language comprehension, text generation, and machine translation [DCLT19, BMR⁺20]. A notable example is BERT [DCLT19], which has served as a foundational model influencing numerous subsequent architectures. In this study, we employ BERT-based models, which build upon the original BERT framework to enhance performance and tailor the model to our specific research requirements.

Definition 2.2.3. *An **attention head** is a component of the self-attention mechanism that computes the attention scores and corresponding weighted values for each token in a sequence [VSP⁺17].*

Definition 2.2.4. *A **multi-head attention layer** applies multiple attention mechanisms in parallel, capturing diverse dependencies within a sequence. The outputs are concatenated and linearly transformed to the desired dimension, enabling the model to focus on different aspects of the input [VSP⁺17].*

Definition 2.2.5. ***Pre-training** a language model is the process of self-supervised learning using a large corpus of unlabeled text to learn general-purpose language representations [DL15, DCLT19].*

BERT (Bidirectional Encoder Representations from Transformers) [DCLT19] is a transformer-based model designed for natural language understanding. It features an encoder architecture with 12 layers and utilizes multi-head self-attention mechanisms to capture contextual relationships in text. The base version of BERT has a hidden size of 768 and contains around 110 million parameters. BERT is pre-trained on large datasets, including Wikipedia and BooksCorpus, enabling it to learn rich bidirectional representations [DCLT19]. It was established to address various linguistic tasks such as question answering [Yin22], sentiment analysis [HBR19], and language inference [ZWZ⁺20]. Additionally, numerous other language models have been developed based on BERT’s transformer architecture with different pre-training settings [Liu19, Lan19, San19]. One such BERT-based model, Prot-BERT [EHD⁺22], is pre-trained on protein sequences and serves as the main model studied in this research, explored in detail in Section 2.3. Additionally, numerous other language models have been developed based on BERT’s transformer architecture with different pre-training settings [Liu19, Lan19, San19]. One such BERT-based model, Prot-BERT [EHD⁺22], is pre-trained on protein sequences and serves as the main model studied in this research, explored in detail in Section 2.3.

Definition 2.2.6. ***Fine-tuning** is the process of adapting a pre-trained model to a specific downstream task by continuing its training on a smaller, task-specific labeled dataset [DL15].*

Definition 2.2.7. ***Transfer learning** is a technique in which a model adapts knowledge acquired from a source task or domain to improve learning efficiency and performance on another target task or domain [TS10].*

Definition 2.2.8. *In the context of BERT-based LMs, an **encoding vector** is the output of the model’s final transformer layer, providing a dense, contextualized representation of each token in the input sequence, capturing its semantic and syntactic relationships based on its surrounding context [DCLT19].*

Definition 2.2.9. *The **encoding space** in this context is a high-dimensional vector space populated by the encoding vectors produced by the model for all input samples.*

There are two conventional transfer learning approaches to using BERT-based language models in downstream classification tasks. First, a pre-trained language model could be utilized as an encoder, producing vectorized encoding for the input proteins. These encodings are fed into a classifier layer learning to predict the targets according to the specified task. This method is also known as encoding construction using a frozen language model. On the other hand, fine-tuning a BERT language model includes updating the weights of the model during the process of training to adjust the encoding according to the downstream task, which is classification in this case. Each of these methods is utilized according to the requirements of the defined task.

2.3 Protein Language Models: ProtBERT-BFD

The challenge of generating reliable features to represent protein sequences has been studied for decades. Traditional protein representations were derived from k-gram-based methods, such as amino acid composition (AAC), dipeptide composition (DPC), and tripeptide composition (TPC). Techniques like position-specific scoring matrices (PSSM) and Hidden Markov Models (HMMs) leveraged evolutionary information from multiple sequence alignments (MSA) to enhance protein encodings [GME87, Kos01]. Furthermore, physicochemical properties of amino acid residues, such as hydrophobicity, electrical charge, and size, have been incorporated as additional information in sequence embeddings [JDHL19]. However, in recent years, advances in machine learning have significantly impacted protein representation generation, as well as many other fields.

Definition 2.3.1. *A **Protein Language Model (PLM)** is a probabilistic model designed to predict the likelihood of sequences of amino acids in protein sequences, interpreting these sequences as analogous to sentences in a natural language, where each amino acid represents a token [FH22].*

Protein sequences in their primary structure are represented as chains of amino acids, similar to how sentences are composed of words in a language. By viewing the 20 amino acids as an alphabet, Protein Language Models (PLMs) leverage advances in Natural Language Processing (NLP) to capture and represent the underlying information in protein sequences [OBL21]. The introduction of transformer models, a breakthrough in NLP, has revolutionized unsupervised learning and shown significant promise for understanding protein sequences [VSP⁺17]. By learning contextual relationships between amino acids,

protein language models capture structural, functional, and evolutionary information, enabling tasks such as protein classification, structure prediction, and functional annotation.

In 2021, Elnaggar et al. [EHD⁺22] introduced six protein language models by training two auto-regressive models (Transformer-XL, XLNet) and four auto-encoder models (BERT, Albert, Electra, T5). This research trained each model on data from UniRef and BFD containing up to 393 billion amino acids on the Summit supercomputer using 5616 GPUs and TPU Pod up to 1024 cores. The ProtBERT model, as one of PLMs introduced in this paper, is a BERT-based protein language model benefiting the attention mechanism [DCLT19]. The advantage of using this PLM’s encoding vectors as exclusive input for several subsequent tasks is validated in this paper, and ProtBERT has shown promising results. Although not claimed as the best language model in this study, the model is widely used due to its accessibility, practical model size, and usability in fine-tuning tasks.

The Big Fantastic Database (BFD), a significant dataset used in training ProtBERT, is created by clustering 2.5 billion protein sequences from UniProt/TrEMBL plus SwissProt, Metaclust, and the Soil Reference Catalog Marine Eukaryotic Reference Catalog assembled by Plass. The BFD represents one of the largest publicly available collections of protein families, consisting of 65,983,866 families represented as MSAs and hidden Markov models (HMMs) covering 2,204,359,010 protein sequences from reference databases, metagenomes, and metatranscriptomes [JEP⁺21].

Definition 2.3.2. *Masked Language Modeling* is a pre-training objective where random tokens in a text sequence are replaced with a special [MASK] token, and the model is trained to predict these tokens based on their context, enabling the learning of deep bidirectional representations for language understanding [DCLT19].

ProtBERT-BFD, as the selected language model for our study, is a BERT model benefiting from the self-attention mechanism trained on BFD-100 datasets using Masked Language Modeling (MLM) task only. Compared to the original BERT publication, the number of layers was increased. In order to first extract useful features from shorter sequences while using a larger batch size, ProtBERT-BFD was first trained for 800k steps for sequences with a maximum length of 512 and then for another 200k steps for sequences with a maximum length of 2k. The optimizer used is Lamb with a learning rate of 0.002, a weight decay of 0.01, a learning rate warm-up for 140k steps, and linear decay of the learning rate after. This model is implemented in 30 layers, 16 attention heads, and 420M parameters.

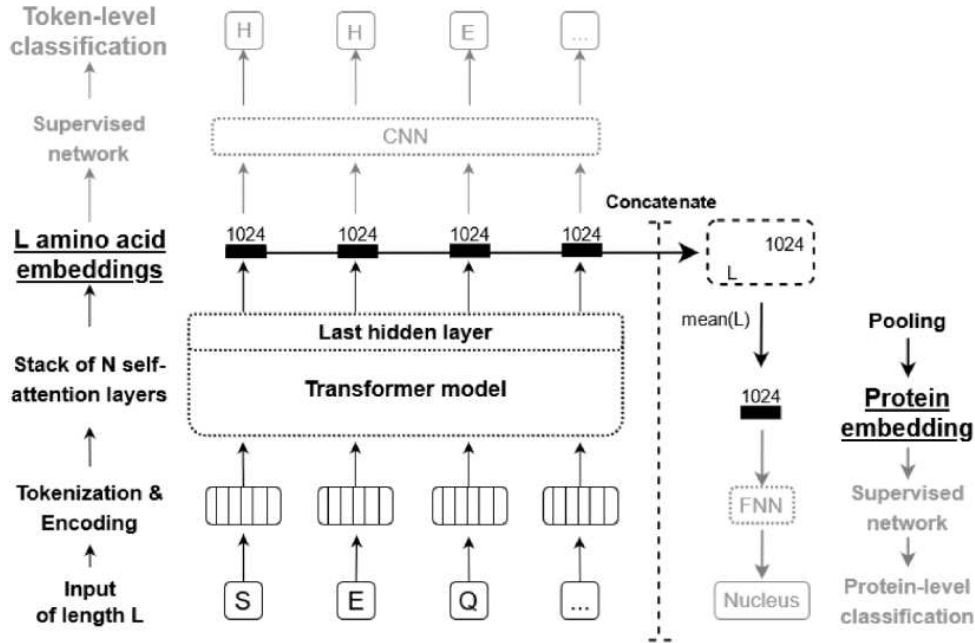


Figure 3: ProtBERT model architecture [EHD⁺22].

ProtTrans models can be used to derive features (embeddings) for arbitrary protein sequences either on the level of single amino acids or whole proteins and how they can be used for classification tasks on both levels. First, an example protein sequence “SEQ” is tokenized and positional encoding is added. The resulting vectors are passed through any of our ProtTrans models to create context-aware embeddings for each input token, i.e., each amino acid. Here, we used the last hidden state of the Transformer’s attention stack as input for downstream prediction methods. Those embeddings can either be used directly as input for prediction tasks on the level of individual tokens, e.g., a CNN can be used to predict an amino acid’s secondary structure. Alternatively, those embeddings can be concatenated and pooled along the length-dimension to get fixed-size embedding irrespective of the input length, i.e., global average pooling is applied. The resulting protein-level embedding can be used as input for predicting aspects of proteins, e.g., a FNN can be used to predict a protein’s cellular localization.

By capturing contextual information from large-scale protein datasets, ProtBERT-BFD and other language models overcome many limitations of earlier sequence representation methods. In the next section, we explore a critical requirement in real-world applications where novel or previously unseen classes of proteins may appear during testing: open-world classification.

2.4 Open-world classification

Definition 2.4.1. *A **classification** task is a supervised learning problem where the objective is to assign an input data point to one of several predefined discrete categories or classes [BN06].*

Classic supervised learning follows the closed-world assumption about classes: all classes that appear in the test data must have appeared in the training set [BB15]. Even though closed-world classification approaches have provided solutions to many real-world problems, in a dynamic and open environment this assumption does not apply.

Definition 2.4.2. ***Closed-world classification** is a classification framework in which the set of target classes is fully defined, finite, and known a priori during both training and test processes.*

Definition 2.4.3. ***Open-world classification** is a classification framework where the model is designed to classify instances from the classes seen during training while identifying and rejecting samples from unseen classes [SXL18].*

To clarify the concept of open-world classification, the basics of probabilistic classifiers are reviewed briefly. A probabilistic classifier assigns labels to the samples based on conditional distributions of the samples over the classes. The conditional distribution $P(Y|X)$ represents the probability that the input $x \in X$ belongs to the class $y \in Y$ given the features or characteristics of the input x . Naturally, the probabilities sum to 1 [HTF09].

In the case of binary classification, two values of 0 and 1 are defined as a label set Y . To assign a label to a sample t , the model calculates two probabilities of p_0 , the probability of the sample belonging to class 0, and p_1 , the probability of the sample belonging to class 1, summing up to 1. The model labels the sample with the class with a higher probability as the predicted label. In the case of multi-class classification, for n class classification

the predicted probabilities are $P = [p_0, p_1, \dots, p_n]$. The model predicts the label as follows [HTF09].

$$\hat{y} = \text{Argmax}_y(P(Y = y|X)) \quad (1)$$

This equation is modified in multi-class multi-label prediction to retrieve not only the maximum probability class as one label, but r labels regarding r top probabilities in the probability vector [ZE02].

Open-world classification or open classification does not hold the assumption of having the same classes in the training and test set and allows the classifier to classify unseen classes in the training set as rejected during the testing process. This assumption could be implemented on the probability vector by setting a threshold, τ , on the retrieved probability on the equation 1. This threshold enables the model to reject the sample and not assign any class labels to it if no probabilities in the vector pass the threshold [SXL18].

The formal definition of open-world classification is as follows. Given the labeled training data $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ where x_i is the i -th example and $y_j \in Y = \{y_0, \dots, y_n\}$ is x_i 's class label from n seen classes, we want to build a model $f(x)$ that can classify each test example x to one of the n seen classes in Y or reject it as unseen [SXL17]. For each sample in the test set t the predicted label $\hat{y}$ is:

$$\hat{y} = \begin{cases} \text{Reject}, & \text{if } \forall y_j \in Y, P(y_j|t) < \tau, \\ \arg \max_{y_j \in Y} P(Y = y_j|t), & \text{otherwise} \end{cases} \quad (2)$$

The problem of classifying specific substrates of transmembrane transport proteins is an open classification problem due to the possibility of locating new substrates in a biological system not seen in reference databases. These novel substrates could use either known or *de novo* transport proteins to pass through the membrane. Therefore to identify the presence of new substrates and their corresponding proteins we need to address the openness of the environment in this classification problem.

While open-world classification allows us to handle unseen classes at test time, real-world datasets also present a different but equally serious challenge: class imbalance. In the following section, we explore the nature of imbalanced datasets, why they arise in protein research, and how they can hinder model performance.

2.5 Imbalanced Dataset

Definition 2.5.1. An *imbalanced dataset* is a dataset where the distribution of instances across different classes is not uniform, resulting in one or more classes being significantly overrepresented (majority classes) while others are underrepresented (minority classes).

Definition 2.5.2. *Slight imbalance* is a type of class imbalance in a dataset where the distribution of examples is uneven by a small amount, with the ratio of the smallest to the largest class being less than 1:4.

Definition 2.5.3. *Extreme imbalance* is a type of class imbalance in a dataset where the ratio of the smallest to the largest class falls between 1:4 and 1:100, indicating a significant disparity in class distributions.

Definition 2.5.4. *Severe imbalance* is a type of class imbalance in a dataset where the ratio of the smallest to the largest class exceeds 1:100, indicating an extreme disparity in class distributions.

The common assumption of the uniform distribution for samples in the classes of the dataset does not hold true in many real-world datasets, causing imbalanced data problems. The ratio of imbalance is defined with the ratio of the smallest to the largest class as $s : l$, with s samples in the smallest and l samples in the largest class [HG09]. This ratio helps with categorizing the imbalanced datasets into three groups: *slight imbalanced*, *extreme imbalanced*, and *severe imbalanced* datasets. *Slight imbalance* where the distribution of examples is uneven by a small amount in the training dataset that is less than 1:4. *Extreme imbalance* is the ratio between 1:4 to 1:100 and any ratio more than 1:100 is considered a *severe imbalance* [Kra16].

Class imbalance could occur in the datasets of both binary and multi-class classification problems. Studying multi-class imbalanced datasets is not a straightforward task due to the complicated relationship between classes in multi-class problems; for example, a class might be majority compared to one class but balanced compared to another class in the dataset [SG12]. Most of the research on class imbalance is developed for the binary classification problems leaving multi-class imbalance a relatively under-studied subject. In addition, the majority of this research work is on *slight imbalance*, resulting in under-developed *severe imbalance* problems [ZXL⁺16]. The problem of specific substrate classification is a severe multi-class imbalanced problem, according to the dataset described in Section 3.7.2.

In Section 2.5.1, a specific group of multi-class imbalanced problems called *long-tailed imbalance* is explored. This group is introduced since the distribution of samples in the studied datasets follows the same pattern (see Section 3.7.2). In Section 2.5.2, the methods to train models of imbalanced datasets are explored: the basic imbalanced learning approaches of resampling and ensemble learning; and cost-effective learning and metric learning methods which are applicable to the long-tailed datasets.

2.5.1 Long-tailed imbalance

Definition 2.5.5. *Long-tailed imbalance* is a specific type of class imbalance in which a dataset demonstrates a long-tailed distribution, characterized by a few majority classes with many samples and numerous minority classes with only a few samples [ZKH⁺23].

Long-tailed or skewed *imbalance* in the dataset indicated a specific imbalance where the samples typically exhibit a long-tailed class distribution with a small portion of classes that have a sizable number of sample points (majority classes) but the others are associated with only a few samples (minority classes). The difference between the general imbalanced datasets and a long-tailed dataset is that in imbalanced datasets, the number of samples in the minority data is not necessarily small but in long-tailed datasets, there are a large number of classes and a tail of the distribution with few samples in them [ZKH⁺23].

2.5.2 Learning from imbalanced data

To solve the problem of class imbalance, three main approaches have been applied: data-level methods, algorithm-level methods, and hybrid methods [Kra16]. Data-level methods modify the collection of examples to balance distributions by resampling; introducing approaches that generate new objects for minority groups (oversampling); and that remove examples from majority groups (undersampling) [FGHC18, MRA20, LWZ08]. Algorithm-level methods directly modify existing learning algorithms to alleviate the bias toward the majority classes. The main approaches are cost-sensitive learning, contrastive learning, and ensemble learning [ZL10, MMB21, KWH15]. Hybrid methods combine the advantages of two previous methods: merging sampling and cost-sensitive learning techniques [WGC14].

The main concern with the severe class imbalance in small minority classes is the risk of poor representation and lack of clear structure in small classes. In such cases, the straightforward application of oversampling methods that rely on relations between

minority objects could result in overfitting the new samples on the weakly represented distribution of current samples [KKP06]. Another obstacle to resampling methods is the overlapping representation of different classes in the dataset. The overlapping of samples from different classes results in the unreliability of the basic resampling techniques. As depicted in Figure 42 the studied dataset suffers from this issue with the sample representations of the majority class overlapping with the minority classes.

Specifically re-balancing datasets with different methods of over/under-sampling is not a feasible solution for our research problem due to two reasons: (1) the high risk of overfitting the class distribution for small-scale classes and (2) overlapping sample representations between classes. Each reason is discussed in the following.

First, a high risk of overfitting the class distribution for minority classes with small sizes (3-10 samples). To balance the classes, this dataset needs to over-sample 400 samples for each minority class to match the majority class with over 1200 samples, with the risk of over-fitting on the distribution of existing samples. Under-sampling also loses information and leaves the dataset with scarce data points to train any machine learning algorithm. The second reason is the minority and majority sample representations are completely overlapping (see Figure 42). According to the 2-D t-SNE representations of minority and majority classes, the representations of the majority class are completely covering the samples of the minority class.

Another solution for learning from imbalanced datasets is the decomposition of the problem and applying ensemble learning techniques to develop a model of multiple binary classifiers in the form of one-vs-all (or one-vs-one) trained on each class (or bi-classes) of data. Although these methods could alleviate significantly the difficulty of the problem they suffer from the basic drawbacks of the potentially high number of base classifiers and the complexity of rebuilding the original problem from decomposed parts [Kra16]. Besides the mentioned problems, the small size of the minority classes in our studied datasets prevents us from benefiting from these approaches. As explained in Section 3.2, the minority classes of this study include 22 classes with only 3 samples and 48 classes of size 4-10 samples, which is considered insufficient data samples for training any binary classifiers.

Cost-sensitive learning methods, [LS08], adjust the training process for different classes to alleviate the bias toward majority classes. In other words, the learning process of a model is modified to incorporate varying penalties for each of the classes in the training set. This model learns the features of the training set according to a problem-specific cost

matrix. This matrix includes different cost penalties for the misclassification of different classes depending on the class size. According to the cost matrix, the model assigns a higher cost to the misclassification of the samples in the minority classes to augment their importance during the learning process.

While cost-sensitive learning with a cost matrix can be an effective approach for addressing class imbalance and adjusting the costs of misclassifications, these models are not the best solution for the proposed problem due to the following reasons [ZL10]. These approaches mostly rely on a cost matrix designed by experts in the field to discriminate the penalty to the prediction in different classes. This matrix is also sensitive to class distribution and any changes in class distribution need to adjust the cost matrix resulting in lower generalizability for minority classes with only a few samples.

In conclusion, the methods of rebalancing the dataset, cost-sensitive methods, and ensemble learning are solutions for many slight imbalance problems but have shortcomings that prevent them from performing reliably on long-tailed class imbalance. Therefore in the following section we focus on low-shot learning, explaining how it tackles the lack of sufficient examples in minority classes.

2.6 Low-shot Learning

Definition 2.6.1. *Few-shot learning* methods aim to solve a supervised task with only a small number of training samples for certain classes in the training set [BPPK24].

Definition 2.6.2. *One-shot learning* refers to solving a supervised task where only one example per rare class is available in the training set [BPPK24].

Definition 2.6.3. *Zero-shot learning* corresponds to solving a problem for which no training data exists for specific classes or tasks. Instead, these classes or tasks are defined using descriptions, often represented as vectors of numeric or symbolic attributes [LEB08].

Definition 2.6.4. *Low-shot learning* learning refers to tasks of zero-, one-, and few-shot learning collectively [LEB08].

Low-shot learning is used as an umbrella term for a model to learn information from few, one, or zero samples from a specific class in the training set [WYKN20]. The main motivation for developing low-shot learning models lies in the fact that these models have many useful applications in real-world problems. Many real-world problems suffer

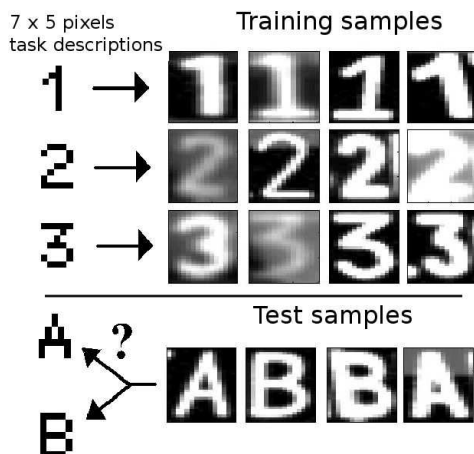


Figure 4: Zero-data learning of new tasks [LEB08].

The figure illustrates a zero-shot learning task. The training set includes three classes of “1”, “2”, and “3” with 4 samples present in each class. The test set includes two classes with zero samples available in the training set: “A” and “B”. The model classifies samples from these classes applying a zero-shot learning approach.

from scarce data samples due to data acquisition and manual labeling process drawbacks. Therefore, low-shot learning models that generalize from a few examples are a matter of interest in state-of-the-art studies [BMR⁺20, LEB08].

Low-shot learning algorithms are generally divided into meta-learning and non-meta-learning approaches. Meta-learning itself has sub-branches such as metric learning, model-based learning, optimization-based meta-learning, hybrid approaches, and so forth. Non-Meta-learning algorithms on the other hand include transfer learning and Task-Agnostic Meta-Learning (TAML) algorithms [SLCS19]. Since exploring all the methods in detail is not feasible in this report, we cover the definitions and methods of metric learning on low-shot classification problems, specifically siamese networks which are a proposed approach to one of our future projects.

Siamese networks have been selected as they are designed to learn a distance metric between the samples which incorporates both the intra- and inter-classes distance of the samples represented by encoding extracted from protein language models (ProtBERT-BFD). The impact of class imbalance on minority classes with few samples is manageable by selecting a suitable loss function, as explained in the following sections. Also, metric learning techniques have been successfully applied to similar protein data problems with promising results [HLS⁺22].

2.7 Metric-based methods

Definition 2.7.1. *Metric learning* is a technique that aims to learn an appropriate distance or similarity function directly from data, so that similar items are mapped closer together while dissimilar items are placed farther apart [BHS15].

Metric learning is the task of learning a distance function over data samples. The use of metric-based methods on low-shot learning relies on leveraging the meta-learning architecture to either learn an embedding function, given a distance function, or to learn both the embedding function and the distance function [PL22]. Four of the main architectures of metric-based few-shot learning methods are siamese networks [KZS⁺15], matching networks [VBL⁺16], prototypical networks [SSZ17], and relation networks [SYZ⁺18].

In the previous studies on long-tailed imbalanced datasets, siamese networks, and prototypical networks have been successfully applied to the few-shot classification of proteins. Siamese networks learn based on similarities between pairs of input. In contrast, the prototypical networks use the prototype representation of the samples in each class, which is prone to poor performance for small-scale classes. Therefore, we only cover siamese neural networks and their application to few-shot learning in protein dataset problems.

The above-mentioned techniques are designed for one-shot or few-shot learning. To classify *de novo* proteins, zero-shot techniques are required to generalize from the seen classes to the unseen. Here we cover ProTranslator as a state-of-the-art zero-shot learning model for protein function prediction in Section 2.9.2.

2.8 Few-shot classification using siamese networks

Siamese neural networks consist of two identical sub-networks joined at their outputs. During training, each sub-networks extract features from one input sample, while the joining neuron measures the distance between the two feature vectors [BGL⁺93]. The parameters between the twin sub-networks are tied assuring that two similar inputs could not possibly be mapped to different locations in feature space because each sub-network computes the same function [KZS⁺15]. The architecture has shown promising results in various image processing tasks including real-time object tracking [HLTZ18], face recognition [RWC16], and recently on natural language processing tasks such as semantic textual similarity [ROM19].

Early developments of few-shot classification were made on image processing tasks of

signature verification and later on face recognition tasks. To cover basic developments of few-shot classification, we start with FaceNet [SKP15] which has introduced a triplet loss function for few-shot learners. Next, we explore developments of the siamese networks on natural language processing by covering the Sentence-BERT paper [RG19].

2.8.1 FaceNet

Although the FaceNet research is defined as a face recognition, verification, and clustering project, this paper is covered in this section due to valuable contributions to metric learning, referred to in the proposal of our future research. FaceNet presents a model that directly learns a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity [SKP15]. A novel loss function, triplet loss, has been introduced to discriminate the matching pairs of samples from unmatching pairs by a distance margin (Figure 5). This loss function is explored in detail in Section 2.12.2.

The selection of the triplets is important for the convergence of the network and triplets violating the triplet constraint in the Equation 15 is crucial to the training process. The paper introduces two different approaches to selecting the triplets from the dataset: online and offline triplet selection. In the offline method, triplets are generated every m step, using the most recent network checkpoint. In the online triplet generation, hard positive/negative exemplars from within a mini-batch are selected.

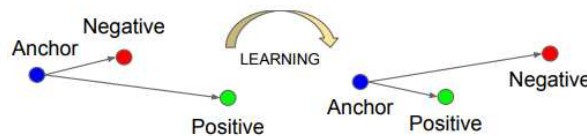


Figure 5: The learning effect using Triplet Loss function.

This loss function minimizes the distance between an anchor and a positive sample from the same class and maximizes the distance between the anchor and a negative from different classes [SKP15]

In the training process, CNN models were trained using the Stochastic Gradient Descent (SGD) algorithm with standard backpropagation, AdaGrad optimizer, and triplet loss function. This model is implemented using ZF-Net and Inception networks separately, trained on 100M-200M thumbnails of faces for about 8M different people [ZF14, SLJ⁺15]. FaceNet outperforms the best-published results by cutting the error rate by 30% and

achieves a record accuracy of 99.63% on Labeled Faces in the Wild (LFW) and 95.12% on YouTube Faces datasets. The use of this methodology and loss function has been expanded to other fields of machine learning such as natural language processing [RG19].

2.8.2 Sentence-BERT

Sentence encodings using siamese BERT-Networks have been introduced in paper [RG19] to overcome the massive overhead of using transformers on sentence-pair regression tasks e.g. semantic textual similarity (STS). According to this research, using a cross-encoder in BERT architecture is unsuitable for various pair regression tasks due to too many possible combinations. For example, they state that finding the highest similarity with BERT in a collection of $n = 10,000$ sentences requires $n(n - 1)/2 = 49,995,000$ inference computations, which require about 65 hours on a **Nvidia V100 GPU**. They also show that fixed-size BERT encodings, by averaging the BERT output layer, or using the output of the first token (the [CLS] token), yield rather bad sentence encodings, often worse than averaging GloVe encodings [PSM14].

Sentence-BERT, or SBERT, uses BERT on a siamese network architecture with a similarity measure like cosine similarity 4 or Manhattan / Euclidean distance 3 to find semantically similar sentences.

$$x, y \in R^d, k \in \mathbb{Z} \quad L_k(x, y) = \left(\sum_{i=1}^d ||x^i - y^i||^k \right)^{1/k} \quad (3)$$

$$x, y \in R^d \quad similarity(x, y) = \cos(\theta) = \frac{x \cdot y}{||x|| ||y||} \quad (4)$$

These similarity measures can be performed highly efficiently on modern hardware, as in the previous example, most similar sentence pairs in a collection of 10,000 sentences are reduced from 65 hours with BERT to the computation of 10,000 sentence encodings (~ 5 seconds with SBERT) and computing cosine similarity (~ 0.01 seconds).

In order to fine-tune the BERT model, the paper suggests creating siamese and triplet networks to update the weights such that the produced sentence encodings are semantically meaningful and can be compared with cosine similarity. The model has been implemented separately for three different objective functions: classification objective function, regression objective function, and triplet objective function.

For the classification objective function, a softmax function of the weighted concatenation of sample encoding along with their element-wise difference for two-sentence encodings is trained using a cross-entropy loss function. For the regression objective, the cosine

similarity between the two sentence encodings is computed with a mean squared-error loss as the objective function. The triplet objective function implements a triplet loss function given an anchor sentence a , a positive sentence p , and a negative sentence n .

The model has been tested on seven Semantic Textual Similarity (STS) tasks which the simple BERT model struggled on and achieved a significant improvement over state-of-the-art sentence encoding methods. Improving the GPU running time 9% from InferSent and 55% from Universal Sentence Encoder as the baseline method, SBERT is proven computationally efficient and capable of solving tasks that are computationally not feasible to be modeled with BERT. This model has been widely used in many studies on semantic textual similarity in the natural language processing field.

Low-shot learning techniques help when data are scarce, enabling models to generalize from just a handful of samples. This is especially relevant in protein sequence datasets where minority classes can be highly underrepresented. Next, we delve into how low-shot classification approaches—particularly in protein sequence analysis—have been applied to solve real-world challenges in protein function and substrate prediction.

2.9 Low-shot classification on protein sequences

Low-shot classification methods refer to the few/one/zero-shot learning approaches in machine learning. Here, we briefly review two of the most relevant state-of-the-art papers related to studies of low-shot learning on protein sequences. The first presented research successfully applies the Triplet loss function with the BERT model in four downstream tasks [LRD21]. The second paper demonstrates a translator model for the zero-shot classification of protein sequences based on their function [XW22].

2.9.1 Deep learning proteins using a Triplet-BERT network

A Triplet-BERT network [LRD21] has been introduced on protein datasets based on a pre-trained BERT model to be fine-tuned on small-scale protein datasets. Spotlighting the scarcity of labeled resources in protein data, this research leverages a pre-trained BERT model to model a collection of regression tasks using only a minimal amount of data. The model has been tested on four downstream tasks: plasma membrane localization, thermostability, peak absorption wavelength, and enantioselectivity. Significantly improving the original BERT baseline as well as the previous state-of-the-art models for

each task, this research demonstrates the benefits of using a triplet network for refining such a large pre-trained model on a limited dataset [LRD21].

This model is developed based on a model pre-trained on a large corpus of proteins as described in [RBT⁺19]. Triplets of anchor, positive, and negative protein sequences are used along with weight-sharing within the BERT model to cluster the data based on a triplet loss in the training process. During the fine-tuning process, the model has been adapted to each of the four defined tasks separately. The results for each downstream task indicate that the triplet-BERT network produced more detailed encoding compared to the original pre-trained BERT encoding and previous baselines. The state-of-the-art results for all downstream tasks are achieved using both strategies of pre-training and metric learning compared to using just one of these approaches.

2.9.2 ProTranslator

ProTranslator [XW22] is a zero-shot classification model developed to solve the task of protein function prediction. This model relates proteins to functions based on the textual descriptions available for the functions. Redefining the problem as a machine translation task, the authors propose a model to translate the description word sequence of a function into the amino acid sequence of a protein. By transferring annotations from functions with similar textual descriptions, ProTranslator can annotate novel functions. The model is claimed to function as a protein function “search engine”, which allows users to retrieve a list of proteins based on free-text queries.

Challenging the restrictions of defining protein function prediction as a closed-world classification problem, ProTranslator considers the potential novel function sets in the problem space. The study includes the following steps:

1. **encoding GO functions:** The model calculates vector representations for Gene Ontology (GO) functions based on their textual descriptions, using the PubMedBERT model [GTC⁺21].
2. **encoding proteins:** Protein encodings are calculated based on sequence, textual descriptions, and protein-protein interaction (PPI) networks. For sequences, convolutional neural networks (CNNs) are used to extract feature vectors. Protein descriptions are retrieved from GeneCards and encoded using a BERT model fine-tuned for the task. The network information is derived from PPI networks constructed using publicly available datasets, such as STRING and BioGRID. Each protein’s

topology in the PPI network is used to generate encodings through a graph neural network approach.

3. **Combining features:** The feature vectors for sequence, description, and network are processed through individual one-depth fully connected layers to ensure that their dimensions are uniform. These processed vectors are then concatenated into a single protein representation vector, which is projected into a shared low-dimensional encoding space along with GO term vectors.
4. **Protein function prediction:** Protein function prediction is performed by calculating similarities between the projected GO term encodings and protein encodings in the shared space. The model is also used to annotate novel functions, sparse functions, and gene sets to pathways.

The performance of ProTranslator is evaluated on multiple datasets for GO term prediction in zero-shot and few-shot settings. It outperforms state-of-the-art methods with improvements in precision, recall, and AUROC. For example, the AUROC values in zero-shot predictions were notably higher than competing models, demonstrating its capability to generalize to unseen functions. However, the paper also notes a performance drop in predicting novel functions (functions with few or no annotations) compared to functions with many annotations, indicating limitations in handling under-studied functions. Despite this limitation, the authors highlight that ProTranslator can significantly narrow down candidate proteins for a given new function, reducing the effort required for wet lab experiments.

While the model shows strong potential, certain limitations exist. ProTranslator relies heavily on the availability of textual descriptions for new functions, which can be challenging for under-studied or poorly described functions. Additionally, the reliance on high-quality PPI network data may restrict its applicability in cases where such data is incomplete or unreliable. The authors suggest that ProTranslator is best suited as a tool to prioritize protein candidates for experimental validation, rather than as a standalone solution for function prediction.

These low-shot approaches illustrate how models can deal with scarce protein data effectively. Building on this, the next section examines substrate class prediction, reviewing existing research to show how open-world classification, imbalanced datasets, and low-shot methods converge in addressing transporter specificity problems.

2.10 Substrate class prediction

The early efforts on transporter identification and substrate specificity prediction mostly rely on sequence homology. For homologs of known transporters, the standard applications of similarity search and HMM models of orthologous protein families have proven effective. TranSyT [LFL⁺23] is a component of Merlin 4.0, the Metabolic Models Reconstruction Using Genome-Scale Information [DRFR10], which identifies transport systems and the compounds carried across membranes, based on the annotations of the Transporter Classification Database (TCDB) [SJRT⁺16]. The tool utilizes the mappings in TCDB and KEGG to determine substrates for transport reactions during the reconstruction of pathway networks by assigning TC identifiers, GO terms, and ChEBI terms to sequences.

In 2014, Mishra et al. conducted the first study of specific substrate prediction using machine learning algorithms, classifying transporter sequences into seven substrate classes: amino acid, anion, cation, electron, protein/mRNA, sugar, and other transporters [MCZ14]. This research utilizes integrative protein sequence features such as amino acid composition, dipeptide composition, physicochemical composition, biochemical composition, and position-specific scoring matrices (PSSM) as inputs of the classifier.

The final model called TrSSP, Transporter Substrate Specificity Prediction Server, reports classification results for a Support Vector Machine(SVM) with an RBF kernel and biochemical composition and PSSM as the input features outperformed other classifiers with prediction accuracy of 76.69% with a Mathews correlation coefficient (MCC) of 0.49. The paper suggests that evolutionary information (i.e., the PSSM) and the AAIndex hold more discriminative information in this problem in comparison to similarity-based methods such as BLAST, PSI-BLAST, and hidden Markov models [MCZ14].

TranCEP (Transporter substrate prediction using Compositional, Evolutionary, and Positional information) paper [AAB20], follows TrSSP research on same the dataset, combining the traditional use of the amino acid composition of the protein, with evolutionary information captured in a multiple sequence alignment (MSA). This research suggests the transmembrane segments (TMSs) as vital segments of the sequences in substrate prediction problem, utilizing them using the transitive consistency score (TCS) [CDTTN12]. Developing an SVM model with TM-coffee MSA algorithm from pair amino acid composition (PAAC) encoding scheme inputs, they outperform TrSSP results with an MCC of 0.63.

The first study to benefit from natural language processing approaches in protein representation for substrate class classification is conducted using the word embedding technique. In this research each protein sequence is split into biological words of the same length in an overlapping fashion; then word embedding vector for each biological word using the skip-gram [MCCD13] or Continuous Bags of Words (CBOW) models [GAL⁺06] is generated. Evaluations show the SVM outperforms Naive Bayes, Random Forrest, and K-Nearest Neighbours classifiers on a dataset of 8 substrate classes with an MCC of 0.88.

The authors of TranCEP expanded the research from predicting transport proteins from 7 substrate classes to 11 substrate classes in the TooT-SC study [AB22a]. The substrate classes in the produced dataset using an ontology-based transporter substrate annotation for benchmark datasets approach include 11 substrate classes of inorganic cations, inorganic anions, water, organic anions, organo-oxygens, amino acids and derivatives, other organonitrogens, nucleotides, organic heterocyclics, miscellaneous, and non-selective. Applying the classification of the SVM model with an RBF kernel using a compositional, evolutionary, and positional information approach from TranCEP, this research reports 0.82 MCC over 11 substrate classes establishing as the state-of-the-art in the case of a number of substrate classes in substrate class prediction [AB19]. This study is presented as the baseline for our research.

Having covered the broader landscape of substrate classification research, we can now focus on how to measure and compare model performance effectively. Therefore, we turn to evaluation metrics in the next section, outlining the classification and clustering measures used throughout this study.

2.11 Evaluation Metrics

In this section, the evaluation metrics utilized in the following chapters of the manuscript are presented. These metrics evaluate the model’s performances on clustering and classification results. To evaluate the consistency of sequence encodings for each class of data, a clustering evaluation metric is utilized. The sample encodings of each class are considered as a cluster in the encoding space. The silhouette distance measures two factors: cohesion, which is the similarity of an object to its own cluster, and separation, which is the similarity of an object to other clusters [Rou87]. To calculate the silhouette coefficient for each data point i in cluster C_I , let the intra-cluster distance of i and all the samples in C_I as $a(i)$ be calculated as follows.

$$\forall i \in C_I \quad a(i) = \frac{1}{|C_I|-1} \sum_{j \in C_I, i \neq j} d(i, j) \quad (5)$$

$d(i, j)$ is the distance between data points i and j in the cluster C_I and $|C_I|$ is the number of data points included in cluster C_I . Next, the mean inter-cluster distance of i and all the samples in C_j as $b(i)$ be calculated as follows.

$$\forall i \in C_I, C_I \neq C_j \quad b(j) = \min(\frac{1}{|C_J|} \sum_{j \in C_J} d(i, j)) \quad (6)$$

In this setting, $a(i)$ evaluates the cohesion while $b(i)$ evaluates the separation in the clusters. Based on these two metrics, the Silhouette coefficient is defined as follows.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad \text{if } |C_I| > 1 \quad (7)$$

or

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & \text{if } a(i) < b(i) \\ 0 & \text{if } a(i) = b(i) \\ \frac{a(i)}{b(i)} - 1 & \text{if } a(i) > b(i) \end{cases} \quad (8)$$

According to the Equation 8 the silhouette coefficient ranges between -1 and 1.

To evaluate the performance of the classification models, five different metrics were considered:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \quad (11)$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (13)$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. Imbalanced

data less influence the Matthews Correlation Coefficient (MCC) and is arguably the best single assessment metric in this case [Din11, WP03b, BDA13]. The overall performance across all classes is the micro-average of the individual results [MRS08] and we used the multi-class version of MCC [Gor04].

A consistent and fair evaluation framework is crucial for benchmarking models on real-world datasets. However, results also hinge on the choice of loss functions during training. In the following section, we discuss the loss functions—such as focal loss, triplet loss, and MSE—that play a pivotal role in optimizing our models under challenging conditions like severe imbalance and open-world settings.

2.12 Loss Functions

Definition 2.12.1. *A **loss function** in deep learning is a mathematical function that quantifies the difference between the predicted outputs of a model and the true target values.*

The goal of training is to minimize the calculated loss by updating the model’s parameters. Different loss functions are employed in the training process of the models presented in this manuscript. This section provides a formal definition of some of the more specialized loss functions and discusses their application in various scenarios. Specifically, six loss functions used in Chapters 4, 5 and 6 are described: Focal loss [LGG⁺17], Triplet loss [SKP15], Mean Squared Error (MSE) loss, weighted Mean Squared Error loss, Mean Absolute Error (MAE) loss [Ber13], and Huber loss function [Hub92].

2.12.1 Focal loss

The focal loss function is designed to address the issue of class imbalance ignored in the commonly used cross-entropy loss. This loss function assigns less weight to well-classified examples and concentrates the model’s learning on more difficult, misclassified examples. The probability of the true class denoted as p_i , represents the model’s estimated likelihood for class i . A modulating factor $(1 - p_i)^\gamma$, where $\gamma \geq 0$ is a tunable focusing parameter, reduces the influence of easy-to-classify samples. The focal loss is expressed as [LGG⁺17]:

$$\text{Focal Loss}(p_i) = -(1 - p_i)^\gamma \log(p_i) \quad (14)$$

With γ parameter, the focus of loss is adjusted to different samples; higher values of γ apply more emphasis on hard examples with low p_i .

2.12.2 Triplet loss

The triplet loss function defines three sets of samples for each class: anchor (a), positive (p), and negative (n). Considering any samples of class x as an anchor, a positive sample is another sample from class x and a negative sample is a data point from a class other than x . Figure 5 shows the training objective using this loss function is to minimize the distance between the anchor and positive samples from the same class and maximize the distance between the anchor and a negative from different classes.

For all possible sets of triplets in the training set, T , with cardinality N , the triplet loss function is defined on Equation 17. a , n , and p are anchor, negative, and positive samples respectively and $f(x_i)$ shows the representation of the sample x_i . The parameter α is defined as a margin that is enforced between positive and negative pairs [SKP15]. The *hard positive* and *hard negative* samples are defined as, given an anchor x_i^a and a positive sample, x_i^p , and a negative sample, x_i^n , $\operatorname{argmax}_{x_i^p} \|f(x_i^a) - f(x_i^p)\|_2^2$ and $\operatorname{argmin}_{x_i^n} \|f(x_i^a) - f(x_i^n)\|_2^2$ respectively.

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2 \quad (15)$$

$$\forall (f(x_i^a), f(x_i^p), f(x_i^n)) \in T \quad (16)$$

$$\text{Triplet Loss} = \sum_i^N \max[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 - \alpha, 0] \quad (17)$$

2.12.3 Mean Squared Error and weighted Mean Squared Error

Mean Squared Error (MSE) or L2 loss calculates the average squared difference between the model's predicted values and the actual values. For n number of samples, y as the actual value, and $\hat{y}$ as the predicted value of each sample, the MSE loss is calculated as follows:

$$\text{MSE Loss} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (18)$$

In many cases, different classes are valued differently during the training process. For example, in the case of an imbalanced dataset, classes with a lower number of samples are given greater priority compared to the majority classes. To address this, a weight w_{c_i} is applied to each class c_i in the MSE loss to account for the varying significance, as shown in Equation 19 [Ber13]:

$$\text{Weighted MSE Loss} = \frac{1}{n} \sum_{i=1}^n w_{c_i} (y_i - \hat{y}_i)^2 \quad (19)$$

2.12.4 Mean Absolute Error

Mean Absolute Error (MAE) or L1 loss calculated average of the absolute difference between the model’s predicted values and the actual values. In contrast to MSE loss, which penalizes larger errors more heavily and is sensitive to outliers, MAE loss does not square the errors, treating all deviations equally. This characteristic makes MAE more robust to outliers in the data. Therefore, MSE is preferred when the goal is to emphasize larger errors and when outliers are a significant aspect of the problem. On the other hand, MAE is useful when dealing with data containing outliers or when a more consistent, robust error measure is desired without highlighting large deviations. For n number of samples, y as the actual value, and $\hat{y}$ as the predicted value of each sample, the MAE loss is calculated as follows:

$$\text{MAE Loss} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (20)$$

2.12.5 Huber Loss

The Huber loss function is introduced to address challenges in the robust estimation of location parameters, making the estimator less sensitive to outliers. This function has combines Mean Squared Error (MSE) and Mean Absolute Error (MAE) loss functions, advantaging in cases where the sample mean is excessively influenced by a few extreme values in long-tailed distributions. Defining y as the actual value, $\hat{y}$ is the predicted value for a sample, and δ as a constant threshold, the Huber loss function is quadratic for small value differences and linear for larger value differences as in the following equation [Hub92].

$$\text{Huber Loss}(y_i, \hat{y}_i) = \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2 & \text{if } |y_i - \hat{y}_i| \leq \delta, \\ \delta (|y_i - \hat{y}_i| - \frac{1}{2}\delta) & \text{otherwise} \end{cases} \quad (21)$$

2.13 Summary

This chapter provides a comprehensive overview of key topics that support this research, covering protein language models (Section 2.3), open-world classification (Section 2.4), imbalanced datasets (Section 2.5), low-shot learning (Section 2.6), evaluation metrics (Section 2.11), and essential loss functions (Section 2.12). The chapter begins by examining protein language models, with a focus on the ProtBERT-BFD model, which uses

natural language processing techniques to derive informative encodings from amino acid sequences.

Next, the chapter introduces open-world classification (Section 2.4), emphasizing its role in biological contexts where unseen protein functions or novel substrates may emerge. It then explores imbalanced datasets (Section 2.5), a frequent challenge in real-world data that often leads to overfitting on majority classes when handled improperly.

To address such challenges—especially when minority classes include few samples—the chapter covers low-shot learning approaches (Section 2.6). These approaches draw on foundational methods from image processing (e.g. FaceNet) and sentence-level NLP (Sentence-BERT), and they extend to protein sequence analysis, where current research shows promising results on scarce and imbalanced biological datasets.

The chapter then presents evaluation metrics (Section 2.11)—such as precision, recall, F1-score, and MCC—that are used throughout this dissertation to measure both classification and clustering performance. These metrics help distinguish genuine improvements from overfitting or data bias. Finally, it highlights a range of loss functions (Section 2.12), including focal loss, triplet loss, and mean squared error. Each loss function has a specific purpose, especially for dealing with skewed data distributions or enhancing the representation of minority classes.

By integrating these topics, the chapter offers a unified view of ongoing research in substrate specificity prediction and the broader field of protein annotation. Protein language models (Section 2.3) advance sequence representation quality, while open-world (Section 2.4) and low-shot (Section 2.6) frameworks address the dynamic and unpredictable nature of biological systems. Although imbalanced datasets (Section 2.5) remain a substantial challenge, specialized metrics (Section 2.11) and targeted loss functions (Section 2.12) enable ongoing progress in modeling and prediction.

Subsequent chapters build on these foundations to propose and evaluate novel approaches for substrate specificity prediction. By incorporating the insights and best practices presented here, this work aims to push the boundaries of computational analysis for transport proteins and their substrates.

Chapter 3

Preliminary Study

3.1 Introduction

In this chapter, our preliminary research on substrate specificity prediction is presented. Due to the lack of machine learning-based research on cross-organism prediction of substrate specificity, there are no available datasets of protein sequences labeled with their transported substrate. In Section 3.2, we present an automated pipeline to construct machine learning ready-to-use datasets for any substrate group of interest from ChEBI ontology. Next, we build two different dataset series using this pipeline: TooT-ICAT and TooT-SPEC (Section 3.5). These datasets are used in our preliminary experiments using transfer learning from the ProtBERT language model and the results are presented (Section 3.6 and Section 3.7).

3.2 Datasets for Substrate Specificity

The lack of research in predicting specific substrates in transport activities results from the absence of a dataset of transport protein sequences with their specific substrate. Here, we describe an automated pipeline used to build a dataset of transport proteins as the data samples which are labeled with their specific transported substrate. Since this pipeline relies on the ChEBI and GO term annotations, it can be used to construct various datasets for proteins and their associated substrates. It is not limited to building datasets solely for transport proteins, but could possibly generate datasets for other domains such as metabolic and biosynthetic processes, as well as protein-ligand binding, and more.

Here, we apply the pipeline to build two separate groups of datasets: ICAT and

SPEC. TooT-ICAT datasets cover Inorganic Cation and Anion Transporters (ICAT) and the TooT-SPEC datasets cover a wider range of the transport proteins and their Specific (SPEC) carried substrate. These datasets are built to include all the transport proteins available not only a specified group of the carried substrate. The pipeline is described in the Section 3.4. The following sections describe ChEBI and GO annotation systems as two resources required for this pipeline.

3.2.1 Chemical Entities of Biological Interest (ChEBI)

Definition 3.2.1. An *ontology* is a structured framework that defines a set of concepts, their attributes, and the relationships between them within a specific domain, enabling consistent data interpretation and integration.

ChEBI is a publicly available database and ontology containing information about molecular entities of biological interest focused on small chemical compounds. This database currently includes over 46,000 molecular entities. A molecular entity is defined as: “Any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, identifiable as a separately distinguishable entity” in this annotation [DDME⁺07]. The term *biological interest* elaborates that these molecular entities are either products of nature or synthetic products involved in the processes of living organisms [HOD⁺16].

ChEBI annotation system incorporates an ontological classification, whereby the relationships between molecular entities or classes of entities are specified with parents and/or children hierarchies. The main ontology relationships are the “is_a” relationship for classification, the “has_role” relationship which links chemical entities to their roles, and “has_part” which links composite entities, e.g. salts, to their component parts [HOD⁺16]. Here, we utilize the “is_a” relationship between components. Figure 6 presents an example of ChEBI ontology hierarchy.

3.2.2 Gene Ontology (GO)

Gene Ontology is a comprehensive framework providing a standardized vocabulary to describe the roles of genes and their products across different species. This knowledge is described in three main domains: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) [ABB⁺00]. Each domain is described as follows: Biological Process (BP) describes large processes, or biological programs accomplished by multiple

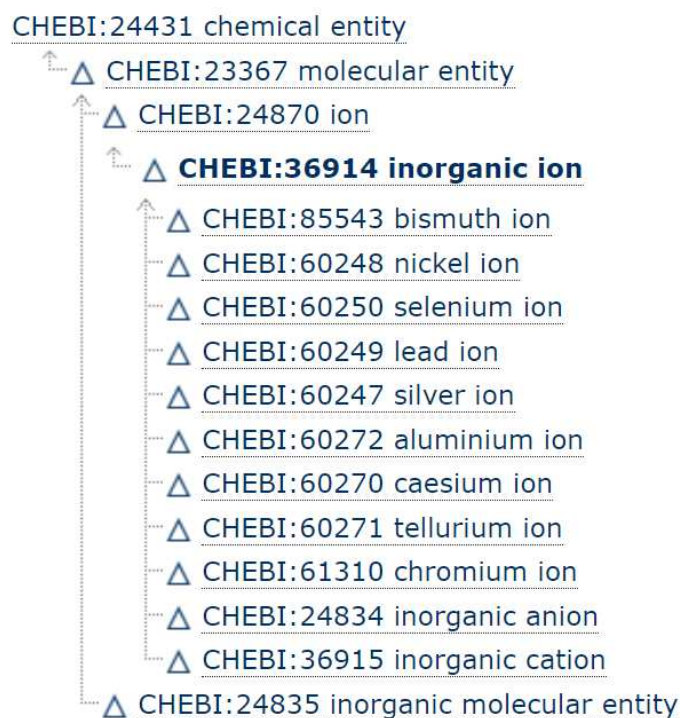


Figure 6: Example of ChEBI ontology hierarchy for inorganic ion. This figure shows the ChEBI term CHEBI:36914- inorganic ion as a descendant of the term CHEBI:24870-ion, in the ChEBI ontology hierarchy

molecular activities, for example, DNA repair or signal transduction. Cellular Component (CC) describes the locations in which a gene product performs a function which may be either cellular compartments (e.g. mitochondrion) or stable macromolecular complexes of which they are parts (e.g. the ribosome). Molecular Function (MF) describes the Molecular-level activities performed by gene products such as catalysis or transport. GO MF terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place [The21a]. Figure 7 presents an example of each domain in Gene Ontology.

3.3 GO and ChEBI cross-reference

Gene Ontology (GO) is designed to describe the action of genes and gene products in a biological context. As a matter of fact, many chemical entities are involved in the

actions described by GO. Therefore, to study many aspects of these actions, a cross-reference semantic relationship from GO terms and their associated chemical entities, e.g. in ChEBI, is required. This semantic relationship has been developed in a collaborative effort between GO and ChEBI ontology developers. The result is a Web Ontology Language (OWL) format file, GO-Plus.owl, ensuring that the representation of chemicals in the GO is both internally consistent and in alignment with the chemical expertise captured in ChEBI [HAB⁺13, ABB⁺00, The21a]. An example GO term in this file is presented in Figure 8. This file also includes cross-references to other related ontologies such as Cell Ontology and Uberon.

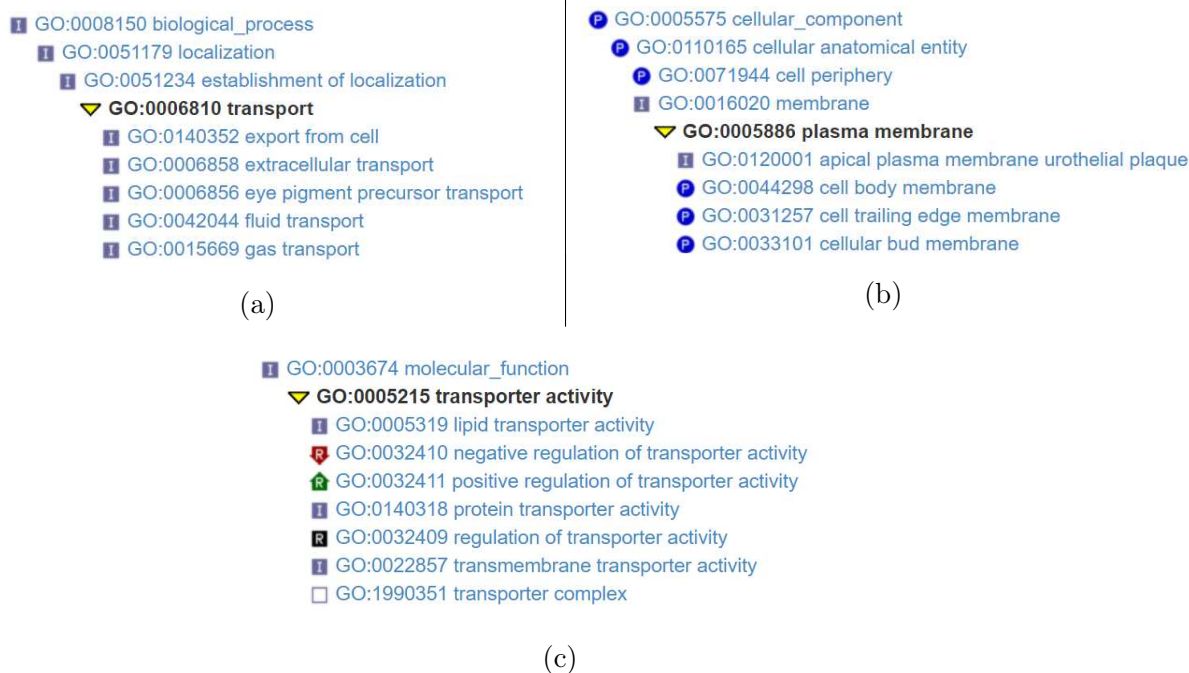


Figure 7: Gene Ontology hierarchy for membrane transport in 3 domains. The figure shows GO terms (a) **GO:0006810-transport** as a biological process, (b) **GO:0005886-plasma membrane** as a cellular component, and (c) **GO:0005215-transporter activity** as a molecular function entity with their ancestors and descendants in the hierarchy.

3.4 SGI Dataset Construction Pipeline

We developed an automated method to construct a dataset of protein sequences labeled with their specific interacting substrates. This tool called the Substrate Group of Interest

```

<owl:Class rdf:about="http://purl.obolibrary.org/obo/GO_0015250">
  <owl:equivalentClass>
    <owl:Class>
      <owl:intersectionOf rdf:parseType="Collection">
        <rdf:Description rdf:about="http://purl.obolibrary.org/obo/GO_0015267"/>
        <owl:Restriction>
          <owl:onProperty rdf:resource="http://purl.obolibrary.org/obo/RO_0002313"/>
          <owl:someValuesFrom rdf:resource="http://purl.obolibrary.org/obo/CHEBI_15377"/>
        </owl:Restriction>
      </owl:intersectionOf>
    </owl:Class>
  </owl:equivalentClass>
</owl:Class>

```

Figure 8: An example of GO and ChEBI cross-reference from GO-Plus.owl file. The figure shows GO term GO:00015250- water channel activity in GO-Plus.owl file. As the figure presents this term is cross-referenced to the ChEBI term CHEBI:15377-water.

(SGI) dataset generator, is designed to build a dataset of proteins that interact with a user-specified group of substrates from ChEBI. The protein sequences are sourced from the Universal Protein Resource (UniProt) [The21b], and the specific chemical substrates (i.e., labels of the protein samples in the dataset) are identified using the ChEBI (Chemical Entities of Biological Interest) ontology [HOD⁺16].

The main challenge in this task is that UniProt sequences are not directly labeled with ChEBI terms; instead, they are associated with Gene Ontology (GO) terms. Following the approach in [AB19] for building ontology-based benchmark datasets, we rely on GO labels to establish a connection between the protein sequences and the substrates. This connection is made using the cross-ontology relationship between GO terms and ChEBI terms in the “GO-Plus.owl” file, as described in Section 3.3 [HAB⁺13]. To use this cross-ontology for labeling the protein sequences, we first need to create a map from ChEBI terms to their cross-referenced GO terms. This map is constructed by retrieving the relevant ChEBI-GO entries associated with the Substrate Group of Interest (SGI) from the “GO-Plus.owl” file.

This map is called C2GO in this document and has ChEBI terms as the range and Gene Ontology terms as the domain. This is a one-to-many map since one ChEBI term could be related to multiple GO terms in the molecular activity domain. Consequently, the ChEBI-GO map entries in the map are in the format of one ChEBI term mapped to a set of GO terms. For example, the ChEBI term for fluoride CHEBI:17051 is related to both GO terms GO:0062054-fluoride channel activity and GO:1903425-fluoride

transmembrane transporter activity. Therefore the map entry is CHEBI:17051 $\mapsto$ {GO:0062054, GO:1903425}.

The scope of transport proteins in the dataset is determined by a user input called Substrate Group of Interest (SGI). This is a list of ChEBI terms that will be used as classes in the dataset. The SGI can be a general term, such as CHEBI:25213-metal cation, which generates a dataset with multiple classes of specific substrates rooted at metal cation in ChEBI ontology, for example, CHEBI:49807-lead(2+), CHEBI:39099-calcium(1+), and so on. Additionally, since the SGI is a list of ChEBI terms the pipeline can accept multiple terms at once. For example, if the SGI is set to [CHEBI:36915-inorganic cation, CHEBI:24834-inorganic anion], the resulting dataset would have specific substrates rooted in both ChEBI terms in ChEBI ontology as the class labels.

To build substrate classes of the dataset as specific as possible, one would naturally rely on the lowest level of terms in the ChEBI “is_a” hierarchy otherwise known as the leaf terms, considering these terms as specific substrate classes. However, relying on ChEBI leaf terms, we encounter the challenge of missing multiple ChEBI-GO map entry relationships due to the special specification in ChEBI ontology. We demonstrate this challenge with the following example.

Consider building a dataset of transport proteins labeled with their specific substrates; starting from the ChEBI root, CHEBI:24431-chemical entity we retrieve all the leaf ChEBI terms from ChEBI ontology. These terms are then mapped into GO terms in the C2GO map and used to label the sequences. We expect water transporters as a significant class of transporters in our dataset. However, using this methodology resulted in capturing zero ChEBI-GO relationship terms related to water.

To investigate this issue we review the ChEBI term CHEBI:15377-water, in the ChEBI “is_a” hierarchy. As Figure 9 illustrates this term is not a leaf term in ChEBI ontology and it has six different descendants which are isotopes of the ‘H2O’ molecule. Due to this special structure of ChEBI ontology, there exist some specific substrates appearing at the higher level from the leaf in the hierarchy. On the other hand, since the isotope molecules in most cases are transported by the same protein, they are not related to the molecular function GO terms in the “GO-Plus.owl” file and only their parent substrate, CHEBI:15377-water in this example, has a cross-reference to the Gene Ontology.

Therefore, a ChEBI leaf term such as hydrogen tritium oxide is not related to any transporter activity in Gene Ontology. The actual relationship for water transporters is from ChEBI term CHEBI:15377-water to the GO term GO:0015250-water channel

activity, which is not captured since the ChEBI term is not a leaf in the ChEBI ontology. Accordingly, relying on only ChEBI leaf terms does not capture all the ChEBI-GO relationships for building the dataset.

To overcome this issue and cover more ChEBI-GO map entries in the C2GO map, we rely on both ChEBI and GO leaf terms implementing a similar procedure from the Gene Ontology hierarchy. Starting at a general GO term related to the group of substrates in the datasets, the GO leaf terms are collected and mapped to their corresponding ChEBI term. This step is explained in detail in Section 3.4.3.

In the end, this map is utilized to label the sequences in the dataset. The pairwise identity of the sequences is reduced using the CD-Hit algorithm [FNZ⁺12]. We illustrate each step of the pipeline in the following sections along with two example datasets: ICAT and SPEC as described in section 3.2. The three main steps of this pipeline are as follows.

1. Build C2GO map
 - (a) Build C2GO_C map from ChEBI leaf terms;
 - (b) Build C2GO_G map to GO leaf terms;
 - (c) Combine C2GO_C and C2GO_G to build C2GO map
 - (d) C2GO map refinements
2. Labeling protein sequences
 - (a) Retrieve the UniProt IDs with the SGI GO terms
 - (b) Label them with the corresponding ChEBI term in C2GO map
 - (c) Eliminate multi-label sequences
3. Pair-wise identity reduction using CD-Hit algorithm

Each step is explained in detail in the following sections. The code for SGI Dataset construction pipeline is available at the following link (<https://github.com/simaataei/SGI-Labeler>)

3.4.1 Algorithms

Before describing the algorithms of the pipeline, we need to demonstrate the defined terms used in the algorithms to assist in the readability. Since the algorithms work with

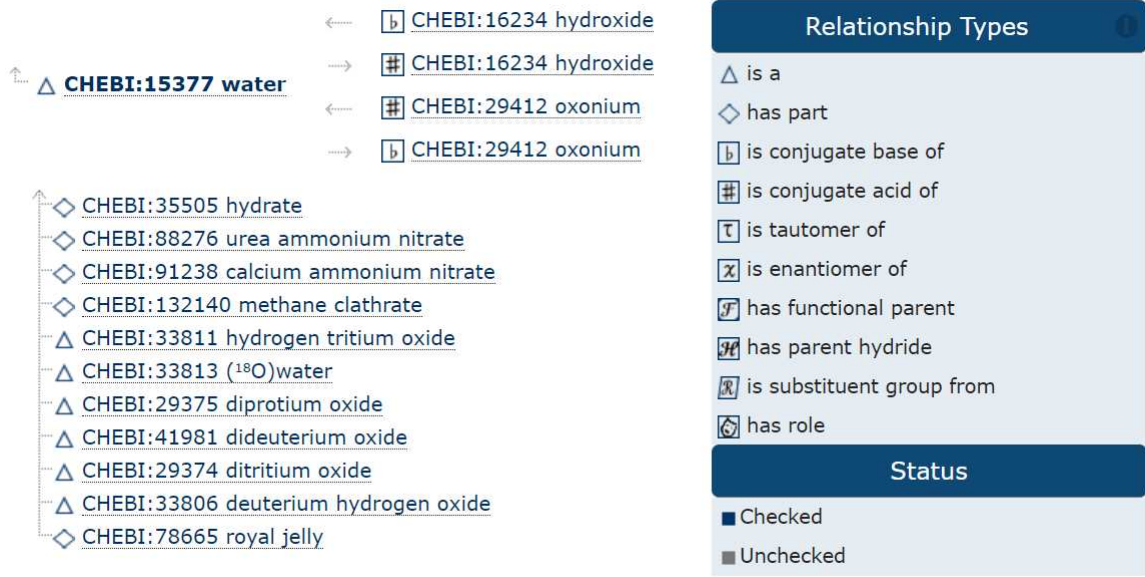


Figure 9: The ChEBI hierarchy for CHEBI:15377-water.

The figure shows the ChEBI term for water and its descendants in the ChEBI hierarchy. This term has 6 children in the “is_a” relationship graph which are the isotopes of H₂O.

ChEBI and GO ontologies, we define relationships in the ontology hierarchy used in the algorithms. First, the child relationship between two nodes of the hierarchy is defined as:

$$child(t_c, t_p) \iff t_c < is\_a > t_p \quad (22)$$

This means a node t_c is the child of node t_p if and only if there is a ‘is_a’ edge from the t_c to t_p in the given ontology. Based on the child definition we can define the parent relationship as:

$$parent(t_p, t_c) \iff child(t_c, t_p) \quad (23)$$

The two relationships of descendant, and ancestor, are defined based on *child* and *parent* relationships as follows.

$$descendant(t) = child^*(t) \quad (24)$$

$$ancestor(t) = parent^*(t) \quad (25)$$

To find the leaf terms in the ontology the relationship *is_leaf* is defined as:

$$is\_leaf(t) \iff !\exists t_c \in Ontology \text{ s.t. } child(t_c, t) \quad (26)$$

To retrieve cross-reference relationships between ChEBI and GO terms in the “GO-Plus.owl” file, any ChEBI terms assigned to a GO term or any of the ancestors of the GO term in

Gene Ontology as a ‘Restriction’ in the file is collected as follows.

$$\begin{aligned} \text{linked}(t_C, t_G) &\iff \exists t_C \in \text{ChEBI}, t_G \in \text{GO} \\ &\text{s.t. } t_C \in \text{Restrictions}(\text{ancestor}(t_G)) \end{aligned} \quad (27)$$

The following terms are defined as the common inputs of the algorithms:

- “*GO_ontology*” is the root node of the ‘*is_a*’ relationships between the Molecular Function (MF) hierarchy of Gene Ontology collected from the Gene Ontology website¹.
- “*GO_Plus*” is a file of GO and ChEBI relationship in *.OBO format collected from Gene Ontology website².
- “*GO_terms*” is a non-empty list of strings containing GO term ids in the valid format of ‘GO:xxxxxxx’ (x holding a digit) describing the molecular function of transport for terms in substrate group of interests (SGI)
- “*chebi_terms*” is a non-empty list of strings containing ChEBI term ids for ChEBI terms of substrate group of interests (SGI) in the format of ‘CHEBI:xxxxx’ (x holding a digit)
- “*chebi_ontology*” is the root node of the ‘*is_a*’ relationship ChEBI ontology collected from ChEBI website³.
- *C2GO* is a one-to-many map of ChEBI terms to sets of GO terms where each key in *C2GO* is a valid ChEBI term and the corresponding value is a set of valid Gene Ontology terms

3.4.2 C2GO_C map from CHEBI leaves

The first step of the pipeline is to build a C2GO_C map from the ChEBI leaf terms. In this step, the subset of the ChEBI substrates and their corresponding GO terms are extracted and added to a predefined empty map called “C2GO_C”. To build this map user needs to specify their Substrate Group of Interest (SGI). The pipeline accepts a list of specific Substrate Group of Interest (SGI) in ChEBI IDs format.

¹<http://purl.obolibrary.org/obo/go/go-basic.obo>

²<http://purl.obolibrary.org/obo/go/extensions/go-Plus.owl>

³<https://ftp.ebi.ac.uk/pub/databases/chebi/ontology/chebi.owl>

The algorithm retrieves the **ChEBI** term and the leaf terms of the descendants rooted on this term for the substrate group of interest from the “is_a” relationships in the **ChEBI** ontology. We use a local version of Open Biomedical Ontologies (OBO) **ChEBI** database as ‘CHEBI.OBO’ file from <https://obofoundry.org/ontology/chebi.html> accessed on March 25, 2022 [HOD⁺16]. The corresponding GO terms for **ChEBI** leaf terms with “linked” relationship are extracted and added to the *C2GO_C* map.

This pipeline allows the selection of multiple substrate groups of interest. For example, the user could search for two substrate groups of **CHEBI:24834-inorganic anion** and **CHEBI:36915-inorganic anion** as the same input. The pipeline will collect the descendants of both groups separately and merge the extracted leaves as one set. The pipeline will need the corresponding GO term for each of the substrate groups of input **ChEBI** terms in the following steps. This process is described in Algorithm 1.

Algorithm 1 *C2GO_G* map from **CHEBI** leaf terms: *C2GO_from_CHEBI*.

Require: *chebi_terms*, and *chebi_ontology*, are defined as in the preamble

Ensure: *C2GO_C* is a one-to-many map where the keys are all the terms in *chebi_terms* and the values are sets of corresponding GO terms collected from the **ChEBI** ontology and each key is a valid **ChEBI** term and the corresponding value is a set of valid GO terms

Input: *chebi_ontology*, *chebi_terms*, *GO_ontology*

Output: *C2GO_C*, a one-to-many map of items in *chebi_terms* to their corresponding GO terms from **ChEBI** ontology

```

1: function C2GO_from_CHEBI(chebi_terms, chebi_ontology, GO_ontology)
2:   C_leaves, C_descendant  $\leftarrow$  empty sets
3:   C_leaves = { $t_C \in chebi\_ontology \mid is\_leaf(t_C)$ }
4:   C_descendant = { $t_C \in chebi\_terms \mid descendant(t_C)$ }
5:   C2GO_C = {( $t_C \mapsto S$ )  $\mid t_C \in (C\_leaves \cap C\_descendant) \wedge S = \{t_G \mid t_G \in$ 
      GO_ontology  $\wedge linked(t_C, t_G)\}$ }
6:   return C2GO_C
7: end function

```

The algorithm first initializes two empty sets, *C_leaves*, and *C_descendant* to collect, the **ChEBI** ontology leaf terms, and **ChEBI** terms descendant respectively. Then with all the leaf nodes in *chebi_ontology* are collected in *C_leaves* and all the descendants of each input term in *chebi_terms* list are collected in *C_descendant*. Then the algorithm

initializes an empty one-to-many map $C2GO\_C$ to collect the ChEBI-GO relationships. It iterates over each combination of a ChEBI term t_C and a GO term t_G where t_C is a leaf node in *chebi_ontology* and in the descendant of at least one term in *chebi_terms*, and where t_C is *linked* to t_G (according to “GO-Plus.OWL” file) and adds the pair $\langle t_C, t_G \rangle$ to $C2GO\_C$. This ensures that only valid combinations are included in the map.

3.4.3 C2GO_G map from GO leaves

To cover more ChEBI – GO entries in the map, the pipeline explores GO terms related to the Substrate Group of Interest (SGI). Having the molecular function (MF) GO term associated with the membrane transport of the substrate of interest, Algorithm 2 extracts the ChEBI terms associated with this GO term and added them to a second map called $C2GO\_G$.

Algorithm 2 C2GO map from GO leaf terms: $C2GO\_from\_GO$.

Require: GO_terms , $GO\_ontology$, $chebi\_term$, and $chebi\_ontology$ are defined as in the preamble;

Ensure: $C2GO\_G$ is a one-to-many map of items in *chebi_terms* to their corresponding GO terms from the GO ontology and each key is a valid ChEBI term and the corresponding value is a set of valid Gene Ontology terms

Input: GO_terms , $GO\_ontology$, $chebi\_term$, $chebi\_ontology$

Output: $C2GO\_G$, a one-to-many map of items in *chebi_terms* to their corresponding GO terms from Gene Ontology

```

1: function    $C2GO\_from\_GO(GO\_terms, \quad GO\_ontology, \quad CHEBI\_terms,$ 
    $CHEBI\_ontology)$ 
2:    $GO\_leaves, GO\_descendant, CHEBI\_descendant \leftarrow$  empty sets
3:    $GO\_leaves = \{t_G \in GO\_ontology \mid is\_leaf(t_G)\}$ 
4:    $GO\_descendant = \{ t_G \in GO\_terms \mid descendant(t_G) \}$ 
5:    $CHEBI\_descendant = \{ t_C \in CHEBI\_terms \mid descendant(t_C) \}$ 
6:    $C2GO\_G = \{(t_C \mapsto S) \mid t_C \in CHEBI\_descendant \wedge S = \{t_G \mid t_G \in (GO\_leaves$ 
    $\cap GO\_descendant) \wedge linked(t_C, t_G) \} \}$ 
7:   return  $C2GO\_G$ 
8: end function

```

The first step is to create two empty sets, GO_leaves , and $GO\_descendant$, to store the GO leaf terms in the Gene Ontology and the descendant terms of the GO_terms

according to gene ontology, respectively. Next, the algorithm fills the *GO_leaves* set with all the leaf terms in the GO ontology. This is done by iterating over all terms t_g in *GO_ontology* and adding them to *GO_leaves* if they are leaves. Iterating over all terms t_g in *GO_terms* the algorithm adds their descendant to *GO_descendant* according to the defined relationship in equation 24.

Finally, the algorithm builds the *C2GO_G* map. This is done using a set comprehension that iterates over all pairs of terms $\langle t_C, t_G \rangle$ where t_C is a term in the ChEBI ontology and t_G is a term in the GO ontology. The condition for adding a pair to the map is that there is a *linked* relationship, as described in the Section 3.4.1, between t_C and t_G according to GO-Plus.OWL file and that t_G is a leaf in the descendants of the GO terms in the GO ontology. The last condition “ $t_C \in CHEBI\_descendant$ ” is preventing the algorithm from collecting ChEBI terms unrelated to the SGI. The map value for each t_C key is a set containing the corresponding t_G terms.

3.4.4 Build C2GO map

The *C2GO* map is built based on both *C2GO_C* and *C2GO_G* maps. This map contains both terms from ChEBI term leaves and term leaves covering all possible relationships mentioned in “GO-Plus.OWL” file for the Substrate Group of Interests. This map is a simple union of the two maps with ChEBI-GO one-to-many map entries. This procedure is described in Algorithm 3.

The algorithm then calls two functions, *C2GO_from_CHEBI* and *C2GO_from_GO*, which both return a one-to-many map of ChEBI terms to sets of GO terms. These functions are introduced and explained in Algorithm 1, Algorithm 2. The algorithm then merges the two maps by performing a set union between the values associated with each key in *C2GO_G* and *C2GO_C* and stores the result in *C2GO*.

3.4.5 Map Refinement

The differences in specifications in the two ontologies of ChEBI and GO, result in the relationship of leaf terms to non-leaf terms on both ChEBI and GO. This means that according to the “GO-Plus.OWL” file, a leaf ChEBI term could be mapped to non-leaf GO terms and the other way around. An example of this issue is explained in Section 3.4 mentioning the differences between the two annotation systems related to **water transporters**. Since

Algorithm 3 Build C2GO map.

Require: *chebi_terms*, *chebi_ontology*, *GO_terms*, *GO_ontology* are defined as in preamble.

The *C2GO_from_CHEBI* function is defined and returns a one-to-many map of the **ChEBI** terms to sets of GO terms collected from **CHEBI**

The *C2GO_from_GO* function is defined and returns a one-to-many map of the **ChEBI** terms to sets of GO terms collected from Gene Ontology

Ensure: *C2GO* is a one-to-many map of **ChEBI** terms to sets of Gene Ontology terms collected from both the Gene Ontology and **ChEBI** ; Each key is a valid **ChEBI** term and the corresponding value is a set of valid Gene Ontology terms

Input: *chebi_ontology*, *chebi_terms*, *GO_terms*, *GO_ontology*

Output: *C2GO*, a one-to-many map of the **ChEBI** terms to sets of GO terms collected from both Gene Ontology and **ChEBI**

- 1: $C2GO\_C = C2GO\_FROM\_CHEBI(chebi\_terms, chebi\_ontology, GO\_ontology)$
 - 2: $C2GO\_G = C2GO\_FROM\_GO(GO\_terms, GO\_ontology, chebi\_term, chebi\_ontology)$
 - 3: $C2GO = C2GO\_C \cup C2GO\_G$
-

the objective of the pipeline is to build a dataset with fine-grained classes, the next algorithms are designed to remove terms associated with a group of chemical substrates rather than specific substrates themselves. Although Algorithm 2 collects additional **ChEBI**-GO map entries it is prone to collecting broad **ChEBI** terms that require refinement. In this algorithm, the leaf terms in the Gene Ontology annotation system are not necessarily mapped to the leaf terms in **ChEBI** annotation systems. Therefore relying on leaf GO terms in Algorithm 2, might result in collecting non-leaf **ChEBI** terms. The refinements are implemented to have classes of **ChEBI** terms as close as possible to the leaf level in **ChEBI** ontology.

For example, consider the **ChEBI** term **CHEBI:25213- metal cation**, Figure 10 shows this term is not a **ChEBI** leaf term however it has been entered to the map from Algorithm 2. Algorithm 4 and Algorithm 5 refine the map to eliminate this term since there exists another descendant of this term, e.g. **calcium(2+)** in this case, in the *C2GO* map which has been collected by Algorithm 1.

The map refinement is implemented in two main steps. First, parent **ChEBI** terms are removed from the map. Considering the map to be a specific substrate labeling, we

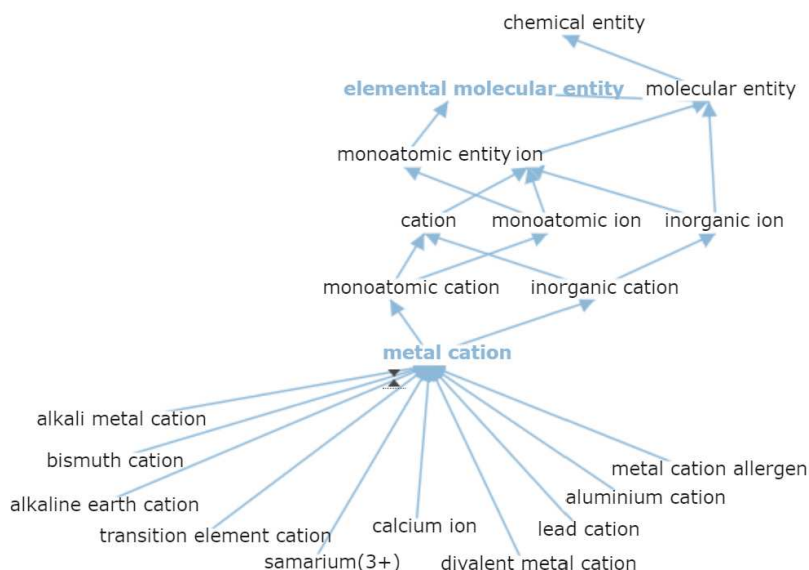


Figure 10: The ChEBI term for metal cation with its ancestors, and descendants. The figure shows the relationships of the term **metal cation** with its ancestors and descendants in the ChEBI ontology hierarchy.

eliminate the CHEBI terms that have children in the map. Therefore, for all CHEBI terms in the C2GO keys, if children of the CHEBI term are present in the C2GO map, the CHEBI term itself is removed from the map. This process is described in Algorithm 4.

Second, the GO terms associated with multiple ChEBI terms are refined. According to Gene Ontology annotations, there are GO terms with more than one CHEBI term associated with them. These groups of GO terms are the molecular functions including more than one chemical substrate. To improve the accuracy of the map, the pipeline removes non-leaf CHEBI terms from GO terms with multiple CHEBI maps. In other words, for GO terms with more than one CHEBI term, non-leaf CHEBI terms are deleted. This process has been described in Algorithm 5.

Algorithm 4 iterates over each ChEBI term t_c in the input *C2GO* map, and for each term, it checks if there exists a parent ChEBI term t_p such that t_c is a descendant of t_p and the entry for t_p in the *C2GO* is not empty, $C2GO(t_p) \neq \emptyset$. If such a parent term t_p and its child term t_c both exist in the *C2GO* map, the parent, t_p , should be filtered out from the map. Otherwise, t_p is added to the refined map with its corresponding GO terms.

Algorithm 4 Map refinement from chebi.

Require: *chebi_ontology*, *chebi_terms*, and *C2GO* are defined as in the preamble.

Ensure: *C2GO* is a filtered map that only contains **ChEBI** terms for which have no child **ChEBI** term in the *C2GO* map;

Input: *chebi_ontology*, and *C2GO* as the output of algorithm 3;

Output: *C2GO*, a refined map of *C2GO*

- 1: $C2GO = \{(t_c \mapsto S) \mid t_c \in \text{ChEBI} \wedge S = \{t_G \mid t_G \in GO \wedge \neg \exists t_p \in \text{ChEBI} \text{ s.t. } t_c \in \text{descendant}(t_p) \wedge C2GO(t_p) \neq \emptyset\}\}$
-

Algorithm 5 Map Refinement from GO.

Require: *chebi_ontology*, *chebi_terms*, and *C2GO* are defined as in the preamble.

Ensure: The algorithm returns a filtered C2GO map with only leaf **ChEBI** terms for GO terms with multiple **ChEBI** terms

Input: *chebi_ontology*, *chebi_terms*, and *C2GO* as the output of algorithm 4;

Output: *C2GO* filtered map

- 1: $GO2C \leftarrow$ empty one-to-many map
 - 2: $GO2C = C2GO^{-1}$
 - 3: $GO2C = \{t_G \mapsto S \mid t_G \in GO \wedge S = \{t_c \mid t_c \in \text{ChEBI} \wedge \neg \exists t_p \in \text{ChEBI} \text{ s.t. } t_c \in \text{descendant}(t_p) \wedge GO2C(t_p) \neq \emptyset\}\}$
 - 4: $C2GO = GO2C^{-1}$
-

Algorithm 5 first creates an empty one-to-many map called *GO2C* to store the reverse mapping of *C2GO* from GO terms to **ChEBI** terms. Then, it iterates over t_G in GO terms and for GO terms with more than one **ChEBI** term, if there exists a t_c which is a descendant of another term t_p in the GO same entry, the parent term t_p is filtered out from the $\{t_c\}$ set in map entry. Finally, the algorithm creates the refined *C2GO* map by reversing the *GO2C* map back. The result C2GO map does not contain the same GO entry in two **ChEBI** terms that share a descendant relationship based on the **ChEBI** ontology.

Algorithm 6 is an optional algorithm to specify any GO terms that the user needs to be excluded from the map. Due to the structure of Gene Ontology, there are terms that are involved in more than one process. These terms might result in collecting C2GO entries and consequently labeling protein sequences that are out of the scope of the study. An example of this issue has been presented in Section 3.5 with GO terms related to **catalytic activity** and **transporter activity**. This algorithm is not essential in the pipeline and could be disregarded if not necessary.

Algorithm 6 Term exclusion.

Require: $GO\_ontology$, $C2GO$ are defined as in the preamble.

$exclude\_term$ is the GO term to be excluded from the map in the format of ‘GO:xxxxxxx’ (x holding a digit)

Ensure: The algorithm returns a filtered C2GO map with GO terms not in $exclude\_term$ or any of its descendants in GO ontology

Input: $GO\_ontology$, $C2GO$ as the output of algorithm 5, and $exclude\_term$ as a GO term to be excluded from the C2GO map

Output: $C2GO$ filtered map

1: $C2GO = \{t_C \mapsto S \mid \forall t_C \in \text{ChEBI}, C2GO(t_C) = C2GO(t_C) \setminus \text{descendant}(exclude\_term)\}$

Algorithm 6 takes C2GO as an input map and $exclude\_term$ which is a GO term that the user needs to be excluded from the map. The goal of the algorithm is that the GO term $exclude\_term$ and all its descendants are removed from the entries of C2GO. To accomplish this, the algorithm iterates over each entry in C2GO and checks if the value set contains any GO term that is a descendant of $exclude\_term$. If so, that entry is excluded from the filtered map. The resulting map contains only those entries from the original C2GO map that do not contain any GO term that is a descendant of $exclude\_term$.

3.4.6 Labeling Data

The final step is to label protein sequences according to the produced C2GO map. This step requires protein samples and their associated GO terms. Here, we collect the data from UniProt-KB, version 2023_2, accessed on August 16, 2023 [The23]. To collect the data from UniProt the GO term for SGI is queried.

For each sequence in the dataset, the pipeline labels the sequences with the substrate name of a ChEBI term if the sequences have common GO terms with the corresponding ChEBI term from the C2GO map. For example, having the C2GO map entry: $\text{CHEBI:17632} \mapsto \{\text{GO:0010542}\}$ in the map, any protein from the dataset with GO term GO:0010542 is labeled with `nitrate`, the substrate name of the ChEBI term CHEBI:17632.

This pipeline could produce both single-label and multi-label datasets; this parameter is selected by the user in the *multi_label* variable as the input of the Algorithm 7. If *multi_label* is set to *True* the pipeline produces a dataset of protein sequences with multiple ChEBI substrates as labels and if it is set to *False*, the final dataset only includes

Algorithm 7 Labeling Data.

Require: *C2GO* is defined as in the preamble.

input_data is a set of protein sequences with corresponding GO terms.

multi_label is a variable determining if the output dataset is multi-labeled if the value is *True*, or single-labeled if the value is *False*.

Ensure: Returns a set of protein sequences from *input_data* labeled with the term(s) of their corresponding specific substrate from **ChEBI**

Input: *input_data*, *C2GO* as the output of Algorithm 5 (or 6), and *multi_label*

Output: *labeled_data*

```
1: labeled_data =  $\{p \mapsto L \mid p \in \text{input\_data} \wedge L = \{t_c \mid t_c \in \text{ChEBI} \wedge \exists t_G \in \text{input\_data}[p].\text{GO\_label} \wedge t_G \in C2GO(t_C)\} \wedge L \neq \emptyset\}$ 
2: if  $\neg \text{multi\_label}$  then
3:   labeled_data = labeled_data  $\setminus \{(p \mapsto L) \in \text{labeled\_data} \mid |L| > 1\}$ 
4: end if
```

protein sequences with one label from **ChEBI**. This means that the algorithm eliminates proteins with more than one label associated with them from the dataset. This setting is flexible to change and the pipeline is able to produce multi-label datasets with samples with more than one label.

3.4.7 Identity Reduction

To discuss the role of sequence homology in the experiments, the pipeline is able to run a CD-Hit clustering algorithm [HNG⁺10] to cluster sequences based on a user-provided pairwise sequence identity level; randomly sampling on sequence from each cluster the method assures a lower level of pairwise sequence identity from the user-specified level in the final dataset.

3.5 Example Datasets

3.5.1 TooT-ICAT Dataset series

The first example is the ICAT dataset series built based on inorganic cation and inorganic anion substrates. The process of building these dataset series is depicted in Figure 17. We apply the pipeline to the **ChEBI** terms:

- CHEBI:36915-Inorganic cation,
- CHEBI:24834-Inorganic anion

and the GO terms

- GO:0022890-Inorganic cation transmembrane transporter activity
- GO:0015103-Inorganic anion transmembrane transporter activity

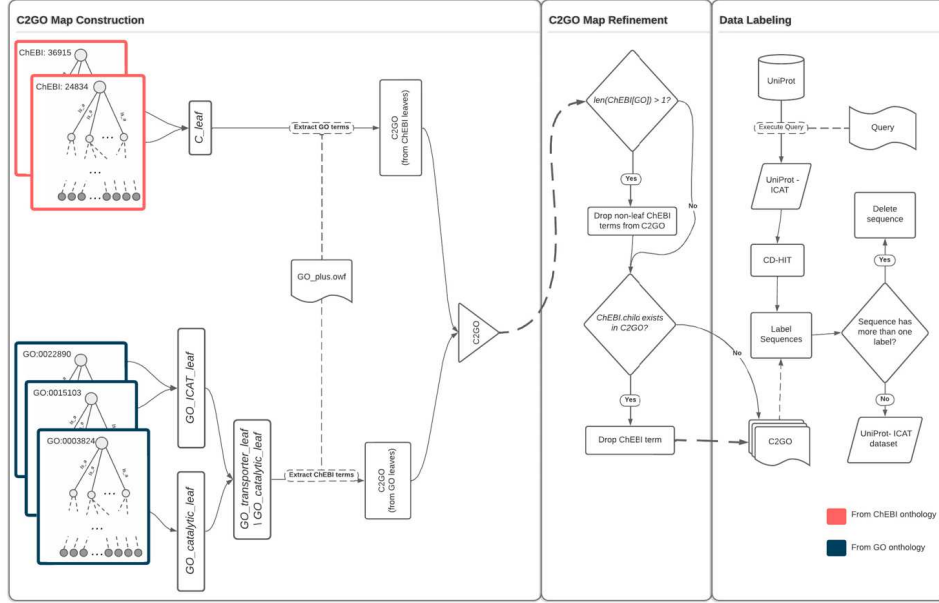


Figure 11: Detailed process for extracting TooT-ICAT dataset series.

The process of constructing the ICAT datasets: C2GO map Construction from Gene Ontology and ChEBI annotation, based on algorithms 1, 2, and 3 (left). The C2GO map is refined in two steps as algorithm 4, and 5 describes (middle). The data is collected from UniProt and the data labeling is performed based on the extracted C2GO map based on algorithm 7 (right)

The Algorithm 1, collects all ChEBI terms in the descendants of both terms CHEBI:36915 -inorganic cation, and CHEBI:24834-inorganic anion in the variable $C\_descendant$. Then all the leaf terms of ChEBI ontology are collected in the set C_leaves . Next, the corresponding GO terms to each ChEBI term in the intersection of $C\_descendant$ and C_leaves . Each ChEBI term and its corresponding GO term is collected to build the initial ChEBI to GO map, $C2GO\_C$. The intermediate results from the output of this algorithm are presented in Table 2.

Next, in Algorithm 2, the leaf terms of Gene Ontology are collected in a set GO_leaves . Then, the descendants of GO terms of SGI, GO:0022890-inorganic cation transmembrane

transporter activity and GO:0015103-inorganic anion transmembrane transporter activity are retrieved from the Gene Ontology in *GO_descendant*. Then the *linked* ChEBI terms to each *GO* term in the intersection of *GO_leaves* and *GO_descendant* is collected and the entry of $t_c \mapsto t_g$ is added to the new ChEBI to GO map *C2GO_G*. The output results of this algorithm are presented in Table 3. In the following steps, Algorithm 3 builds the *C2GO* map combining two maps of *C2GO_C* and *C2GO_G* and Algorithm 4, Algorithm 5 refine the map as explained.

In the *C2GO* map, we noticed the existence of GO terms related to **catalytic activity**. To investigate the reason for their inclusion in the *C2GO* we reviewed the definition of the term in Gene Ontology. Catalytic or enzyme activity according to Gene Ontology’s definition is “*Catalysis of a biochemical reaction at physiological temperatures. In biologically catalyzed reactions, the reactants are known as substrates, and the catalysts are naturally occurring macromolecular substances known as enzymes. Enzymes possess specific binding sites for substrates and are usually composed wholly or largely of protein, but RNA that has catalytic activity (ribozyme) is often also regarded as enzymatic.*”

The enzymes which catalyze the movement of ions or molecules across membranes or their separation within membranes are called translocases. They have been recently added to the Enzyme Commission(EC) system under the class “*EC 7*” in 2018 [MT21]. Accordingly, a group of catalytic activities in Gene Ontology is involved in translocation-type transport mechanisms for the uptake of a variety of nucleotide precursors happening in the membrane. These activities are categorized under both catalytic activity and transmembrane transport activity. The intersection of two GO terms, GO:0022857-transmembrane transporter activity and GO:0003824-catalytic activity includes GO terms of this kind. An example of this intersection is presented in Figure 12.

Studying enzymes and catalytic activity requires different approaches from transmembrane transport proteins and it does not fall into the scope of this research. Therefore, the GO terms in the intersection of **catalytic activity** and **transmembrane transport activity** are eliminated from the *C2GO* map. To do so, the GO term GO:0003824-catalytic activity and its descendant are retrieved from Gene Ontology and excluded from the *C2GO* map. In the TooT-ICAT *C2GO* map, the intersection of **catalytic activity** and **transmembrane transport activity** included the five GO terms presented in Table 4.

The final step is to label protein sequences according to the produced *C2GO* map. In this step we download the proteins with transmembrane transporter activity and the

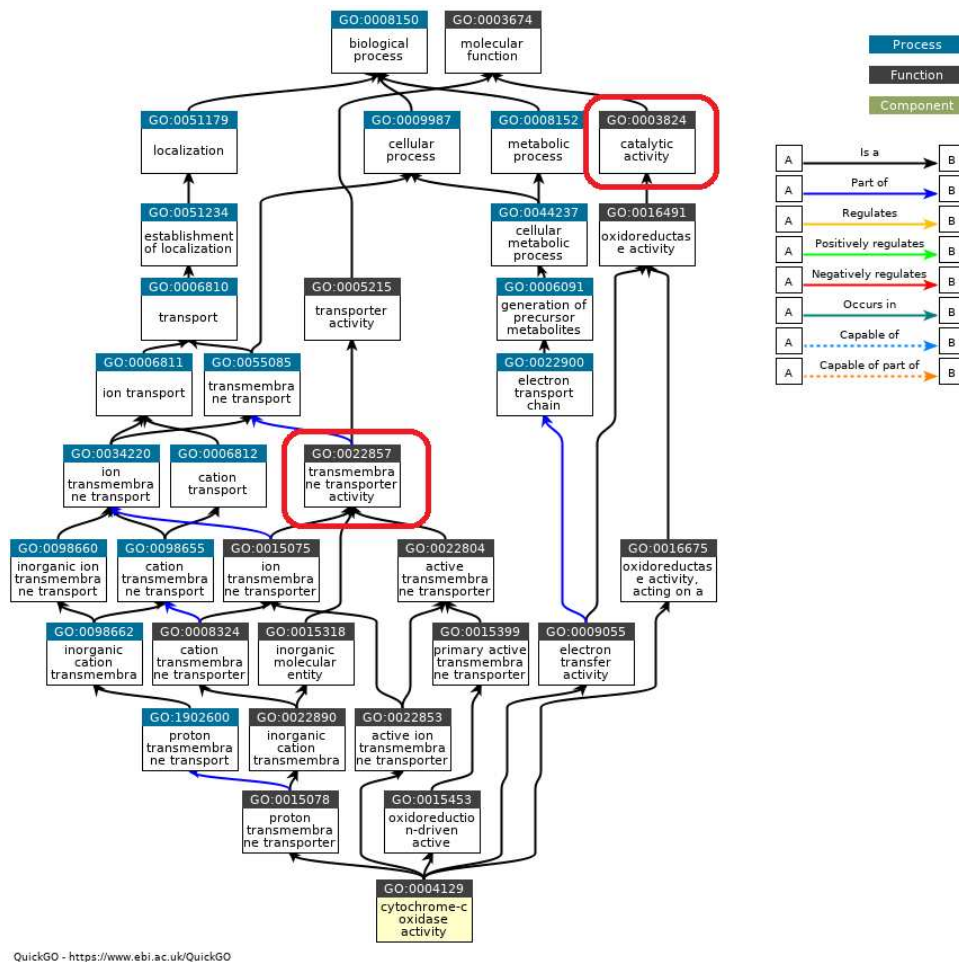


Figure 12: The ancestor chart for GO term GO:0004129 from Gene Ontology. This figure shows an example of the intersection of transmembrane transporter activity and catalytic activity. This term is related to both GO:0022857 and GO:0003824, marked in red.

evidence at the protein level from UniProt with the following query:

$\langle goa:("transmembrane\ transporter\ activity\ [22857]")\ existence:\ "Evidence\ at\ the\ protein\ level\ [1]" \rangle$

and Swiss-Prot with:

$\langle goa:("transmembrane\ transporter\ activity\ [22857]")\ existence:\ "Evidence\ at\ protein\ level\ [1]" reviewed:yes \rangle$

For each sequence in the UniProt (or SwissProt) dataset, we label the sequences with the ChEBI term if the sequences have common GO terms with the corresponding ChEBI term from the C2GO map. Finally, we remove sequences with more than one label or sequences without any label.

The pairwise sequence identity is set to two different numbers. First, we extract a dataset with 100% pairwise identity to exclude exactly identical sequences from the dataset. Next, a dataset with 60% pairwise identity is built to only include protein sequences with less than 60% pairwise identity. At last, a threshold of a minimum of 10 sequences in each class is applied and classes with less than 10 samples are excluded from the dataset.

By setting two levels of pairwise identity level on UniProt and SwissProt, we create four datasets of inorganic cation and anion transporters:

1. UniProt-ICAT-100, the pipeline applied to UniProt with CD-HIT at 100 percent identity for sequences;
2. Swiss-Prot-ICAT-100, the pipeline applied to Swiss-Prot with CD-HIT at 100 percent identity for sequences;
3. UniProt-ICAT-60, the pipeline applied to UniProt with CD-HIT at 60 percent identity for sequences; and
4. Swiss-Prot-ICAT-60, the pipeline applied to Swiss-Prot with CD-HIT at 60 percent identity for sequences.

Details of these 4 datasets are presented in Table 5 and Table 6.

3.5.2 TooT-SPEC Dataset series

This study intends to predict all of the specific substrates carried by any transmembrane protein, therefore we selected to start at the root node of ChEBI ontology to cover all of the existing substrates. Accordingly, the term

- CHEBI:24431-chemical entity

from ChEBI and GO term

- GO:0022857-transmembrane transporter activity (MF)

are selected as the inputs to the pipeline.

Similar to TooT-ICAT dataset construction, the Algorithm 1 collects all the ChEBI terms in the descendants of CHEBI:24431-chemical entity in the variable $C\_descendant$, which in this case results in all the nodes in ChEBI ontology. Then the leaf terms of ChEBI ontology are collected as C_leaves and the ChEBI their intersection with $C\_descendant$ is retrieved, which in this case equals the set C_leaves itself. The corresponding GO terms with a *linked* relationship to any of the ChEBI terms in C_leaves are collected from the “GO-Plus.OWL” file and added to the $C2GO\_C$ map.

Algorithm 2 collects the leaf terms in the descendants of the GO term of the SGI GO:0022857-transmembrane transporter activity (MF) and collects the related ChEBI terms to these from *linked* relationships in “GO-Plus.OWL” file. The map entries of ChEBI term to a set of GO terms are added to the $C2GO\_G$ map. Then, Algorithm 3 builds the final map of $C2GO$ combining two maps of $C2GO\_C$ and $C2GO\_G$. At last, the Algorithm 4 and Algorithm 5 modified the $C2GO$ map in order to have specific classes of substrates.

As explained in the TooT-ICAT dataset construction, according to Gene Ontology, catalytic activity has an intersection with transmembrane transporter activity. However, since this type of activity is not a matter of interest in this study, catalytic activity related GO terms are removed from the $C2GO$ map. Similar to TooT-ICAT dataset construction, the GO terms in the descendants of GO:0003824-catalytic activity have been eliminated from the GO leaf set. Here, 41 GO terms are eliminated from the map as presented in Table 3.5.2.

GO term	Name
GO:0090586	protein-phosphocysteine-N-acetylglucosamine phosphotransferase system transporter activity
GO:0090587	protein-phosphocysteine-glucosamine phosphotransferase system transporter activity
GO:0090588	protein-phosphocysteine-N-acetylmuramate phosphotransferase system transporter activity
GO:0090589	protein-phosphocysteine-trehalose phosphotransferase system transporter activity
GO:0090590	protein-N(PI)-phosphohistidine-D-glucosamine phosphotransferase system transporter activity
GO:0090591	protein-N(PI)-phosphohistidine-N-acetyl-mannosamine phosphotransferase system transporter activity
GO:0090562	protein-N(PI)-phosphohistidine-N,N'-diacetylchitobiose phosphotransferase system transporter activity
GO:0090564	protein-phosphocysteine-glucose phosphotransferase system transporter activity
GO:0090565	protein-phosphocysteine-mannitol phosphotransferase system transporter activity
GO:0090566	protein-phosphocysteine-N,N'-diacetylchitobiose phosphotransferase system transporter activity
GO:0090581	protein-phosphocysteine-mannosylglycerate-phosphotransferase system transporter activity
GO:0090582	protein-phosphocysteine-D-fructose-phosphotransferase system transporter activity
GO:0090583	protein-phosphocysteine-D-sorbitol-phosphotransferase system transporter activity
GO:0090584	protein-phosphocysteine-galactitol-phosphotransferase system transporter activity
GO:0090585	protein-phosphocysteine-L-ascorbate-phosphotransferase system transporter activity
GO:0018801	glutaconyl-CoA decarboxylase activity
GO:0022856	protein-N(PI)-phosphohistidine-sorbitol phosphotransferase system transporter activity
GO:0022855	protein-N(PI)-phosphohistidine-glucose phosphotransferase system transporter activity
GO:0022870	protein-N(PI)-phosphohistidine-mannose phosphotransferase system transporter activity
GO:0022871	protein-N(PI)-phosphohistidine-sorbose phosphotransferase system transporter activity
GO:0022872	protein-N(PI)-phosphohistidine-mannitol phosphotransferase system transmembrane transporter activity
GO:0022869	protein-N(PI)-phosphohistidine-lactose phosphotransferase system transporter activity
GO:0022873	protein-N(PI)-phosphohistidine-maltose phosphotransferase system transporter activity
GO:0022874	protein-N(PI)-phosphohistidine-cellobiose phosphotransferase system transporter activity
GO:0022875	protein-N(PI)-phosphohistidine-galactitol phosphotransferase system transmembrane transporter activity
GO:0022876	protein-N(PI)-phosphohistidine-galactosamine phosphotransferase system transporter activity
GO:0022877	protein-N(PI)-phosphohistidine-fructose phosphotransferase system transporter activity
GO:0022878	protein-N(PI)-phosphohistidine-sucrose phosphotransferase system transporter activity
GO:0022879	protein-N(PI)-phosphohistidine-trehalose phosphotransferase system transporter activity
GO:0022880	protein-N(PI)-phosphohistidine-N-acetylglucosamine phosphotransferase system transporter activity
GO:0022881	protein-N(PI)-phosphohistidine-N-acetylgalactosamine phosphotransferase system transporter activity
GO:0022882	protein-N(PI)-phosphohistidine-beta-glucoside phosphotransferase system transporter activity
GO:0046932	sodium-transporting ATP synthase activity, rotational mechanism
GO:0046933	proton-transporting ATP synthase activity, rotational mechanism
GO:0004129	cytochrome-c oxidase activity
GO:0003957	NAD(P)+ transhydrogenase (B-specific) activity
GO:0140571	transmembrane ascorbate ferrioreductase activity
GO:0009486	cytochrome bo3 ubiquinol oxidase activity
GO:0009496	plastoquinol-plastocyanin reductase activity
GO:0008121	ubiquinol-cytochrome-c reductase activity
GO:0008137	NADH dehydrogenase (ubiquinone) activity

The table shows the GO term ID and name for the 41 GO terms in the intersection of catalytic activity and transporter activity which have been eliminated from the C2GO map for SPEC datasets

The data of transmembrane transport proteins are collected from the UniProt server⁴, to build the dataset. To collect the proteins with transmembrane transporter activity and the evidence at the “protein level” the following query has been received:

⟨goa: “transmembrane transporter activity [22857]” existence: “Evidence at protein level [1]” ⟩

for UniProt and

⟨goa: “transmembrane transporter activity [22857]” existence: “Evidence at protein level [1]” reviewed:yes ⟩

⁴<https://www.uniprot.org/>

for SwissProt separately. The collected data is labeled using the extracted C2GO map. The proteins are labeled with the ChEBI substrate name mapped to their associated GO term. Sequences with multiple tags are eliminated from the dataset. Similar to TooT-ICAT datasets, two pairwise identity levels of 100% and 60% are applied using the CD-Hit algorithm. A threshold of 3 sequences in each class has been applied and substrate classes with less than 3 sequences are eliminated from the datasets.

Applying two levels of identity to UniProt and SwissProt data, we create four datasets:

1. UniProt-SPEC-100, the pipeline applied to UniProt with CD-HIT at 100 percent identity for sequences;
2. SwissProt-SPEC-100, the pipeline applied to Swiss-Prot with CD-HIT at 100 percent identity for sequences;
3. UniProt-SPEC-60, the pipeline applied to UniProt with CD-HIT at 60 percent identity for sequences; and
4. SwissProt-SPEC-60, the pipeline applied to Swiss-Prot with CD-HIT at 60 percent identity for sequences.

The final datasets and the substrate classes are presented in detail in Table 43.

3.6 TooT-ICAT

In this section, we describe the project of classification of inorganic cation and anion transport proteins based on their specific transported substrate. First, we describe the TooT-ICAT datasets as the dataset series constructed for this study. Second, the methodology and classification experiments are explored. At last, the final results and discussion are presented. The code and the dataset of this project are publicly available at https://github.com/simaataei/Inorganic_Ion_Predictor.

3.6.1 Dataset description

Four TooT-ICAT datasets are extracted from UniProt and SwissProt with two identity levels of 60% and 100%. A threshold of 10 sequences in each class has been applied and substrate classes with less than 10 sequences are eliminated from the datasets. Each dataset has been divided into the training set and hold-out test set with a ratio of 80:20

for training and test purposes (Table 5). The train and test split is randomly stratified. The final datasets and the substrate classes are presented in detail in Table 6.

3.6.2 Methodology

TooT-ICAT model utilizes the ProtBERT-BFD language model to classify sequences of the ICAT dataset. We implement three different approaches based on the ProtBERT-BFD language model to classify protein sequences: a frozen ProtBERT-BFD language model to extract protein representations followed by a Logistic Regression algorithm as a classifier; frozen ProtBERT-BFD representations followed by a Feed-forward Neural Network to classify sequences; and fine-tune ProtBERT-BFD model with a softmax function which classifies the ICAT dataset.

3.6.3 BERT Representation

To produce representations of protein sequences we used the ProtBERT-BFD pre-trained model publicly available on HuggingFace website [WDS⁺20]. In this approach, we freeze the ProtBERT-BFD architecture and extract encodings of proteins from the last hidden layer. For each sequence, we collect the last hidden layer representation of the first token [CLS] which contains the encodings of all the amino acids in the sequence. The maximum length of sequences has been set to 1024 as the pre-trained model and larger sequence lengths are truncated. The ProtBERT-BFD representations are passed to a Logistic Regression model for classification.

3.6.4 Logistic Regression

Logistic Regression (LR) is selected as a basic model to implement the classification [DOM02]. A multi-class classification has been implemented for 5-fold cross-validation and hold-out testing. This model is implemented using scikit-learn python package [PVG⁺11].

3.6.5 Fine-tuned BERT language model

In this method, pre-trained ProtBERT-BFD has been fine-tuned on the extracted datasets separately to update the model weights for the specified classification task. Our model includes a single-layer Feed-forward network followed by a softmax function to predict the probabilities of each class in the dataset. The configuration for the network is as

follows; Cross entropy loss function and Adam optimizer are utilized with learning rate = 5×10^{-5} for batch size 1 [KB15]. The model converges in 7 training epochs according to the loss reduction and evaluation metrics on the validation set presented in Figure 13. The implementations are applied using the Pytorch package (version 2.5) and the pre-trained ProtBERT-BFD model is collected from the HuggingFace website⁵ [WDS⁺20, PGM⁺19].

3.6.6 Cross-validation

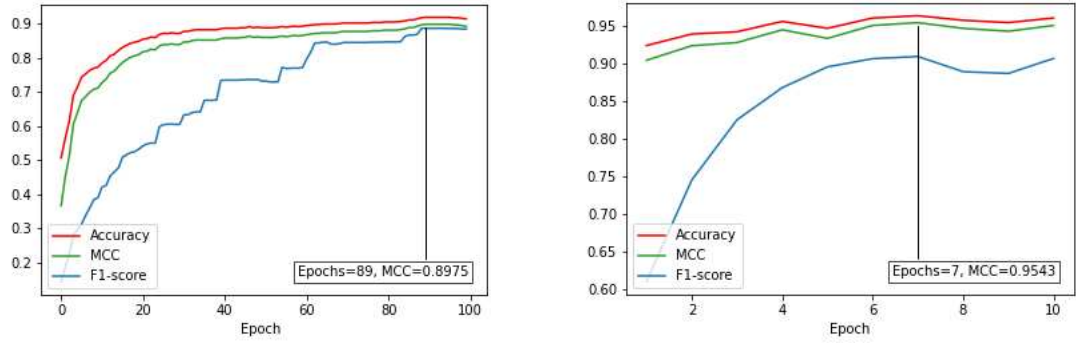
A 5-fold cross-validation has been performed on the training set for Logistic Regression classification. Then the dataset is trained on the full training set and tested on the hold-out test set. After each epoch of the training process, the other two methods are evaluated on the validation set. Then the model is trained on the full training set and tested on the hold-out test set.

3.6.7 Results and discussion

Five-fold cross-validation has been implemented on each dataset separately. As the Table 7 presents, the best results for cross-validation are from fine-tuned ProtBERT-BFD trained on UniProt-ICAT which has been classified with an MCC of 0.808. As the Table 7 presents for each dataset, the variation in performance is small, as shown by the standard deviation. Figure 13 shows the validation performance in the training process for both Feed-forward Neural Network and fine-tuned Prot-BERT using UniProt-ICAT-100. As Figure 13a shows the Feed-forward Neural Network reaches maximum performance on the validation set in epoch 89 with an MCC of 0.900. Similarly, Figure 13b shows the learning curve maximum on epoch number 7 with an MCC of 0.950.

Table 8 presents the overall results of the hold-out test set for three methods implemented on each of the datasets. This table suggests the fine-tuned model outperforms the other methods in all the metrics for four datasets. Table 9 presents detailed results for the fine-tuned ProtBERT-BFD model on the UniProt-ICAT-100 dataset for each class in an hold-out test set. Despite the small size of the dataset, Copper ion, `copper(1+)`, is predicted completely correctly. Three classes of `zinc(2+)`, `iron(2+)`, and `nitrate` have scored 100% in precision, and most of the classes achieve very strong MCC scores. From Table 9 we can see the number of classes has less effect on the dataset than the number of samples. This table shows that UniProt-ICAT-100 with 12 classes has better results in

⁵https://huggingface.co/Rostlab/prot_bert_bfd



(a) Feed-forward Network training process (b) Fine-tuning process of ProtBERT-BFD

Figure 13: Comparison of validation performance in the training of TooT-ICAT. The figure shows the performance of the models with two classification methods on the validation set after each epoch of training on UniProt-ICAT-100.

all the metrics than Swiss-Prot-ICAT-100 with 11 classes because of the higher number of sequences in this dataset.

Table 2: The intermediate results: map $C2GO\_C$.

#	CHEBI term	GO terms
0	CHEBI:49713	'GO:0010348'
1	CHEBI:24636	'GO:0015528', 'GO:0005366', 'GO:0015494', 'GO:0015517', 'GO:0015653', 'GO:0034216', 'GO:0015345', 'GO:0044693', 'GO:0003957', 'GO:0009671', 'GO:0015312', 'GO:0005360', 'GO:0062158', 'GO:0015333', 'GO:0008506', 'GO:0015531', 'GO:0015492', 'GO:0010348', 'GO:0015650', 'GO:0045119', 'GO:0008900', 'GO:0015493', 'GO:0005358', 'GO:0015346', 'GO:0015385', 'GO:0015295', 'GO:0015506', 'GO:0015317', 'GO:0005280', 'GO:0015360', 'GO:0004129', 'GO:0015309', 'GO:0015311', 'GO:0005352', 'GO:0015314', 'GO:0015518', 'GO:0009672', 'GO:0015519', 'GO:0015366', 'GO:0015529', 'GO:0009675', 'GO:0005351', 'GO:0015561', 'GO:0015495', 'GO:0055053', 'GO:0015386', 'GO:0009679', 'GO:0055054', 'GO:0097079', 'GO:0042962', 'GO:0015532', 'GO:0009676', 'GO:0140211', 'GO:0015310', ...
2	CHEBI:29108	'GO:0022849', 'GO:1990028', 'GO:0048763', 'GO:0086007', 'GO:0008331', 'GO:0008273', 'GO:0140417', 'GO:0086059', 'GO:0015369', 'GO:0015275', 'GO:0008086', 'GO:0097364', 'GO:1905058', 'GO:0015278', 'GO:1905060', 'GO:0015085', 'GO:1905055', 'GO:0097365', 'GO:0008332', 'GO:0086057', 'GO:1905057', 'GO:0086038', 'GO:1905059', 'GO:0005432', 'GO:0086058', 'GO:0099511', 'GO:0098695', 'GO:0086039', 'GO:0099626', 'GO:0005262', 'GO:0015279', 'GO:0005220', 'GO:0098697', 'GO:0005388', 'GO:0072345', 'GO:0005245', 'GO:0015368', 'GO:1905054', 'GO:0099635', 'GO:1905056', 'GO:0086056', 'GO:0099604', 'GO:0005219', 'GO:0072346'
3	CHEBI:29105	'GO:0000006', 'GO:0022883', 'GO:0000007', 'GO:0015341', 'GO:0005385'
4	CHEBI:29033	'GO:0061840', 'GO:0015093'
5	CHEBI:29103	'GO:0086090', 'GO:0017032', 'GO:0060072', 'GO:0099102', 'GO:0015386', 'GO:0015387', 'GO:0015079', 'GO:0009674', 'GO:0005251', 'GO:1902282', 'GO:0022819', 'GO:0098782', 'GO:0008273', 'GO:0015271', 'GO:0005250', 'GO:0086041', 'GO:0015467', 'GO:0015277', 'GO:0086088', 'GO:0086008', 'GO:0008556', 'GO:0086037', 'GO:0022841', 'GO:0008511', 'GO:0016286', 'GO:0140107', 'GO:0005249', 'GO:0062156', 'GO:0086089', 'GO:0005228', 'GO:0005242', 'GO:0015503', 'GO:0099101', 'GO:0086086', 'GO:0015269', 'GO:0015379', 'GO:0022821', 'GO:0008900', 'GO:0015272', 'GO:0022894', 'GO:0005252', 'GO:0005267', 'GO:0070089', 'GO:0005391', 'GO:0086087'
6	CHEBI:49552	'GO:0140581'
7	CHEBI:28938	'GO:0015400', 'GO:0008519', 'GO:0015398'
8	CHEBI:29036	'GO:0043682'
9	CHEBI:29101	'GO:0044669', 'GO:0005334', 'GO:0015370', 'GO:0015655', 'GO:0009674', 'GO:0005367', 'GO:0005335', 'GO:0140161', 'GO:0101013', 'GO:0086063', 'GO:0046962', 'GO:0008554', 'GO:0005332', 'GO:0015280', 'GO:0086006', 'GO:0015498', 'GO:0005415', 'GO:0015375', 'GO:0015373', 'GO:0008510', 'GO:0015081', 'GO:0086040', 'GO:0017153', 'GO:0005369', 'GO:0061768', 'GO:0086037', 'GO:0086060', 'GO:0005436', 'GO:0008511', 'GO:0046932', 'GO:0140679', 'GO:0005362', 'GO:0015389', 'GO:0086061', 'GO:0015401', 'GO:0022818', 'GO:0015390', 'GO:0005283', 'GO:0086038', 'GO:0015361', 'GO:0005432', 'GO:0005307', 'GO:0005412', 'GO:0044736', 'GO:0008507', 'GO:0005343', 'GO:0015382', 'GO:0005295', 'GO:0005309', 'GO:0005272', 'GO:0043887', 'GO:0005298', 'GO:0015501', 'GO:0008523', ...
10	CHEBI:29034	'GO:0015092', 'GO:0015091', 'GO:0015408', 'GO:0015090'
11	CHEBI:17632	'GO:0015414', 'GO:0015513', 'GO:0010542', 'GO:0015112', 'GO:0080054', 'GO:0009671'
12	CHEBI:46502	'GO:1901238'
13	CHEBI:49709	'GO:0015107'
14	CHEBI:29242	'GO:0008490', 'GO:0015105'
15	CHEBI:16301	'GO:0015514', 'GO:0015113'
16	CHEBI:16189	'GO:0005340', 'GO:0009676', 'GO:0015419', 'GO:0009675', 'GO:0015116', 'GO:0008512', 'GO:0008271', 'GO:0015382', 'GO:0015383', 'GO:0015381'
17	CHEBI:16382	'GO:0008507', 'GO:0015111'
18	CHEBI:17996	'GO:0005334', 'GO:0015108', 'GO:0005335', 'GO:0019182', 'GO:0061797', 'GO:0005330', 'GO:0022851', 'GO:0061778', 'GO:0022852', 'GO:0015377', 'GO:0062158', 'GO:0008511', 'GO:0072320', 'GO:0008068', 'GO:0016934', 'GO:0005247', 'GO:0015379', 'GO:0015274', 'GO:0005260', 'GO:0015378', 'GO:0005254', 'GO:0005229'
19	CHEBI:35404	'GO:0015109'
20	CHEBI:15075	'GO:0140413'
21	CHEBI:17359	'GO:0000319'
22	CHEBI:17051	'GO:0062054', 'GO:1903425'

The intermediate results of Algorithm 1 for the TooT-ICAT dataset presenting a ChEBI to GO map from ChEBI leaves. This map retrieves 22 entries relating ChEBI terms to GO terms.

Table 3: The intermediate results: map $C2GO\_G$.

	CHEBI term	GO terms
0	'CHEBI:16189	'GO:0015382', 'GO:0015383', 'GO:0015419'
1	'CHEBI:29101	'GO:0015382', 'GO:0015375', 'GO:0015371', 'GO:0005298', 'GO:0015390', 'GO:0005307', 'GO:0005309', 'GO:0140679', 'GO:0005328', 'GO:0043887', 'GO:0005330', 'GO:0005332', 'GO:0043882', 'GO:0005334', 'GO:0005335', 'GO:0051978', 'GO:0005367', 'GO:0005369', 'GO:0044668', 'GO:0044669', 'GO:0008507', 'GO:0008511', 'GO:0008510', 'GO:0008508', 'GO:0015498', 'GO:0015319', 'GO:0015657', 'GO:0015655', 'GO:0061768', 'GO:0008273', 'GO:0015401', 'GO:0008520'
2	'CHEBI:24636	'GO:0042962', 'GO:0005274', 'GO:0005278', 'GO:0009671', 'GO:0097079', 'GO:0015369', 'GO:0005297', 'GO:0015386', 'GO:0008900', 'GO:0015387', 'GO:0015366', 'GO:0030171', 'GO:0015540', 'GO:0015538', 'GO:0015518', 'GO:0015532', 'GO:0015531', 'GO:0015529', 'GO:0015517', 'GO:0005352', 'GO:0015528', 'GO:0015537', 'GO:0015520', 'GO:0005364', 'GO:0140407', 'GO:0005366', 'GO:0015561', 'GO:0015535', 'GO:0015533', 'GO:0015519', 'GO:0140211', 'GO:0015360', 'GO:0015333', 'GO:0015504', 'GO:0015495', 'GO:0015653', 'GO:0015309', 'GO:0008506', 'GO:0015493', 'GO:0015492', 'GO:0097434', 'GO:0090448', 'GO:0015345', 'GO:0055054', 'GO:0062158', 'GO:0015394', 'GO:0044693', 'GO:0015317', 'GO:0010348', 'GO:0055055', 'GO:0015314', 'GO:0009672', 'GO:0010486', 'GO:0015312', 'GO:0015313', 'GO:0015650', 'GO:0015311', 'GO:0015649', 'GO:0055053', 'GO:0015310', 'GO:0015308'
3	'CHEBI:17632	'GO:0015414', 'GO:0009671', 'GO:0010542'
4	'CHEBI:35780	'GO:0015527', 'GO:0015364', 'GO:0051408', 'GO:0015415', 'GO:0051407', 'GO:0015121', 'GO:0061513'
5	'CHEBI:35404	'GO:0015109'
6	'CHEBI:29108	'GO:0015369', 'GO:0008273'
7	'CHEBI:29103	'GO:0015386', 'GO:0008900', 'GO:0015387', 'GO:0098782', 'GO:0086090', 'GO:0022841', 'GO:0008511', 'GO:1902282', 'GO:0062156', 'GO:0008273', 'GO:0017032', 'GO:0086088', 'GO:0086089', 'GO:0086086', 'GO:0086087'
8	'CHEBI:17996	'GO:0005330', 'GO:0005334', 'GO:0005335', 'GO:0019182', 'GO:0008511', 'GO:0062158'
9	'CHEBI:25197	'GO:0015097'
10	'CHEBI:29036	'GO:0043682'
11	'CHEBI:36264	'GO:0015412'
12	'CHEBI:16301	'GO:0015514'
13	'CHEBI:15075	'GO:0140413'
14	'CHEBI:63056	'GO:0140413', 'GO:0015633', 'GO:0016463'
15	'CHEBI:25155	'GO:0140613', 'GO:0015410', 'GO:0010486'
16	'CHEBI:37239	'GO:0042888'
17	'CHEBI:29242	'GO:0008490'
18	'CHEBI:35172	'GO:0015100'
19	'CHEBI:16382	'GO:0008507'
20	'CHEBI:49709	'GO:0015107'
21	'CHEBI:17359	'GO:0000319'
22	'CHEBI:49552	'GO:0140581'
23	'CHEBI:25516	'GO:0015413'
24	'CHEBI:39127	'GO:0015444'
25	'CHEBI:49713	'GO:0010348'
26	'CHEBI:63063	'GO:0015434', 'GO:0008551'
27	'CHEBI:23336	'GO:0032778'
28	'CHEBI:60253	'GO:0015445'
29	'CHEBI:17051	'GO:0062054'
30	'CHEBI:46502	'GO:1901238'
31	'CHEBI:29034	'GO:0015408'

The intermediate results of Algorithm 2 for the TooT-ICAT dataset presenting a ChEBI to GO map from Gene Ontology leaves. This map retrieves 31 entries relating ChEBI terms to GO terms.

Table 4: Catalytic activity GO terms eliminated from C2GO map of ICAT datasets.

GO term	Name
GO:0004129	cytochrome-c oxidase activity
GO:0003957	NAD(P)+ transhydrogenase (B-specific) activity
GO:0046932	sodium-transporting ATP synthase activity, rotational mechanism
GO:0008121	ubiquinol-cytochrome-c reductase activity
GO:0046933	proton-transporting ATP synthase activity, rotational mechanism

Five GO terms in the intersection of **catalytic activity** and **transmembrane transport activity** are eliminated from the TooT-ICAT C2GO map due to out-of-scope descriptions

Table 5: Four extracted TooT-ICAT dataset series.

Dataset	Size	Trainset	Testset	# Classes
UniProt-ICAT-100	4,112	3,289	823	12
UniProt-ICAT-60	1,429	1,143	286	11
SwissProt-ICAT-100	2,140	1,712	428	11
SwissProt-ICAT-60	1,098	878	220	11

The number of sequences and classes in each of the four datasets. Each dataset has been divided into a training set and test set with an 80:20 ratio randomly and stratified.

Table 6: Size of substrate classes for each TooT-ICAT dataset.

Class	CHEBI	Substrate	UP-ICAT-100	UP-ICAT-60	SP-ICAT-10	SP-ICAT-60
0	CHEBI:24636	proton	1307	600	883	484
1	CHEBI:29108	calcium(2+)	868	210	350	154
2	CHEBI:29103	potassium(1+)	726	193	299	146
3	CHEBI:17996	chloride	458	129	177	75
4	CHEBI:29101	sodium(1+)	429	123	235	99
5	CHEBI:16189	sulfate	98	49	38	27
6	CHEBI:29105	zinc(2+)	80	51	67	47
7	CHEBI:28938	ammonium	55	28	36	23
8	CHEBI:17632	nitrate	28	18	27	18
9	CHEBI:29033	iron(2+)	24	16	16	13
10	CHEBI:35780	phosphate ion	22	12	12	12
11	CHEBI:49552	copper(1+)	17	<i>9</i>	<i>7</i>	<i>6</i>
Total	-	-	4,112	1,429	2,140	1,098

The twelve specific substrates and their number of sequences in each of the four datasets: UniProt-ICAT-100, Swiss-Prot-ICAT-100, UniProt-ICAT-60, and Swiss-Prot-ICAT-60 respectively. Numbers in italic are below the threshold of 10 sequences, so that substrate class is not included in experiments with that dataset.

Table 7: Five-fold cross-validation results for TooT-ICAT datasets.

Dataset	#Classes	Accuracy	Precision	Recall	F1-score	MCC
UniProt-ICAT-100	12	0.848 ± 0.01	0.679 ± 0.05	0.561 ± 0.02	0.598 ± 0.02	0.808 ± 0.02
UniProt-ICAT-60	11	0.727 ± 0.03	0.586 ± 0.08	0.430 ± 0.04	0.467 ± 0.05	0.631 ± 0.04
SwissProt-ICAT-100	11	0.825 ± 0.01	0.628 ± 0.04	0.512 ± 0.04	0.542 ± 0.04	0.766 ± 0.02
SwissProt-ICAT-60	11	0.723 ± 0.03	0.476 ± 0.06	0.395 ± 0.04	0.415 ± 0.05	0.617 ± 0.05

The table presents the results of a fine-tuned ProtBERT-BFD the 5-fold cross-validation for each dataset in TooT-ICAT with five metrics of accuracy, precision, recall, F1-score, and MCC.

Table 8: Hold-out test set results on TooT-ICAT datasets.

Dataset	# Classes	Accuracy			Precision			Recall			F1-score			MCC		
		LR	FNN	FTB	LR	FNN	FTB	LR	FNN	FTB	LR	FNN	FTB	LR	FNN	FTB
UniProt-ICAT-100	12	0.975	0.985	0.993	0.852	0.911	0.959	0.630	0.832	0.881	0.680	0.858	0.913	0.813	0.889	0.948
UniProt-ICAT-60	11	0.952	0.963	0.982	0.734	0.794	0.899	0.415	0.597	0.868	0.442	0.664	0.903	0.640	0.726	0.867
SwissProt-ICAT-100	11	0.971	0.981	0.991	0.839	0.895	0.951	0.509	0.748	0.879	0.548	0.783	0.907	0.785	0.861	0.936
SwissProt-ICAT-60	11	0.950	0.956	0.979	0.723	0.759	0.886	0.425	0.595	0.757	0.465	0.598	0.802	0.616	0.682	0.850

The table presents the results for hold-out test sets from classifiers: Logistic Regression (LR), Feedforward Neural Network (FNN), and Fine-tuned ProtBERT-BFD (FTB) for each of the four datasets in TooT-ICAT

Table 9: Detailed results for UniProt-ICAT-100 dataset classification.

Substrate	Trainset	Validation	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	837	209	261	245	17	16	545	0.960	0.935	0.939	0.937	0.908
calcium(2+)	555	139	174	157	12	17	637	0.965	0.929	0.902	0.916	0.893
potassium(1+)	465	116	145	136	15	9	663	0.971	0.901	0.938	0.919	0.901
chloride	293	73	92	86	9	6	722	0.982	0.905	0.935	0.920	0.910
sodium(1+)	274	69	86	75	14	11	723	0.970	0.843	0.872	0.857	0.840
sulfate	62	16	20	14	4	6	799	0.988	0.778	0.700	0.737	0.732
zinc(2+)	51	13	16	14	0	2	807	0.998	1.000	0.875	0.933	0.934
ammonium	35	9	11	11	1	0	811	0.999	0.917	1.000	0.957	0.957
nitrate	18	4	6	4	0	2	817	0.998	1.000	0.667	0.800	0.816
iron(2+)	15	4	5	2	0	3	818	0.996	1.000	0.400	0.571	0.631
phosphate ion	15	3	4	3	1	1	818	0.998	0.750	0.750	0.750	0.749
copper(1+)	11	3	3	3	0	0	820	1.000	1.000	1.000	1.000	1.000

The table shows the detailed results for fine-tuned ProtBERT-BFD model trained on the UniProt-ICAT-100 and tested on the hold-out test set. The results show 100% accurate predictions on the smallest class (`copper(1+)`).

3.7 TooT-SPEC

This project is conducted on the substrate specificity of transmembrane transport proteins, predicting the specific substrate carried by a transporter. The dataset constructed for this study is the TooT-SPEC series of datasets, which is an imbalanced dataset with small-scale minority classes. The experiments are conducted using a fine-tuned large language model, ProtBERT-BFD with a classifier layer to predict the specific substrate. The code and the studied dataset of this project are publicly available at the following link <https://github.com/simaataei/TooT-BERT-SC>.

3.7.1 Dataset description

The TooT-SPEC dataset series is utilized in this research to study the substrate specificity prediction task. One of the objectives of this study is to include as many classes as possible while testing the pre-trained model’s ability to handle low sample sizes. Based on preliminary studies, a threshold of 10 sequences is applied in the TooT-ICAT datasets. This threshold is reduced to 3 sequences in the TooT-SPEC datasets to maximize substrate coverage while ensuring that each class has at least one sample in the train, test, and validation sets. Table 43 from Appendix B presents detailed information about the number of substrates in each extracted dataset. Table 10 shows the four datasets split into training and test sets. A 66:34 split is chosen to accommodate the smallest class in the dataset with 3 sequences. This ratio ensures that one sequence from the smallest class is included in the test set using a stratified split.

Table 10: Four extracted TooT-SPEC datasets.

Dataset	Size	Trainset	Testset	# Classes
UniProt-SPEC-100	4,455	2,970	1,485	96
UniProt-SPEC-60	1,628	1,085	543	69
SwissProt-SPEC-100	2,340	1,560	780	86
SwissProt-SPEC-60	1,262	841	421	62

The table presents four different datasets extracted for the TooT-SPEC project, with their total sample size and the number of classes. Each dataset has been split into train/test subsets with a ratio of 80:20 randomly and stratified.

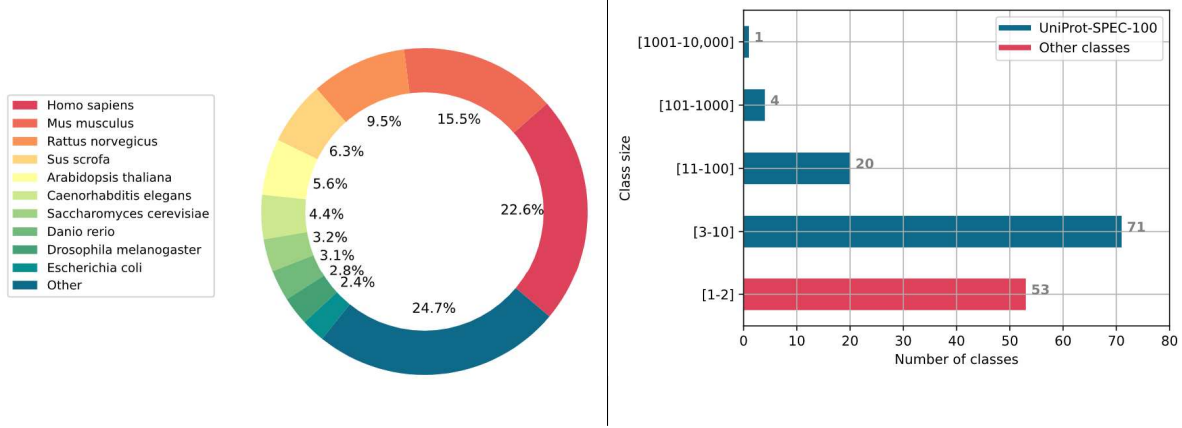


Figure 14: UniProt-SPEC-100 dataset analysis.

The left plot shows the analysis of different organisms' share in the UniProt-SPEC-100 dataset; The list of organisms is presented in the legend and organisms with less than 100 samples in the dataset are categorized as "Other". The right side plot is an analysis of class frequency for different class size ranges in UniProt-100.

3.7.2 Dataset Analysis

The UniProt-SPEC-100 is the main dataset of this research; An analysis of this dataset is presented. Figure 14 presents the distribution across species and over classes of the dataset. As depicted in Figure 14 and Table 43 51.6% of the data occurs in 4 classes with sizes in $[101, 1000]$ sequences. These classes are CHEBI:29108-calcium(2+), CHEBI:29103-potassium(1+), CHEBI:17996-chloride, and CHEBI:29101-sodium(1+). Furthermore, 27.5% of the data is in one majority class: CHEBI:24636-proton with 1,226 samples.

The UniProt-SPEC-100 dataset includes 71 classes with 3 to 10 sequences. These classes form 8.4% of the dataset. They include CHEBI:29036-copper(2+), CHEBI:29034-iron(3+) and CHEBI:17992-sucrose which are considered important substrates for cell metabolism. Figure 14 also shows that there are 20 classes with sizes between 11 to 100 which include 12.5% percent of the dataset. The 53 substrate classes with less than 3 sequences are not included in the UniProt-SPEC-100 dataset. In conclusion, 79.1% of the data samples are distributed in the 5 largest classes and the remaining 20.9% occur in 91 smaller classes.

This dataset is considered a long-tailed imbalanced dataset as depicted in Figure 15. The ratio of samples from the smallest to the largest class in the UniProt-SPEC-100 dataset is 1:408 which shows not only that the data is unevenly distributed between the classes but also that the class size span is high. According to [Kra16], the smallest to the

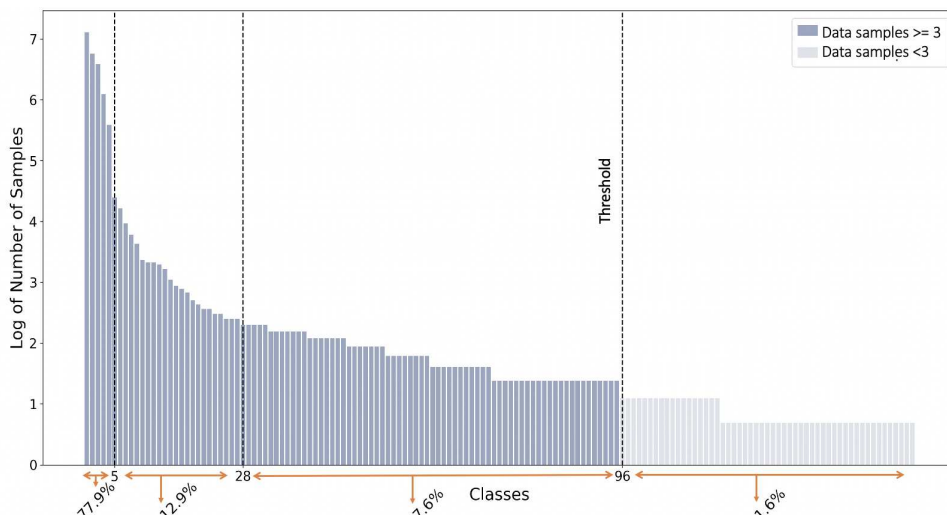


Figure 15: Log-scale of sample distribution on each class for UniProt-SPEC

The figure shows the log scale of sample distribution (bases of 10) on each transmembrane transport protein class retrieved from UniProt. The classes with more than 3 samples (Dark gray) are included in the UniProt-SPEC series of datasets. As depicted, the distribution follows a long-tailed shape with 5 head classes holding and 91 tail classes. The head classes are substrate transporters of **proton**, **calcium(2+)**, **potassium (1+)**, **chloride**, and **sodium (1+)**. Classes with less than 3 samples (light gray) are not included in the UniProt-SPEC series of dataset.

largest class ratio of more than 1:100 is considered severe class imbalance.

One of the challenges in building this dataset is the lack of experimental validation in the UniProt dataset. According to the results from the UniProt, 12,589 transmembrane protein sequences exist at the protein level but only 19 sequences are assigned GO molecular function terms where specific transporter activity is inferred from experiments.

3.7.3 Methodology

Here, we examine the performance of transfer learning on imbalanced datasets with small-scale minority classes. We fine-tuned the ProtBERT-BFD protein language model to classify the protein sequences based on their specific carried substrates. Three-fold cross-validation is performed to validate the performance results.

3.7.4 Fine-tuning ProtBERT-BFD model

In this study, pre-trained ProtBERT-BFD model has been fine-tuned on the extracted datasets separately to update the model weights for the specified classification task. Our model includes a single-layer Feed-forward network followed by a soft-max function to predict the probabilities of each class in the dataset. We used the Cross entropy loss function and Adam optimizer with learning rate = 5×10^{-5} for batch size 1 [KB15]. After 15 epochs of training, the model converges according to the validation loss. The implementations are applied using the Pytorch package and the pre-trained ProtBERT-BFD model is collected from the HuggingFace website [WDS⁺20, PGM⁺19].

3.7.5 Cross-validation

Due to the existence of small-scale size classes in the datasets, the results are expected to be highly affected by the randomness of the train-test split. Especially for the classes of size three, there are two sequences in the training set and only one sequence in the testset. Accordingly, a 3-fold cross-validation experiment has been implemented to examine the consistency of the results. The UniProt-SPEC-100 dataset has been divided into 3-folds with the stratified approach to include all of the classes in both test and training sets. Then, the fine-tuning process has been implemented on each training and testset separately. The performance of the method was determined using 3-fold cross-validation and reported the performance variations between the runs by computing the standard deviation.

3.7.6 Results and discussion

Table 11 shows the overall performance evaluation results of fine-tuned ProtBERT-BFD on the test sets of 4 different datasets. This table includes two columns: “Total Classes” and “Effective Classes”. These columns highlight that some classes in the dataset, although included in the training process, do not have their sequences identified in the test set. The term “Total classes” refers to the number of classes in the full data set and the “Effective Classes” refers to the subset of classes for which predictions are made in the test set. For instance, UniProt-SPEC-100 contains 96 substrate classes used during training. However, only 83 of these classes are predicted in the test set. Specifically, 13 substrate classes have no True Positive (TP) or False Positive (FP) predictions, despite being present in both the training and test sets. In this case, the number of “Total classes” is 96 while

the number of “Effective classes” is 83. Please note that the reported metric values are calculated for effective classes. Please note that the reported metric values are calculated for effective classes. The results for UniProt-SPEC-100 outperform other datasets in all of the metrics, so it is selected as the TooT-SPEC model for specific substrate prediction in transmembrane transporter datasets. The model has converged after 10 epochs and reaches the maximum performance in epoch number 13.

Table 11: TooT-SPEC results on test sets for 4 different datasets.

Dataset	# Total Classes	# Effective Classes	F1-score	MCC
UniProt-SPEC-100	96	83	0.923	0.908
UniProt-SPEC-60	69	52	0.849	0.818
SwissProt-SPEC-100	86	73	0.872	0.848
SwissProt-SPEC-60	62	44	0.786	0.743

The table shows the results for the hold-out test set on 4 different datasets with two metrics of F1-score and MCC. Column #Total classes shows the number of classes and column # Effective classes shows the number of classes retrieved in the test results. The metrics are calculated for the number of effective classes.

Table 12: TooT-SPEC 3-fold cross-validation results for UniProt-SPEC-100.

Fold	# Total Classes	# Effective Classes	F1-score	MCC
1	96	83	0.923	0.908
2	96	84	0.939	0.927
3	96	83	0.902	0.884
Average	96	83.33 ± 0.58	0.921 ± 0.02	0.906 ± 0.02

The table presents the F1-score and MCC for each of the 3 folds in the UniProt-SPEC-100 dataset in the cross-validation process. As the table shows there are 96 classes in the dataset from which 83, 84, and 83 classes have been retrieved in the test processes of fold 1,2, and 3 respectively.

To investigate the issue of 13 unpredicted substrates, the detailed results for each class are presented in Appendix C.2. In Table 45 classes with zero prediction are reported with NaN values for Precision and MCC metrics. As this table indicates small-scale classes are

the majority of classes with zero prediction. All 13 classes with NaN values have at most 5 sequences in the dataset.

The unpredicted classes in the results were foreseeable according to the imbalance in the dataset and the low number of training set samples in these classes. The compelling part of the result table is the performance of TooT-SPEC on predicted small-scale classes. According to Table 45, 34 out of 71 classes with less than 10 sequences in the dataset are predicted completely correctly with an MCC of 1.0. This table also shows 25 minority classes with positive values of MCC for their prediction. This observation supports the ProtBERT-BFD language model’s potential in the classification of small-scale classes in imbalanced datasets. However, among the remaining minority classes, partial misclassifications (e.g. false positives and/or false negatives) occur, reflecting the complexity of learning from very limited data. While the model often exhibits strong precision in minority classes, minimizing FP, false negatives remain a challenge due to the limited variety of training examples for rare substrates, which can result in reduced recall for these classes.

However, it is important to mention the impact of randomness in the test-train ratio on the final results. Due to the minimum of 3 sequences set on the number of samples in each class, there are 62 classes with only one sequence in their test set. The prediction of one test sequence based on two sequences in the training set may be dependent on the randomness of the test-train split.

Therefore, a 3-fold cross-validation is implemented to assess the impact of randomness. For small classes with only three samples, each sample is used as the test sample in one fold. This approach eliminates the effect of randomness in test set selection, ensuring that every sample in small classes is included in the test set once, thereby making the evaluation more reliable. The 3-fold cross-validation experiment results are presented in Table 12. As the results indicate, the overall performance of the TooT-SPEC has been consistent on the three different folds. The model also predicted one extra class reaching 84 predicted classes for the second fold. The unpredicted classes vary from fold to fold, however, the total number is consistent between 12 to 13. This indicates the effect of randomness in the test-train split that an un-predicted class in one random split could be predicted in another split (Table 13).

An observation from TooT-SPEC results is that the deficient performance of transfer learning on imbalanced data is mainly due to the low recall of minority classes. Despite this, precision on minority classes is surprisingly high. The model has almost perfect precision on minority classes, suggesting that the model is conservative in classifying

samples into minority classes; once it makes such a prediction we can be confident it is correct.

Table 13: MCC comparison of unclassified classes in different folds.

Substrate	Fold 1	Fold 2	Fold 3
copper(1+)	0.925	0.913	NaN
2-oxoglutarate(2-)	0.816	0.816	NaN
maltose	0.707	NaN	0.499
iron(3+)	1.000	NaN	0.407
hydrogencarbonate	0.515	NaN	0.499
N-ribosylnicotinamide	NaN	0.707	1.000
gamma-aminobutyric acid zwitterion	0.707	NaN	NaN
L-arabinose	NaN	0.499	NaN
bacteriocin	0.707	1.000	NaN
L-histidine zwitterion	NaN	-0.001	NaN
L-methionine zwitterion	NaN	NaN	NaN
boric acid	NaN	1.000	0.707
maltodextrin	NaN	1.000	1.000
folate(2-)	NaN	NaN	NaN
molybdate	NaN	NaN	NaN
N-retinylidenephosphatidylethanolamine	1.000	NaN	NaN
sn-glycerol 3-phosphate(2-)	NaN	NaN	NaN
flavin adenine dinucleotide	NaN	NaN	0.707
CMP-N-acetyl-beta-neuraminate(2-)	0.707	1.000	NaN
L-tryptophan zwitterion	-0.001	NaN	0.577
phosphonatoenolpyruvate	1.000	NaN	1.000
guanine	NaN	1.000	1.000
xanthine	NaN	1.000	1.000
glycine betaine	NaN	NaN	NaN

The table presents the detailed MCC of unclassified substrate classes in each fold of cross-validation. There are 5 substrate classes unclassified in all 3 folds.

Chapter 4

TooT-Open-ICAT: Open classification of inorganic ion transport proteins

4.1 Introduction

Inorganic ions, as ions lacking carbon-hydrogen bonds, are crucial compounds in biological systems [BS20]. They play key roles in processes such as nerve transmission, muscle contraction, cell signaling, and drug uptake [RAAS11, PZA07]. Inorganic ion transporters and channels facilitate the transport of these essential ions. Despite their importance and potential as drug targets, these proteins remain understudied due to the costly and time-intensive nature of wet lab experiments [KMPG08, SUG⁺17]. Therefore, computational methods offer a potentially valuable approach for studying these proteins.

Predicting the specificity of inorganic ion transport proteins with machine learning models is challenging due to the composition of public datasets. These datasets typically contain three groups of samples. The first group includes well-characterized proteins with many annotated examples, such as `sodium` transport proteins. This group comprises 1,010,507 entries in UniProt-KB from which 5,071 entries are reviewed and included in SwissProt-KB, accessed on December 19, 2024, based on the specified query [The23].

$$\langle \text{goa:}(\textit{“transmembrane transporter activity [22857]”}) \\ \text{AND any: “Sodium ion”} \rangle$$

The second group consists of proteins with fewer annotated sequences, such as `chromate(2-) transporters`, which has 41,104 entries in UniProt-KB with only 7 entries reviewed in SwissProt-KB, accessed on the same date using the following query [The23].

$\langle \text{goa:}(\text{"transmembrane transporter activity [22857]"})$
 $\text{AND any: "chromate(2-)"}$

The limited data availability for the second group of substrate transporters stems from several factors, such as their rarity in biological processes or the lack of research focus on them. The third group contains unannotated *de novo* sequences with no known references. These proteins may transport known or unknown inorganic ions. Based on this notation we define four groups of specific substrate transporter classes as follows.

Definition 4.1.1. *Known classes* are a group of transport protein sequences transporting known chemical compounds available in **ChEBI** ontology across the membrane.

Definition 4.1.2. *Unknown classes* are a group of transport protein sequences transporting unknown chemical compounds not included **ChEBI** ontology across the membrane.

Definition 4.1.3. *Seen classes* are a set of classes that a machine learning model is trained on, using their corresponding samples during the training phase.

Definition 4.1.4. *Unseen classes* are a set of classes that are absent from the training set, with no samples included during the training phase of a machine learning model.

This research aims to solve the inorganic ion transporter prediction problem by developing a deep learning model that (1) identifies inorganic ion transport proteins from proteins transporting other substrates based on their primary structure, (2) predicts the specific substrates transported by the identified protein sequences, and (3) detects the unseen substrate classes as a novel class through rejection.

The dataset and code for this project are publicly available on GitHub at the link: <https://github.com/simaataei/TooT-Open-ICAT>.

4.2 Problem statement: Inorganic ion prediction

The problem of predicting specific inorganic ion substrates carried by the transporters is redefined as an open classification problem. The openness of the problem enables the model to identify inorganic ion transporters in the transportome and reject other transporters. The model will either assign a label from the label set to the sample or reject the sample from any of the classes in the training set. This means that the model could differentiate the samples from other classes of proteins not included in the training set.

Given a labeled set of n transmembrane transport protein sequences as:

$$D = \{(p_1, l_1), (p_2, l_2), (p_3, l_3), \dots, (p_n, l_n)\},$$

p_i representing the sequence of the i -th protein, as a chain of amino acids, associated to a label, l_i , in the label set L of seen classes; build a model, $f(p)$, to classify samples of the test set, D_{test} , as:

$$D_{test} = \{p_1, p_2, p_3, \dots, p_z\}$$

associated to a label set as:

$$L = \{l_1, l_2, l_3, \dots, l_n\}$$

or reject the sample. This means for a transportome of an organism, the model is to predict (1) the specific substrate it transports across the cell by assigning it to one of the classes in the label set L or (2) reject the protein as a novel class. Note that in this notation, each sample is associated with a single label. As detailed in Sections 3.4 and 3.6, transport proteins with multiple substrates are excluded from the ICAT dataset series, including those involved in transporting inorganic cations and anions.

4.3 Challenges

Predicting substrate specificity for inorganic ion transporters using machine learning presents significant challenges. Although ion transporters represent a substantial portion of the transportome in any organism, many of these proteins are poorly characterized. This results in small datasets with limited samples per class, making it challenging for machine learning algorithms to perform accurately. The lack of comprehensive data limits predictive power and model effectiveness.

In addition, the focus of pharmaceutical research and the varying prevalence of ion transporters in biological systems have produced imbalanced datasets. Some transporters are well-studied, with large numbers of samples available, while others are less explored. For instance, the number of annotated **Sodium** ion transporters is 5,071, which is significantly higher than the 7 annotated **Chromate(2-)** transporters in SwissProt-KB (details in Section 4.1) [The23]. This imbalanced sample distribution further complicates substrate specificity prediction. Addressing these limitations requires innovative approaches to develop reliable machine-learning models for inorganic ion transporter prediction.

4.4 Methodology

4.4.1 Introduction

Our previous model, TooT-ICAT (explained in Section 3.6), predicts inorganic ion transporters for 12 specific substrates [AB22b]. However, a full transportome of an organism includes all transport proteins, including both inorganic ion transporters and other substrate transporters. The TooT-Open-ICAT model extends the previous approach to build a model for analyzing the transportome including the single substrate transporters in each organism. This model leverages open-world classification by removing the closed-world assumption, where all test classes must have appeared in the training set [BB15]. It allows the classifier to handle test samples from unseen classes by rejecting them as *Novel classes*.

The initial proposal for TooT-Open-ICAT involved setting a threshold on the TooT-ICAT model’s classifier. However, the model did not perform as expected. While it successfully classified 12 known classes, it rejected too few sequences to account for the diversity of transport proteins in the transportome. This approach overlooked a broader group of transport proteins that transport substrates other than inorganic ions. This issue is mainly due to the limited coverage of the UniProt-ICAT-100 dataset, which only includes sequences from 12 inorganic ion classes and may not represent the true diversity of transportomes.

We developed a two-component network: the ICAT detector and the specific substrate predictor to address this. The ICAT detector filters out non-ICAT transport proteins (proteins that do not transport inorganic ions). The ICAT transporters are then passed to the specific substrate predictor, which either assigns one of the 12 known substrate labels or rejects it based on a threshold for proteins transporting other inorganic ions not in the UniProt-ICAT-100 dataset as a novel class. The structure of these components is detailed in Section 4.4.2.

4.4.2 Network Structure

TooT-Open-ICAT consists of two components, both based on ProtBERT-BFD networks. This model, introduced in [EHD⁺22], is a BERT model with thirty layers of sixteen attention-head transformer encoders, comprising a total of 420 million parameters. It is

pre-trained on a corpus of protein sequences from the BFD database¹. Each model is fine-tuned independently on a separate dataset by adding a fully connected layer and a softmax layer as a classifier.

The first ProtBERT-BFD network, referred to as the ICAT-transporter detector, is fine-tuned on the UniProt-ICAT-binary dataset to differentiate inorganic ion transport proteins (ICAT transport proteins) from proteins transporting other substrates (non-ICAT transport proteins). Details about this dataset can be found in Section 4.4.3. The ICAT transport proteins identified by this model are passed to a second classifier, called the specific substrate predictor, which predicts their specific substrate. This second ProtBERT-BFD network is fine-tuned on the UniProt-ICAT-100 dataset, with a rejection option to handle ICAT transport proteins not included in the 12 classes of the training set.

The rejection option is implemented using a threshold, denoted by τ , applied to the output vectors of the classification model. This threshold enables the model to reject a sample and refrain from assigning a class label if none of the class output values exceed τ . Two different approaches are used to determine this threshold:

- **Fixed Threshold:** A single, fixed threshold is determined and uniformly applied across all classes. This threshold is set for the outputs of the softmax layer. In this approach, the threshold τ is fixed at 0.5, meaning that a sample is rejected if the maximum softmax output is below this value. Consequently, any prediction based on a softmax score lower than 0.5 is disregarded, and the corresponding sample is rejected as a novel class sample.
- **Class-Specific Thresholds:** Different τ values are computed for each class in the dataset. These thresholds are determined using the validation set, which is a stratified random subset of the full dataset containing samples from all classes (see Section 3.6 for more details). The class-specific thresholds (τ_i) for each class C_i in the validation set are calculated as follows.

$$\tau_i = \min_{j=1}^k \left(\min_{x \in \text{TruePositiveSet}_i^{(j)}} s_i(x) \right)$$

$$\forall i \in \{1, 2, \dots, N\}, \quad \forall j \in \{1, 2, \dots, k\}$$

¹<https://bfd.mmseqs.com>

where N is the number of classes; k is the number of folds; i and j are indexes. $\text{TruePositiveSet}_i^{(j)}$ is the set of correctly predicted validation samples for class i in fold j . The variable x is the output of the penultimate layer for a correctly predicted validation sample, and $s_i(x)$ is the softmax score for class i for the sample x .

Figure 16 illustrates an example of class-specific thresholds determined based on the softmax outputs of the validation samples. Specifically, for each class, the model’s softmax scores for all validation samples are collected. The predicted class for each sample corresponds to the index of the highest softmax value (i.e., the argmax of the softmax output). To establish the threshold for a given class, the minimum softmax value among all correctly predicted validation samples of that class is identified. This minimum value serves as the class-specific threshold, ensuring that only samples with a softmax score above this threshold are considered confidently classified.

The final class-specific threshold is the minimum of all thresholds in 3 folds of cross-validations per each class. During the test phase, each test sample is first classified by the model into a specific class, denoted as class i . The softmax score corresponding to class i is then compared to the class-specific threshold determined during validation. If the softmax score is lower than this threshold, the sample is rejected and considered to belong to a novel class unseen during training. Additionally, this approach is extended to the outputs of the penultimate layer, rather than being limited to the softmax layer.

In both approaches, the threshold τ is calculated and used for rejection during testing. For each test sample, the values from the final layers are compared to the corresponding threshold, τ . If the highest class value at the softmax layer (or at the penultimate layer in a separate experiment) is greater than the threshold, the sample is assigned to that class using the argmax function. Otherwise, if the highest class value is smaller than the threshold, the sample is considered dissimilar to the seen classes. This suggests that it may belong to new classes and is rejected from the training dataset. Further details on these methods and their application can be found in Section 4.4.5.

4.4.3 Dataset

The UniProt-ICAT-100 is a dataset of non-identical inorganic cation and anion transport (ICAT) protein sequences, including both channels and transporters, extracted from

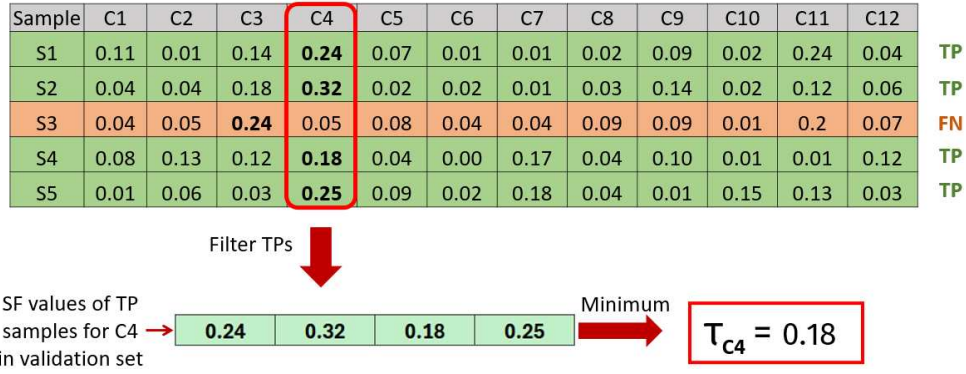


Figure 16: An example of class-specific threshold calculation in one fold. The figure illustrates class-specific threshold calculation for C4. The table shows softmax (SF) values, where the highest value (bolded) indicates the predicted class. Among five validation samples, S3 is misclassified as C3 (false negative, FN). The threshold τ for C4 is determined by selecting true positive (TP) samples, extracting their softmax values, and setting τ as the minimum value. The figure represents one fold of a 3-fold cross-validation, with the final threshold per class defined as the minimum τ across all folds.

UniProt-KB [The21b]. The dataset construction steps are described in Section 3.4 [AB22b]. This results in a dataset of 4,112 samples from 12 specific substrate classes of **proton**, **calcium(2+)**, **potassium(1+)**, **chloride**, **sodium(1+)**, **sulfate**, **zinc(2+)**, **ammonium**, **nitrate**, **iron(2+)**, **phosphate ion**, and **copper(1+)** transport proteins (ordered in decreasing number of samples). The details of samples for each class of this dataset are available in Table 15. As the table shows this dataset is an imbalanced dataset with a majority of the class of **proton**.

A second dataset, UniProt-ICAT-binary, is constructed to differentiate the inorganic ion transport proteins and other transport sequences. This dataset includes two classes: **ICAT transport proteins**, as all the sequences in UniProt-ICAT-100, and **non-ICAT transport proteins** class. The second class includes 4,000 samples extracted randomly from UniProt, resulting in a size of 8,112 sequences for the full dataset.

To construct the UniProt-ICAT-binary dataset, the following steps are undertaken. First, protein sequences with Gene Ontology annotations indicating transport activity, excluding those categorized as inorganic cation and anion transporters, are collected from UniProt-KB using the specified query:

Query = ‘GO:0022857 *transmembrane transporter activity*’ and NOT ‘GO:0022890

inorganic cation transmembrane transporter activity’ and NOT ‘GO:0015103 inorganic anion transmembrane transporter activity’

This query retrieves 14,133 reviewed protein sequences. Integrating these with the ICAT-transport protein class creates an imbalanced dataset containing potentially identical sequences. To address this, the CD-HIT algorithm is applied with a similarity threshold of 100% [LG06] to reduce the pairwise identity level.

To balance the binary dataset, a random downsampling algorithm is employed on the new class of non-ICAT transport proteins. This step is implemented using the `scikit-learn` package in Python, setting `n_samples` to 4000 and `random_state` to 42 [PVG⁺11]. The resulting dataset is named UniProt-ICAT-binary, with detailed class statistics provided in Table 14.

Table 14: UniProt-ICAT-binary dataset details

Class	Number of samples
ICAT transport proteins	4,112
Non-ICAT transport proteins	4,000
Total	8,112

The class of “ICAT transport proteins” includes transport proteins of inorganic cation and anions and the class of “non-ICAT transport proteins” includes transport proteins transporting any substrate other than inorganic cation and anions.

4.4.4 Training

Each dataset is split into training, validation, and hold-out test sets in a 64:16:20 ratio. The `train_test_split` function from `scikit-learn` [PVG⁺11] performs the stratified random split to ensure all classes are represented in all three subsets. Models are implemented in Python, version 3.9.1, using the `Transformers` package [WDS⁺20]. A grid search with `Optuna` selects the weighted focal loss function and Adam optimizer [ASY⁺19, LGG⁺17, KB15]. The ICAT transporter detector and specific substrate predictor are trained for 10 and 13 epochs, respectively.

During testing, sequences are first processed by the ICAT transporter detector. Samples classified as ICAT transport proteins are then passed to the specific substrate predictor, which performs open classification. It assigns sequences to one of 12 classes or rejects them as other inorganic ion transport proteins. Figure 17 illustrates this process.

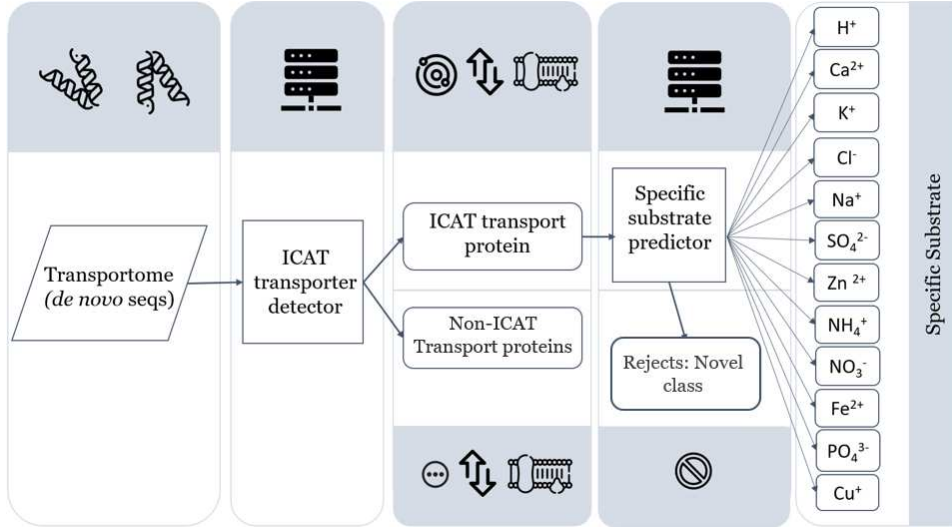


Figure 17: The testing process of TooT-Open-ICAT.

The input is an organism’s transportome, including unseen protein sequences. First, the ICAT transporter detector model identifies inorganic ion transport proteins within the transportome. The detected transport proteins are then passed to the specific substrate predictor model, which either assigns them to one of 11 known classes or rejects them as a novel class.

4.4.5 Thresholded classification layer

The open classification model, with a sample rejection option, is implemented by applying a threshold to the outputs of either the final or penultimate layers. The final layer of the model as described in Section 5.3 the final layer is a softmax layer that calculates the probabilities for each class. The softmax function computes the probability of a sample belonging to each class using the following formula:

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^n \exp(x_j)} \quad (29)$$

A softmax classifier uses the “Arg max” strategy to return the class with the highest probability as the predicted class. In open classification, a sample can be rejected from all seen classes by applying a threshold to the selected probability. This means the maximum probability is not the single constraint for retrieving the predictions and the value of the probability is also considered in making the prediction. Following [SXL17], we propose setting a threshold, τ , on the softmax output to adjust the model’s predictions. This threshold is applied to the last fully connected layer of the network, denoted as d , and to d_i , the values corresponding to l_i in d .

$$\hat{y} = \begin{cases} \text{Reject}, & \text{if } \forall l_i \in L \text{ softmax}(d_i) < \tau \\ \text{Argmax}_{l_i \in L}(\text{softmax}(d_i)), & \text{otherwise} \end{cases}$$

The threshold τ is calculated in two different approaches: (1) a fixed threshold set to all of the classes and (2) a variable threshold for each class calculated based on the validation set.

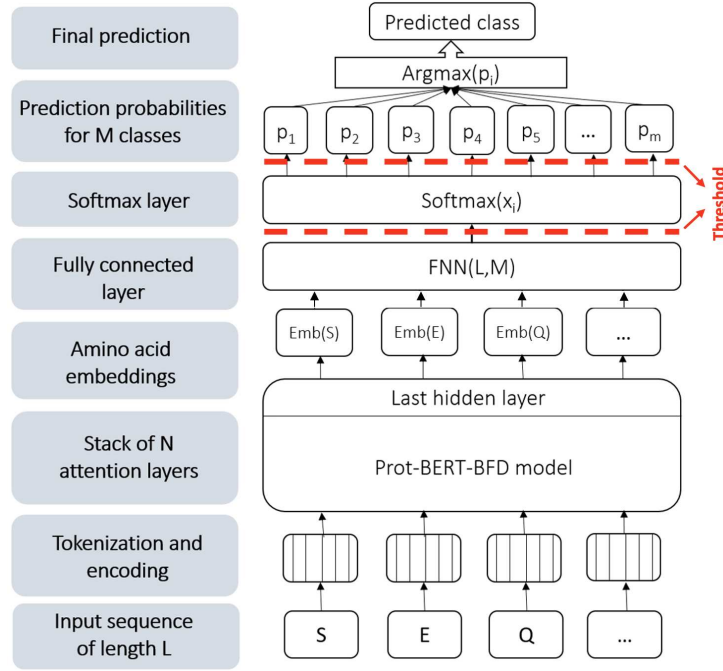


Figure 18: Thresholded network structure in TooT-Open-ICAT.

The figure shows the TooT-Open-ICAT model with layers of the thresholds shown as red dashed lines. The softmax layer (as the final layer) and the penultimate layer are thresholded separately resulting in two different models of TooT-Open-ICAT.

In the fixed threshold model, only strong softmax probabilities are accepted as predictions. A fixed threshold of $\tau = 0.5$ is applied to the softmax layer outputs across all classes, rejecting samples with probabilities below this threshold. In another approach, thresholds are calculated individually for each class based on the validation set. The softmax probabilities for True Positive (TP) samples from the validation set are collected, and the minimum probability is set as the threshold for each class. The minimum values from 3-fold cross-validation are used as the τ for each class. The same method is applied

to penultimate layer values independently. The network structure with thresholds is presented in Figure 18. Table 15 shows the calculated thresholds for each class, based on the validation set, for both the softmax and penultimate layers.

Table 15: Calculated thresholds for each class based on cross-validation.

Class	Substrate	Trainset	Validation	Testset	SF layer threshold	PU layer thershold
0	proton	837	209	261	0.916959	2.523486
1	calcium(2+)	555	139	174	0.564509	3.649124
2	potassium(1+)	465	116	145	0.404441	1.954091
3	chloride	293	73	92	0.983802	3.815068
4	sodium(1+)	274	69	86	0.380023	3.088192
5	sulfate	62	16	20	0.487306	2.189013
6	zinc(2+)	51	13	16	0.923111	3.312487
7	ammonium	35	9	11	0.840211	2.915448
8	nitrate	18	4	6	0.472045	1.248217
9	iron(2+)	15	4	5	0.576323	1.250634
10	phosphate ion	15	3	4	0.354505	1.040461
11	copper(1+)	11	3	3	0.674061	1.391682

The table presents the thresholds calculated from the 3-fold cross-validation of the UniProt-ICAT-100 dataset. Two thresholds for the softmax layer (SF) and penultimate layer (PU), calculated separately, are presented in the last two columns.

4.5 Baseline: Adding miscellaneous class

As a baseline for handling unseen classes in classification, we implemented the traditional approach of adding an extra class for data samples not included in the current classes. Following the method in [AB22a], we introduced an additional class labeled **miscellaneous** for other unlabeled transport protein sequences. Although the samples in this class vary structurally, the results from this experiment serve as a valuable baseline.

As Section 3.6 describes, building the UniProt-ICAT-100 dataset, classes with fewer than 10 samples were discarded. Here, we reintroduce these discarded samples as a single class labeled **miscellaneous**. However, the sample size for this **miscellaneous** class is significantly larger than that of the other classes in the UniProt-ICAT-100 dataset. To achieve balance, we apply undersampling techniques to reduce the size of the **miscellaneous** class.

Specifically, we use the Edited Nearest Neighbours method [Wil72] to undersample the **miscellaneous** class. This method cleans the class by removing samples close to the decision boundary, ensuring a greater distance between the samples of the **miscellaneous** class and the 12 UniProt-ICAT-100 classes. The undersampled class is then split into training, validation, and test sets. Each subset is combined with its corresponding subsets from the UniProt-ICAT-100 dataset. The TooT-ICAT model (as defined in Section 3.6), trained on this new dataset, serves as the baseline of this work.

4.6 Results

The results for the ICAT transporter detector are presented in Table 16. The model identifies the ICAT transporters with an MCC of 0.941. The results for the open classifier on substrate specificity are presented in Table 17. The first row of results shows the result for TooT-ICAT, a simple closed-world classifier. The second model is TooT-ICAT with an extra class of **miscellaneous** added. The third row shows the results for the TooT-Open-ICAT with a fixed threshold of $\tau = 0.5$ on softmax layer outputs. The last row of the table shows the results for TooT-Open-ICAT with different cross-validation thresholds on softmax layer output for each class as presented in Table 15.

Table 16: Results of the ICAT transporter detector.

Substrate	Trainset	Validation	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
ICAT transporters	2631	658	823	813	38	10	762	0.970	0.955	0.988	0.971	0.941
Non-ICAT transporters	2560	640	800	762	10	38	813	0.970	0.987	0.953	0.969	0.941
Overall	5191	1298	1623	1575	48	48	1575	0.970	0.970	0.970	0.970	0.941

The results for the hold-out test set of the UniProt-ICAT-binary dataset.

Table 17: Results of the hold-out test set of specific substrate predictor.

Model	Trainset	Validation	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
TooT-ICAT	2,631	658	823	786	37	37	9016	0.993	0.955	0.883	0.913	0.944
TooT-ICAT + miscellaneous class (baseline)	2,732	684	855	527	328	328	9,932	0.941	0.616	0.505	0.537	0.620
TooT-Open-ICAT with fixed τ	2,631	658	823	784	39	39	9,837	0.993	0.953	0.844	0.816	0.941
TooT-Open-ICAT with cv τ	2,631	658	823	786	37	37	9,839	0.993	0.955	0.883	0.844	0.944

Results for UniProt-ICAT-100 fine-tuned using weighted focal loss function and 3-fold validation thresholds calculated on the softmax layer

Table 18: Results TooT-Open-ICAT trained with cv threshold.

Substrate	Trainset	Validation	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
Proton	837	209	261	258.0	14.0	3.0	548.0	0.979	0.949	0.989	0.968	0.953
calcium(2+)	555	139	174	168.0	4.0	6.0	645.0	0.988	0.977	0.966	0.971	0.963
potassium(1+)	465	116	145	136.0	7.0	9.0	671.0	0.981	0.951	0.938	0.944	0.933
chloride	293	73	92	86.0	6.0	6.0	725.0	0.98	0.94	0.94	0.94	0.93
sodium(1+)	274	69	86	81.0	2.0	5.0	735.0	0.99	0.98	0.94	0.96	0.95
sulfate	62	16	20	18.0	0.0	2.0	803.0	1.00	1.00	0.90	0.95	0.95
zinc(2+)	51	13	16	16.0	0.0	0.0	807.0	1.00	1.00	1.00	1.00	1.00
ammonium	35	9	11	11.0	0.0	0.0	812.0	1.00	1.00	1.00	1.00	1.00
nitrate	18	4	6	2.0	0.0	4.0	817.0	1.0	1.0	0.3	0.5	0.6
iron(2+)	15	4	5	3.0	0.0	2.0	818.0	1.0	1.0	0.6	0.8	0.8
phosphate ion	15	3	4	4.0	0.0	0.0	819.0	1.0	1.0	1.0	1.0	1.0
copper(1+)	11	3	3	3.0	0.0	0.0	820.0	1.0	1.0	1.0	1.0	1.0
Overall	2631	658	823	786.0	37.0	37.0	9839.0	0.993	0.955	0.883	0.844	0.944

Results of the hold-out test set of UniProt-ICAT-100 fine-tuned using weighted focal loss function and 3-fold validation thresholds for softmax layer

4.7 Discussion

In analyzing the TooT-Open-ICAT model, there are two key conditions for test samples: whether the class of the test sample is included in the training process (or model has *seen* the class of the test sequence), and whether the test sample is experimentally annotated (or it is from a *known* class of transporters). Classes that have annotated sequences and were included in the training of TooT-Open-ICAT are called *known* and *seen* classes. Classes that are experimentally annotated but were excluded from training due to their small size are referred to as *known* and *unseen* classes. Finally, classes that are not annotated and therefore not included in the training are considered *unknown* and *unseen* (Figure 19).

4.7.1 Micellaneous class effect

Adding an extra class with various samples is a common approach to handling unseen samples. In traditional models, this extra class, here referred to as **miscellaneous**, is added to the dataset with the expectation that unseen samples will be assigned to it during testing. Here, we examine the accuracy of this approach and its impact on the model’s performance.

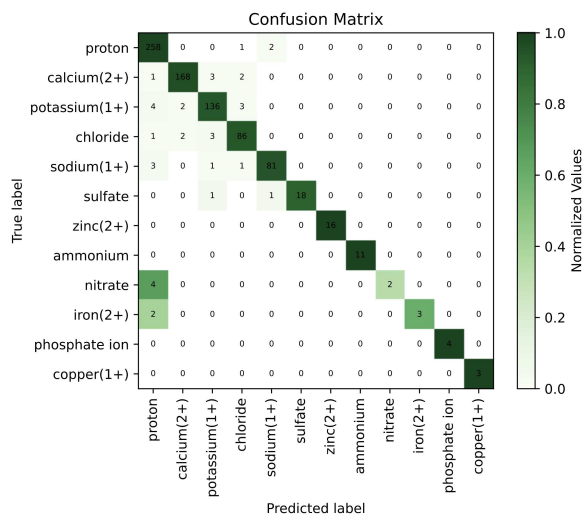
As Table 17 shows, adding the miscellaneous class to the UniProt-ICAT-100 causes a significant performance drop in the TooT-ICAT model, with the MCC decreasing from

	Known	Unknown
seen	Annotated sequences from UniProt included in UP-ICAT dataset	NA
Unseen	Annotated sequences from UniProt NOT included in UP-ICAT dataset (low numbers)	Unannotated <i>de novo</i> protein sequences

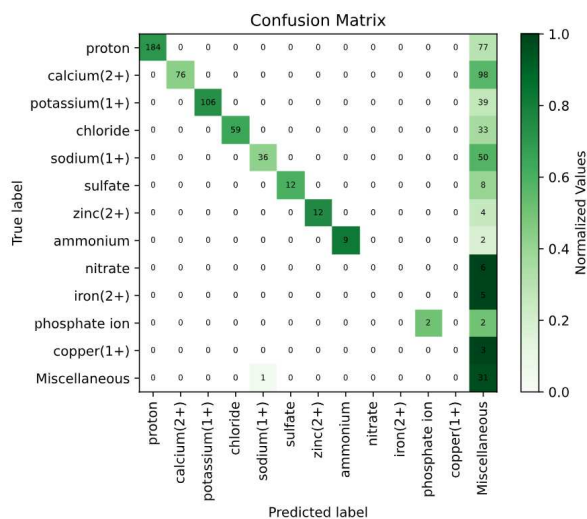
Figure 19: The test sample groups included in TooT-Open-ICAT analysis.

Samples from classes in UniProt-ICAT-100 are experimentally annotated and included in the training of TooT-Open-ICAT referred to as seen and known test samples. There are other annotated samples in UniProt-KB but not included in the training process of TooT-Open-ICAT due to their class size referred to as known and unseen samples. The last group of test samples is *de novo* sequences unannotated in UniProt-KB and not included in training called unknown and unseen samples. There are no unannotated samples included in the training (seen and unknown). The model detects unseen-known and unseen-unknown proteins as a novel class.

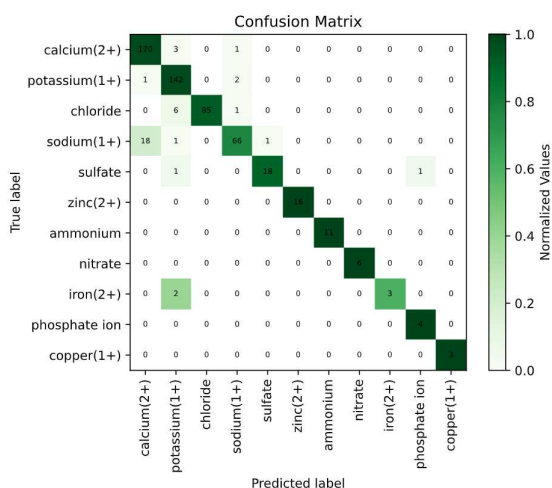
0.944 to 0.620. Figure 20b elaborates the issue, showing an excessive number of False Positives for the **miscellaneous** class. This result is consistent with findings from previous studies [AB22a], where the **miscellaneous** class consistently showed lower performance metrics compared to other classes. Despite using the Edited Nearest Neighbours method [Wil72] to minimize distribution overlap between this class and known classes, our results still reflect the same problem. This experiment demonstrates the trade-off of adding an extra class: while it offers a potential label for unknown samples, it reduces the model’s performance on other seen classes.



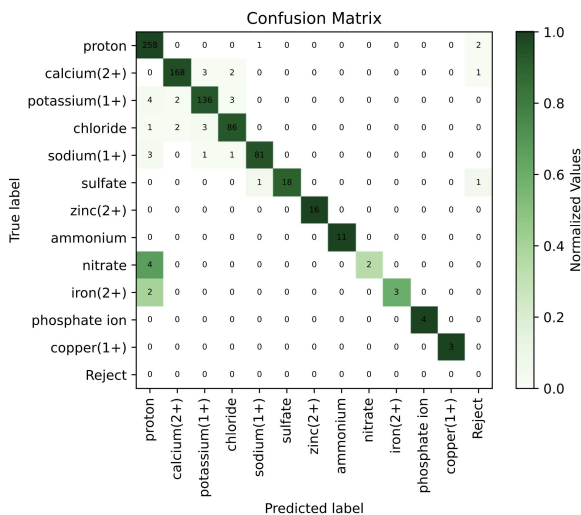
(a) Results for TooT-ICAT.



(b) Results for TooT-ICAT + miscellaneous.



(c) Results for TooT-ICAT without proton.



(d) Results for TooT-Open-ICAT.

Figure 20: Confusion matrix for the hold-out test set results.

The confusion matrices of performance for four models of inorganic ion prediction on the UniProt-ICAT-100 hold-out dataset are presented. In each subfigure, the x-axis represents the true labels of the test samples, and the y-axis shows the labels predicted by the model. The values in the table represent the number of test samples, but the colors are based on normalized values, with darker shades of green indicating higher normalized values. Normalization is performed by dividing the value in each cell by the total number of test samples for the corresponding class in the row.

4.7.2 Imbalanced classes in open-classification

The second subject of investigation is the impact of open classification on model performance in imbalanced datasets. As shown in Figure 20a, test samples from smaller classes, such as **iron(2+)** and **nitrate**, are frequently misclassified into the majority class, **proton**. For example, Figure 21 illustrates that several **nitrate** test samples exhibit high activation in the **proton** column, leading to false positives (FP) for **proton** and false negatives (FN) for **nitrate**. Similarly, **iron(2+)** samples often show high activation in the **potassium** column instead of their correct class, further underscoring the challenges posed by dataset imbalance.

This misclassification is explored in detail. Results show that the majority class, **proton**, has the highest number of false positives, with 14 FPs. As depicted in Figure 20d, 4 out of 6 test samples of the minority class **nitrate** are misclassified as **proton**. Similarly, 2 out of 5 **iron(2+)** test samples are incorrectly predicted as **proton**. Together, these two classes contribute to 42% of the false positives for **proton**. More importantly, this misclassification significantly affects model performance on these minority classes, resulting in the lowest class Matthews Correlation Coefficients (MCC) of 0.6 for **nitrate** and 0.8 for **iron(2+)**.

To further investigate the role of the majority class in false negative predictions for minority class test samples, an ablation study is conducted. In this study, the majority class, **proton**, is excluded from the UniProt-ICAT-100 dataset during both training and testing, and the classification process is repeated. The results, shown in Figure 20c, demonstrate that **nitrate** test samples are correctly classified in the absence of **proton**. However, **iron(2+)** test samples are misclassified as **potassium**, the second most likely option, rather than being correctly identified as **iron(2+)**.

The goal is to determine whether applying the rejection option in open classification, as implemented in the TooT-Open-ICAT model, reduces the misclassification of sequences of minority classes. This requires careful evaluation of the computed threshold and the network layer where the threshold is applied. Here, we investigate the model’s performance in handling imbalanced classification using the rejection option, with thresholds applied at two different layers: the penultimate layer and the softmax layer.

In the softmax layer, the denominator sums over all classes to ensure the probabilities total 1. However, unknown classes will appear during testing in an open set, making this approach ineffective. Therefore, the penultimate layer is selected to apply the threshold for each class separately, with thresholds calculated from the true positive (TP) values of

the validation set. Figure 21 presents the output values of the test set for the penultimate layer. This figure also confirms the penultimate layer values resulted in the misclassification of the smaller classes to the majority, showing higher values on **proton** for the samples of **nitrate** and **iron(2+)**.

Comparing fixed thresholds with cross-validation (CV) calculated thresholds shows that CV-calculated thresholds depend on the composition of the validation set. This dependency is especially pronounced for small classes in our dataset. For instance, in small classes like **copper(1+)** the rejection threshold is determined based on only three samples in the validation set, making the model more prone to errors.

According to the results, thresholding the softmax or penultimate layers does not significantly improve the misclassification problem. This is because thresholding the last two layers offers no benefit once the model has already assigned high values to an incorrect label. This issue persists whether the threshold is applied to softmax probabilities or penultimate layer values.

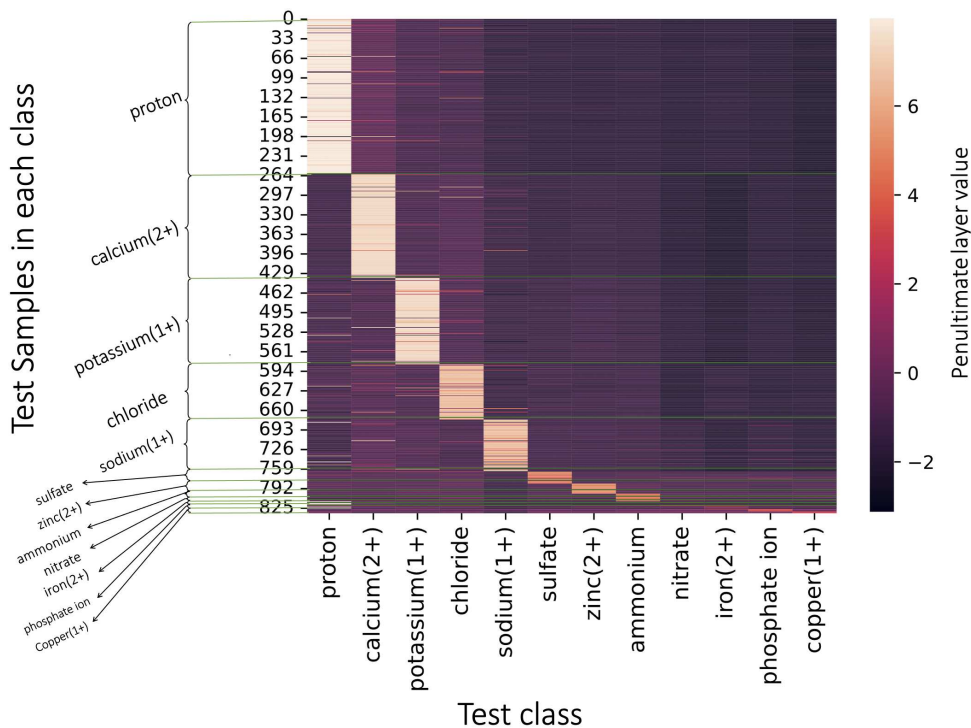


Figure 21: Heatmap of prediction values for penultimate layer for each test sample. The Y-axis shows the test samples in the test set of UniProt-ICAT-100 and the X-axis shows the value of the penultimate layer for each class. Lighter colors show higher values. Green horizontal lines separate classes of the test samples.

4.7.3 Novel class detection in open-classification

The primary objective of developing the TooT-Open-ICAT model is to effectively detect the novel class including *de novo* protein samples from a machine learning perspective. The model is applied to the transportomes of five model organisms, spanning diverse biological groups: a bacterium (*Escherichia coli* K-12 (Es)), a fungus (*Saccharomyces cerevisiae* (Sc)), an insect (*Drosophila melanogaster* (Dm)), a mammal (*Homo sapiens* (Hs)), and a plant (*Arabidopsis thaliana* (At)). Because no reference annotations or manual curation are available for these sequences, fully evaluating the model’s performance on transportomes remains challenging. However, the analysis of its predictions, as summarized in Table 45, offers valuable insights into the model’s capabilities.

The first insight is the consistent ratio of proteins predicted as ICAT transporters to non-ICAT transporters across 5 model organisms. The second observation is the prediction of zero calcium transporters in *Escherichia coli*. According to UniProt-KB, no proteins in the transportome of *E. coli* are annotated as calcium transporters, despite evidence that calcium transport occurs in this organism. While this prediction is consistent with current annotations, caution is warranted since it reflects only one class of transporters. Consequently, although TooT-Open-ICAT’s performance appears to align with annotated data, further investigation is required to confirm these predictions, particularly for under-annotated transporter classes.

4.8 Conclusion

This study addresses the challenge of identifying inorganic ion transport proteins and predicting their specific substrates within the transportome using an open-world classification method. We developed a two-component model, TooT-Open-ICAT, consisting of an ICAT protein detector to identify inorganic ion transport proteins and a specific substrate predictor to classify their substrates or reject them as other inorganic ion transporters. The rejection is based on a threshold τ , transitioning the model from a closed-world to an open-world classification approach, better reflecting real transportome data. Both components achieved a high MCC of 0.94, effectively distinguishing ICAT transporters from non-ICAT proteins and predicting their specific substrates.

The traditional method of adding an extra class to manage unseen classes during testing was also examined. It revealed a trade-off, as it reduced the model’s performance on seen classes, emphasizing the challenge of balancing seen and unseen class management

(Section 4.7.1). In contrast, the open-classification approach, which uses thresholding in network layers, shows promise for handling unseen classes. However, it requires careful attention to two key factors: the representativeness of training datasets in capturing transportome diversity and the precise calculation of thresholds for small, imbalanced classes (Section 4.7.2). Compared to the closed-world classification model, TooT-ICAT, the rejection mechanism in open-world classification did not significantly reduce misclassifications in the imbalanced dataset (Section 4.7.2).

This research explores rejection using thresholding two final layers of the network: the softmax layer and the penultimate layer. However, the ProtBERT encoding outputs, which occur before the fully connected layer, could also be considered for thresholding. This layer may offer advantages by applying thresholds to the vectors before they enter the classifier, potentially addressing the issues discussed in Section 4.7.2. However, this option requires thorough investigation, including the use of Out-Of-Distribution (OOD) learning methods to analyze test sample distributions in relation to the training dataset. This approach involves calculating a threshold by measuring the divergence between the training and validation set distributions, then using this threshold to reject test samples [YZLL21]. Future work will focus on applying OOD learning methods to improve open classification and better handle unseen classes.

This research focuses on single-substrate ion transporters, acknowledging that some biological processes involve multiple ions. An important direction for future work is expanding the dataset to include multi-substrate transport proteins. Addressing multilabel classification problems offers various methodological options to accommodate these complexities. In the current TooT-Open-ICAT model, classification is performed using a softmax layer. To modify this approach, instead of relying on the label with the highest probability, the model could be adapted to retrieve the top m probabilities as multiple labels. Exploring multi-label ion transporters represents a promising area for future research.

Chapter 5

TooT-Triplet-SPEC: One-shot classification of transmembrane transport proteins

5.1 Introduction

This chapter evaluates the effectiveness of a distance-based model in the one-shot classification of transmembrane transport proteins in the presence of imbalanced classes. This study evaluates the effect of the triplet sampling strategy and loss function on a ProtBERT-BFD model followed by a K-NN classifier. The encoding space of the ProtBERT-BFD model (Section 2.3) has been modified using a triplet loss function and sampling strategy to reduce the intra-class sample distances while decreasing the inter-class distances [SKP15]. Two sampling strategies, online and offline sampling, along with three different distance metrics, Manhattan, Euclidean, and Cosine similarity, have been incorporated and examined both in encoding space modification and the K-NN classification (see Sections 2.8.1, 2.8.2, and 2.9.1). To address the computational challenges associated with the models, the Low-Rank Adaption of Large Language Models (LoRa) algorithm has been implemented across various layers (described in Section 5.6.1).

The following sections include stating the problem in detail (Section 5.2), and the methodology applied to solve the problem (Section 5.3) explaining the online and offline sampling strategies, training with LoRa algorithm, dataset, and classification. The results of each model are presented in Section 5.7 followed by the discussion points about sampling strategies, interpreting the sample distances, the effect of imbalance in training, and

the relationship between binding sites and attention heads in Section 5.8. The model’s performance is further validated on additional datasets in Section 5.9. The last section, Section 5.10 provides the conclusion of the chapter. The dataset and code for this project are publicly available on GitHub (<https://github.com/simaataei/TooT-Triplet-BERT>)

5.2 Problem statement: specific substrate prediction

The problem of specific substrate prediction is a supervised learning problem on an imbalanced dataset with a long-tailed sample distribution format. We define this problem as a classification problem with two constraints on the dataset. Given a labeled dataset of n transmembrane transport protein sequences as $D = \{(p_1, l_1), (p_2, l_2), (p_3, l_3), \dots, (p_n, l_n)\}$, p_i representing the primary structure of each protein, as a chain of amino acids, associated to a label, l_i , in the label set L of seen classes; build a model, $f(p)$, to classify samples of the test set, D_{test} associated to a label set as $L = \{l_1, l_2, l_3, \dots, l_n\}$.

Here, the number of samples associated with class i is shown by n_i . There are two specific conditions to define on the dataset: imbalanced and minority classes. The dataset suffers the *severe imbalanced problem* (as defined in Section 2.5) when $\exists l_i$ and $\exists l_j$ in the dataset where a number of samples associated with l_j , is at least 100 times greater than the samples associated with l_i [Kra16]. In other words,

$$\exists l_i, l_j \in L \text{ where } n_j \geq 100 \times n_i \quad (1)$$

In this study, the term *minority class* in the dataset is defined as a class with less than 10 samples. The threshold of 10 is selected based on preliminary studies that indicate the faulty performance of general classification methods on small classes. So,

$$\text{class } c_i \text{ is a minority class} \iff n_i < 10$$

5.3 Methodology

Definition 5.3.1. A *triplet* is a set of 3 samples: a sample from any class as an anchor, a sample from the same class as the anchor (positive), and a sample from a different class than the anchor (negative)

To overcome both issues of the dataset, TooT-Triplet aims to separate the minority classes from the majority in an n -dimensional space by reducing the intra-class and increasing inter-class distances. This model relies on triplets of anchor, positive and negative

in the dataset. Each sample in the dataset is considered at least once as an anchor during the training process, matched with a positive sample with the same label as the anchor and a negative sample with different labels from the label. The ProtBERT-BFD model weights are updated based on these criteria during the training process to minimize the anchor-positive distance and maximize the anchor-negative distance.

The sample encodings are updated as the model weights are adjusted in each epoch. The final encodings of samples in the training set are utilized to train a K-Nearest Neighbor model, which is subsequently tested using the final encodings of the test set. This process is elaborated in Section 5.5. The distance-based classification and efficient performance of the K-NN algorithm advantaged this algorithm over other classification methods.

Definition 5.3.2. *The **manhattan distance** between two vectors is the sum of the absolute differences between their corresponding components. It is also known as the L_1 -norm of the difference between the vectors, as defined in Equation 3, Section 2.8.2.*

Definition 5.3.3. *The **euclidean distance** between two vectors is the square root of the sum of the squared differences of their corresponding components. It is also known as the L_2 -norm of the difference between the vectors, as defined in Equation 3, Section 2.8.2.*

Definition 5.3.4. *The **cosine distance** between two vectors is calculated as the complement of their normalized directional similarity, where the similarity is determined by the cosine of the angle between the vectors (Equation 4, Section 2.8.2).*

The distance metric plays a crucial role in both the triplet loss calculation and the training and testing of the K-NN model. However, it's known that evaluating vector similarities using metric distances in high dimensions can be challenging [BGRS99]. The curse of dimensionality related to distance metrics, particularly focusing on L_k norm distance metrics (Equation 3), is discussed in [AHK01]. The research suggests that the L_1 norm or Manhattan distance is often preferred over the L_2 distance or Euclidean distance in high dimensions. In this study, we utilize both Manhattan and Euclidean metrics separately to train and test models, evaluating their effectiveness in distinguishing similarity in 1024-dimensional ProtBERT-BFD encodings. Additionally, we incorporate cosine similarity as a third distance metric to train and evaluate the models. All three distance metrics are introduced in the background chapter, Section 2.8.2.

Definition 5.3.5. *In machine learning, **sampling** refers to the process of selecting a subset of data points from a larger dataset to use for training, validation, or testing the model.*

The model’s performance is directly influenced by the choice of triplets during the training process, highlighting the main challenge of this method: selecting the strategy to generate the triplets. Here, two sampling strategies, offline and online sampling, are implemented to select the triplets and train two separate models. In the offline sampling process, the triplets are fixed during training and selected randomly, as described in Section 5.3.1. On the other hand, the online sampling strategy selects a new set of *hard* triplets in each epoch during training. A *hard* triplet includes an anchor (any sample in the dataset), a *hard positive* (a sample with the same label farthest from the anchor), and a *hard negative* (a sample with a different label closest to the anchor), as defined in Section 2.8.1. The online sampling strategy is detailed in Section 5.3.2.

5.3.1 Offline sampling strategy

In the offline sampling strategy, the triplets of anchor, negative, and positive are sampled randomly from the training set. In this approach, every sample in the training set is once selected as the anchor. The positive sample for this anchor is selected randomly from all the samples of the training set with the same label as the anchor. Then the positive and negative samples for this anchor are selected randomly from the remaining samples in the dataset with the same label and different labels from the anchor respectively. Figure 22 presents training process using the offline sampling strategy

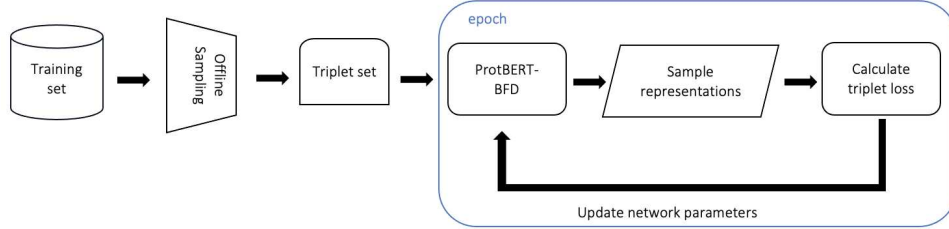


Figure 22: Training process in TooT-Triplet with offline sampling strategy.

The triplets are sampled at the beginning of the training process. The weights of the Prot-BERT-BFD are updated based on the triplet loss calculated from the fixed set of Triplets. One epoch in each training process is presented by the blue rectangle.

5.3.2 Online sampling strategy

In online sampling, the strategy is to train the network using the *hard positive* and *hard negative* samples. In this sampling technique, for each sample in the training set referred

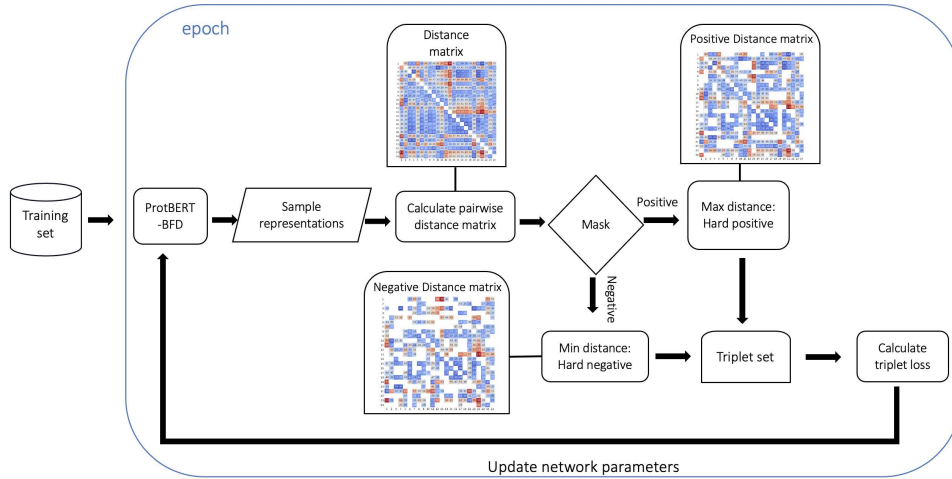


Figure 23: Training process in TooT-Triplet with online sampling strategy.

The triplet sampling is dynamically performed at the beginning of each epoch. The sequence representation is extracted from the ProtBERT-BFD model and the pairwise distance matrix of the samples is calculated based on these representations. A positive and negative mask is applied for each random anchor. A positive mask ignores samples with different labels from the selected anchor and a negative mask ignores the samples with the same labels as the anchor. After building the full triplet set in one epoch, the triplet loss function is calculated and ProtBERT-BFD is updated based on the loss. One epoch in each training process is presented by the blue rectangle.

to as the anchor, the *hard positive* sample is selected as the sample with the same label as the anchor with the maximum distance from it. Also, a *hard negative* sample for an anchor is the sample with a different label with the minimum distance from the anchor (definition details in Section 2.8.1). The triplets are selected during the training at the beginning of each epoch from the sample encodings updated in the previous epoch.

For each epoch training process, first, the *hard* triplets are sampled as described, then the triplet loss function is calculated for each batch of triplets, and the model weights are updated at the end of each epoch based on this loss. The online sampling strategy produces different triplets for training on each epoch. This is because the distance matrix is updated based on the new sequence encodings extracted from the previous epoch model. The process is presented in Figure 23.

5.3.3 Training

Training of the TooT-Triplet model involves two main components. First, the protein language model is updated to separate the class representations in a n-dimensional space,

focusing on consistent representation for minority classes. This language model then is used to extract encodings for protein sequences in the dataset. The second component is a K-NN model trained based on the final encodings of protein sequences, classifying each sequence based on their corresponding substrates.

TooT-Triplet uses the Prot-BERT-BFD protein language model to extract encodings of protein sequences. Based on previous research, this model is widely used due to its reliable BERT-based structure and extensive pretraining data, BFD-100, as described in Section 2.3. For each protein sequence, Prot-BERT-BFD applies 30 layers of self-attention mechanisms using 16 attention heads. Here, the encoding of each sequence is extracted from the CLS token from the last layer of this network.

Triplets of anchor, positive, and negative samples are used in the process of training. These triplets are either fixed triplets sampled randomly at the beginning of the training in offline sampling strategy (Section 5.3.1) or selected dynamically at the beginning of each epoch, and change based on the previous epoch encodings, in online sampling (Section 5.3.2). The triplet loss function is applied to update the Prot-BERT-BFD model weights. As described in Section 2.8.1 this loss function calculates the anchor-positive and anchor-negative distances on each epoch. This function is implemented using `TripletMarginWithDistanceLoss` class from the `Torch` package in Python with three custom distance metrics of Manhattan, Euclidean, and Cosine similarity [PGM⁺19].

The models are developed in Python, version 3.9.1, using the `PyTorch` and `Transformers` packages [PGM⁺19, WDS⁺20]. The hyperparameter optimization has been implemented using the `Optuna` framework [ASY⁺19]. `Optuna` is an MIT-licensed, state-of-the-art optimization framework designed for hyperparameter tuning in machine learning models, employing an efficient, and flexible approach to search spaces to streamline and enhance the performance of model training and evaluation processes. A study with 20 different trials is implemented to find the best hyperparameters for the TooT-Triplet model. The search space has been presented in Table 19.

All offline models are trained and tested on Concordia University computing cluster **High-Performance Computing (HPC) Facility (Speed)** using one **NVIDIA Tesla P6 GPU**, with 16 GB of memory. The online models are trained and tested on Concordia University computing cluster, **Virya** using 4 GPUs of **NVIDIA Tesla V100-PCIE** (in parallel) with 32 GB of memory on each.

To complete the classification task, the CLS token from the last layer encodings of the fine-tuned models are extracted. These encodings along with the substrate labels are used

Table 19: Hyperparameter search space for TooT-Triplet with offline sampling strategy.

Hyperparameter	Suggested Values
Learning Rate	1e-6 to 1e-2
Number of epochs	5 to 100
Optimizer	AdamW, SGD
Distance Metric (in Triplet Loss Function)	Manhattan, Euclidean, Cosine

The search space is explored with 20 trials to find the optimal hyperparameters including the learning rate, number of epochs, optimizer, and the distance metric implemented in calculating triplet loss function for the training process of TooT-Triplet with offline sampling strategy.

to train a K-NN model implemented using `Scikit-learn` package in Python [PVG⁺11]. A 1-Nearest Neighbor algorithm is applied to find the predictions for each test sample (described in Section 5.5)

5.3.4 Baseline

Three baseline models are proposed to evaluate the effectiveness of the TooT-Triplet model. The TooT-Triplet approach aims to adjust the model weights to minimize intra-class distances and maximize inter-class distances in the embedding space. To assess whether the model has successfully modified the embedding space, we use the pre-trained ProtBERT-BFD model with frozen weights as an encoder to compute protein sequence encodings before any weight modifications. This approach is explained in Section 2.3. To evaluate the effect of different distance metrics on classification performance, the sequence encodings obtained from the frozen ProtBERT-BFD model are used to train three K-nearest neighbors (K-NN) models. Each K-NN model is trained using one of the distance metrics in TooT-Triplet: Manhattan, Euclidean, and Cosine, with K set to 1. These K-NN models serve as baselines: Baseline 1, Baseline 2, and Baseline 3, respectively.

5.4 Dataset

The main dataset utilized in the training and testing process of the TooT-Triplet project is the UniProt-SPEC-100 dataset, as described in Section 3.5.2. This dataset includes 4,455 non-identical protein sequences from 96 different classes of specific substrates. It is long-tailed and imbalanced, with the majority class, `proton`, comprising 1,226 samples, while 23 minority classes have only 3 samples each (see Table 43 in Appendix B).

To accommodate the limited size of the minority classes, the dataset is randomly stratified and divided into training and test sets with a ratio of 0.66 to 0.33. This results in each minority class having two samples in the training set and one sample in the test set. The same ratio is applied to divide the dataset into three folds for cross-validation.

To evaluate the effect of the majority class **proton**, a second dataset is generated from UniProt-SPEC-100, with the same samples and classes except for the class of **proton**. The UniProt-SPEC-100-ProtonExcluded dataset has 3,229 non-identical protein sequences from 95 classes of specific substrates. The dataset is divided into training and test sets with the same method described for the UniProt-SPEC-100 dataset.

5.5 Classification

The classification process relies on the distances between samples in the encoding space, aligning with the training process’s objective of minimizing intra-class and maximizing inter-class distances. The k-nearest neighbor algorithm is employed as a distance-based classification method to predict the labels of sequences in the test set. During the prediction phase, the process begins by extracting sample encodings from the final ProtBERT-BFD model for the training set. Subsequently, a K-Nearest Neighbor model is trained using these encodings. In the next step, sample encodings for the test set are extracted from the same ProtBERT-BFD model. Finally, the K-NN model predicts labels for the test sequences by considering their K nearest neighbors in the encoding space.

Setting the number of neighbors, K, in the K-NN algorithm is crucial depending on the dataset and the specific problem. In the UniProt-SPEC-100 dataset, the minority classes consist of only 3 samples overall, with 2 samples in the training set and 1 sample in the test set. Consequently, choosing K greater than 2 could lead to failures in predicting minority class instances. On the other hand, selecting K=2 often results in an equal number of different votes among the nearest neighbors for test samples, which can cause difficulties in decision-making. Therefore, to ensure accurate predictions across both minority and majority classes, K=1 is chosen as the optimal value in this scenario.

5.6 Challenges

Using the online sampling strategy introduces both memory and time computational overheads into the training process. The memory limitation challenge arises from the

need to keep at least three samples with gradients in memory to calculate the triplet loss and backward propagate the calculated loss to update the network weights. The time overhead stems from the need to calculate all the sequence encodings and their pairwise distances at each epoch. Both these challenges and their implemented solutions are explored in the following sections.

5.6.1 Memory limitaion

One of the downsides of the powerful Prot-BERT-BFD model is its large size. This model has 420 million parameters. During the fine-tuning, in addition to input tensors, and model parameters, the loss tensor, intermediate activation, and gradients for each sequence need to be loaded into GPU memory for the model to compute forward and backward passes. Given the limited resources available for this research, only one sequence fits in memory for computation per training epoch. This poses a problem, as the TooT-Triplet requires batches of at least three sequences $\langle anchor, positive, and negative \rangle$ in memory simultaneously to perform the forward pass, calculate the triplet loss, and backpropagate.

To overcome this challenge, we employed LoRA (Low-Rank Adaptation of Large Language Models) [HSW⁺22]. LoRA works by freezing the pre-trained model weights and injecting trainable *rank* decomposition matrices into each layer of the Transformer architecture. Specifically, LoRA introduces two low-rank matrices, A and B, to approximate the original high-dimensional weight matrices, where A is used to project the input down to a lower dimension and B is used to project it back up to the original dimension. This reduces the number of trainable parameters in LLMs for downstream tasks.

LoRA introduces three key parameters: *rank*, *alpha*, and *dropout*. The rank determines the dimensionality of the low-rank matrices A and B, which are used to decompose the original weight matrices. By choosing a smaller *rank*, we reduce the number of parameters and computational cost. *Alpha* is a scaling factor applied to these low-rank matrices, controlling the magnitude of their updates and ensuring they do not disrupt the pre-trained model’s stability. *Dropout* is a regularization technique that helps prevent overfitting by randomly setting a fraction of the input units to zero during training, promoting robustness and generalization.

Here, we have applied the LoRA algorithm to the Key, Query, Value, and the encodings matrices on each attention head with *rank* = 1. The *dropout* and *alpha* are set to 0.1 and 1 based on the default recommendations of the package. Applying LoRA to the Prot-BERT-BFD model decreased the number of trainable parameters from 420,115,456

to 184,320. The configuration details are presented in Table 20. This reduction allows the model to increase the batch size to 31 sequences in memory per epoch, enabling the triplet loss function to work effectively.

Table 20: LoRA configuration details.

Task Type	Rank	Alpha	Dropout	Target Modules
Feature Extraction	1	1	0.1	encoding, query, key, value

Low-Rank Adaptation of Large Language Models (LoRa) is implemented in the training process of the TooT-Triplet model with parameter values as presented in the table.

5.6.2 Computational time limitation

The second challenge is the computational time required for the training. The online sampling strategy requires, first, calculating a new encoding for each sample in the training set on each epoch from the model updated on the previous epoch. Second, the pairwise distances of the samples are calculated on each epoch. Lastly, the *hard positive* and *negative* samples are retrieved based on the distances for each anchor. Normally, implementing this strategy calculates n^3 possible triplets and their distances, where n is the number of samples in the training set. However, only a subset of such possible triplets are valid as anchor, positive, and negative. The triplets of $\langle anchor, positive, positive \rangle$ and $\langle anchor, negative, negative \rangle$ are invalid according to the loss function.

To reduce the computational time overhead of the model, a distance matrix of all the samples in the training set is calculated at the beginning of each epoch. This matrix computes the pairwise distances using the dot product of the full training set encoding matrices. This implementation avoids the use of nested loops and reduces the computational time for distance calculation. Additionally, positive and negative masks are designed to find the *hard* triplets in the dataset for each sample. The positive mask discards the samples with different labels from the anchor, enabling the method to select the sample with the maximum distance as the *hard positive* sample for the anchor. Similarly, the negative mask retains all the negative samples to find the one with the minimum distance from the anchor in the distance matrix as the *hard negative* (Figure 23). These masks disregard invalid triplets of $\langle anchor, positive, positive \rangle$ and $\langle anchor, negative, negative \rangle$.

5.7 Results

This section presents the experiment results for the TooT-Triplet model. In Section 5.7.1, the results of the models trained using the offline sampling strategy with three different distance metrics of Manhattan, Euclidean, and Cosine similarity are presented. Section 5.7.2 presents the detailed results of the model trained with an online sampling strategy using Manhattan distance. The metrics of Accuracy, Precision, Recall, F1-Score, and Mathew’s Correlation Coefficient (MCC) (details in Section 2.11) are utilized to evaluate the models.

5.7.1 Offline Sampling Results

The hold-out test set prediction results suggest that the Manhattan distance outperforms other distance metrics in classifying sequences using the 1-Nearest Neighbor algorithm. Table 21 and Figure 24 present the results for models trained and tested on the UniProt-SPEC-100 dataset for 90 epochs using three distance metrics: Manhattan, Euclidean, and Cosine. The Manhattan distance outperforms the Euclidean and Cosine distance metrics in Precision, Recall, and Matthew’s Correlation Coefficient (MCC). However, the F1-score comparison shows closely matched values for Euclidean and Manhattan, with Manhattan having a slight advantage.

Table 21: TooT-Triplet (offline) results on hold-out test set.

Distance metric	Accuracy	Precision	Recall	F1-Score	MCC
Manhattan	0.998	0.910	0.756	0.712	0.893
Euclidean	0.998	0.906	0.744	0.715	0.889
Cosine	0.998	0.894	0.716	0.683	0.875

TooT-Triplet is trained with three different distance metrics classified with a 1-Nearest Neighbor algorithm. The table presents the results for the models trained with each distance metric. The presented distance metric is used in both the triplet loss function and K-NN algorithm training. Bold fonts represent the numerically largest entry.

To further investigate the results differences using three distance metrics, the one-way ANOVA test has been performed on 3-fold cross-validation test results using all distance metrics [SW89, BHH⁺78]. Table 22 presents the p-value calculated. The p-values indicate that there is no statistically significant difference in model performance, as all values exceed 0.05, failing to reject the null hypothesis.

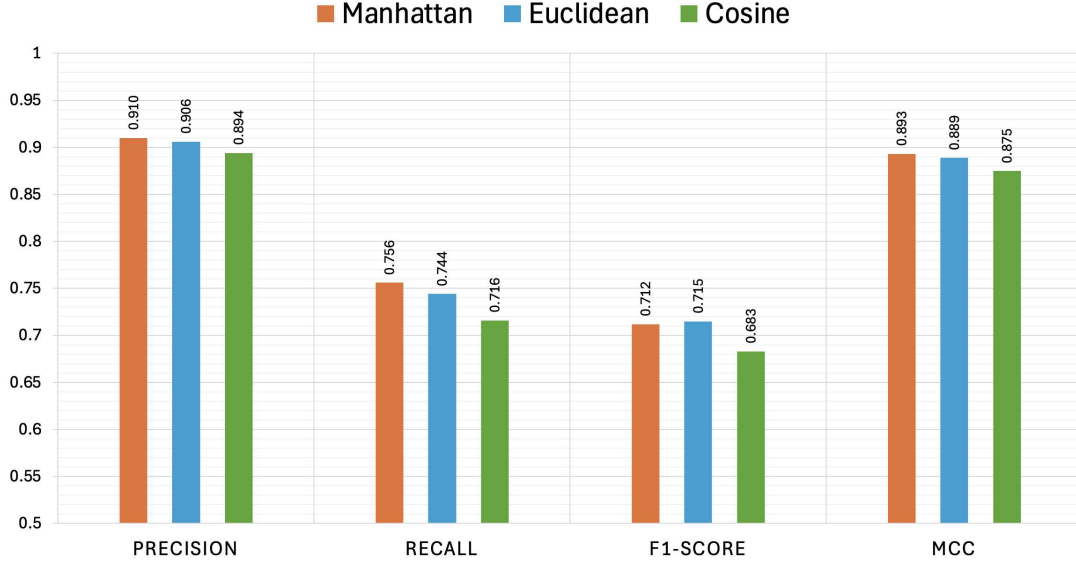


Figure 24: TooT-Triplet results for the hold-out test set trained on three distance metrics. TooT-Triplet is trained with three different distance metrics classified with a 1-Nearest Neighbor algorithm. The figure presents the results of three models trained with each distance metric on the hold-out test set of UniProt-SPEC-100. The presented distance metric is used in both triplet loss function and K-NN algorithm training.

Table 22: TooT-Triplet results for 3-fold cross-validation.

Metric	Accuracy	Precision	Recall	F1-score	MCC
Manhattan	0.998 ± 0.000	0.883 ± 0.022	0.732 ± 0.020	0.701 ± 0.018	0.861 ± 0.027
Euclidean	0.998 ± 0.000	0.885 ± 0.022	0.722 ± 0.027	0.695 ± 0.019	0.863 ± 0.027
Cosine	0.997 ± 0.001	0.864 ± 0.032	0.681 ± 0.036	0.653 ± 0.030	0.840 ± 0.039
p-value	4.921e-01	7.052e-01	2.487e-01	1.517e-01	7.159e-01

TooT-Triplet is trained with three different distance metrics classified with a 1-Nearest Neighbor algorithm. The table presents the results of 3-fold cross-validation for the models trained with each distance metric. The presented distance metric is used in both triplet loss function and K-NN algorithm training. The last row presents the p-value calculated using the ANOVA algorithm to assess the statistical significance of the metric values over 3-fold results. In this test P-values smaller than 0.05 present a statistical significance.

Figure 25 presents the evaluation metrics during the training process for the best TooT-Triplet model with the offline sampling strategy. This model is trained using the Triplet loss function with the Manhattan distance metric and AdamW optimizer with a learning rate of $1e-5$ for 90 epochs. Table 22 presents the results of 3-fold cross-validation for this model. As the table suggests this model reaches the average MCC of 0.874

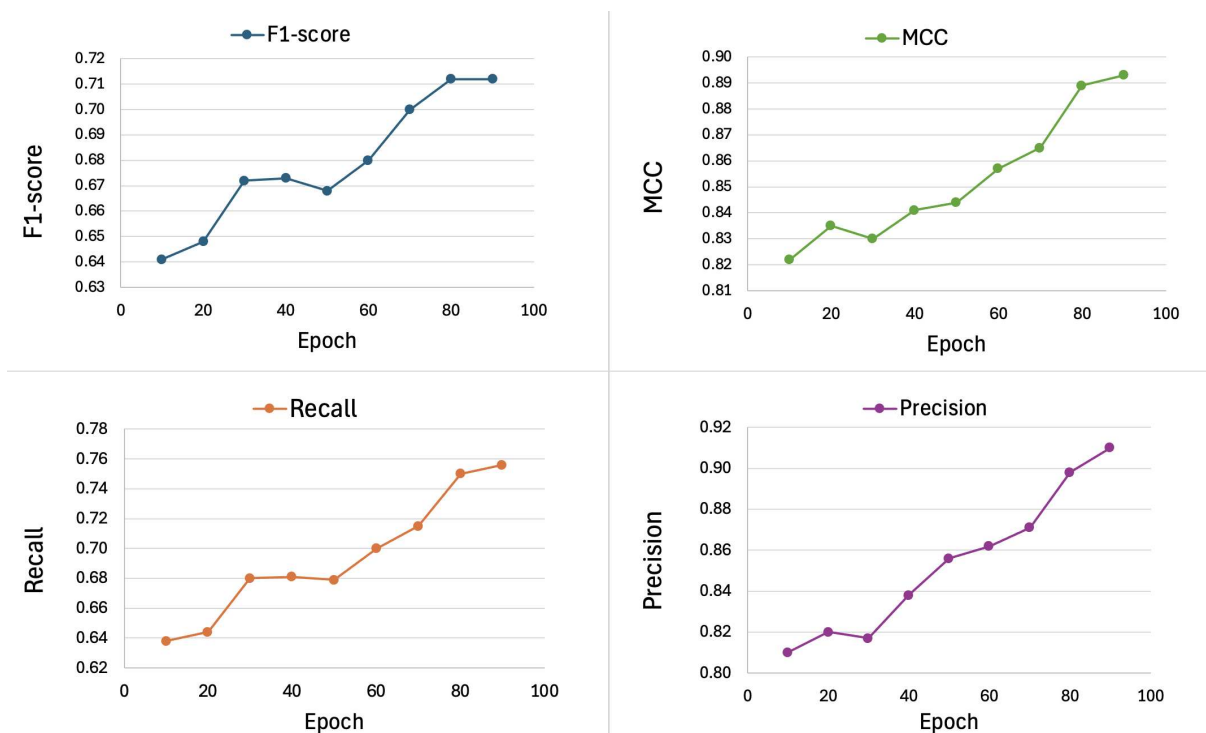


Figure 25: Evaluation in training TooT-Triplet using Manhattan distance.

The reported values for four metrics of Precision, Recall, F1-score, and MCC are presented during the training process of TooT-Triplet from epoch 10 to epoch 90.

with 2.24e-02 standard deviation. Table 23 presents the detailed results for each class of substrates predicted by this model.

Table 23: Hold-out test results for TooT-Triplet: offline sampling, 90 epochs, Manhattan.

Substrate	Train	Test	TP	FP	FN	TN	Acc	Prec	Recall	F1	MCC
proton	818	408	383	17	26	1059	0.971	0.958	0.936	0.947	0.927
calcium(2+)	577	288	280	10	8	1187	0.988	0.966	0.972	0.969	0.961
potassium(1+)	484	242	230	15	12	1228	0.982	0.939	0.950	0.945	0.934
chloride	294	148	141	11	6	1327	0.989	0.928	0.959	0.943	0.937
sodium(1+)	177	89	79	7	10	1389	0.99	0.92	0.89	0.90	0.90
zinc(2+)	53	27	22	3	4	1456	0.99	0.88	0.85	0.86	0.86
sulfate	45	22	16	6	6	1457	0.99	0.73	0.73	0.73	0.72
ammonium	35	17	18	3	0	1464	1.00	0.86	1.00	0.92	0.93
water	29	14	15	0	0	1470	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	25	12	8	5	4	1468	0.99	0.61	0.67	0.64	0.64
protein	19	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
retinol	18	9	7	2	2	1474	1.0	0.8	0.8	0.8	0.8
glutathione derivative	18	9	9	0	0	1476	1.0	1.0	1.0	1.0	1.0
nitrate	17	9	7	0	2	1476	1.0	1.0	0.8	0.9	0.9
iron(2+)	16	8	5	1	3	1476	1.0	0.8	0.6	0.7	0.7
pyruvate	13	7	5	0	2	1478	1.0	1.0	0.7	0.8	0.8

Continued on next page

Table 23: Hold-out test results for TooT-TripleT: offline sampling, 90 epochs, Manhattan.

Substrate	Train	Test	TP	FP	FN	TN	Acc	Prec	Recall	F1	MCC
ATP(4-)	12	6	5	0	1	1479	1.0	1.0	0.8	0.9	0.9
copper(1+)	11	6	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
urea	11	5	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
dehydroascorbide(1-)	9	5	4	2	0	1479	1.0	0.7	1.0	0.8	0.8
choline	8	5	4	3	0	1478	1.0	0.6	1.0	0.7	0.8
cholesterol	8	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
L-ornithinium(1+)	8	4	2	0	2	1481	1.0	1.0	0.5	0.7	0.7
citrate(3-)	7	3	2	0	2	1481	1.0	1.0	0.5	0.7	0.7
bile acid	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
L-cystine zwitterion	7	4	0	1	4	1480	1.0	0.0	0.0	0.0	0.0
lipid-linked peptido-glycan	7	4	2	3	2	1478	1.0	0.4	0.5	0.4	0.4
cobalamin	6	3	0	0	3	1482	1.0	-	0.0	0.0	NaN
oleate	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
heme	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
S-adenosyl-L-methionine zwitterion	6	3	3	3	0	1479	1.0	0.5	1.0	0.7	0.7
UDP-D-glucose(2-)	6	3	1	1	2	1481	1.0	0.5	0.3	0.4	0.4
L-glutamate(1-)	6	4	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
biotinate	6	2	2	4	1	1478	1.0	0.3	0.7	0.4	0.5
urate salt	5	3	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
acetyl-CoA(4-)	5	2	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
hydrogencarbonate	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
iron(3+)	5	2	0	1	3	1481	1.0	0.0	0.0	0.0	0.0
phosphoglycerate	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
maltose	5	2	0	0	3	1482	1.0	-	0.0	0.0	NaN
2-oxoglutarate(2-)	5	2	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
sucrose	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
glutathionate(1-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
copper(2+)	5	3	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
1-4-butanediammonium	5	3	0	1	3	1481	1.0	0.0	0.0	0.0	0.0
UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
fatty acyl-CoA	4	1	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
fluoride	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
L-proline zwitterion	4	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
NAD	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
L-argininium(1+)	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-arabinose	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
prostaglandin	3	1	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
D-methionine zwitterion	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7

Continued on next page

Table 23: Hold-out test results for TooT-TripleT: offline sampling, 90 epochs, Manhattan.

Substrate	Train	Test	TP	FP	FN	TN	Acc	Prec	Recall	F1	MCC
boric acid	3	1	1	3	0	1481	1.0	0.2	1.0	0.4	0.5
L-methionine zwitterion	3	1	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
L-histidine zwitterion	3	1	0	1	2	1482	1.0	0.0	0.0	0.0	0.0
glycine zwitterion	3	1	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
bacteriocin	3	1	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
long-chain fatty acid	3	1	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid zwitterion	3	2	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
N-ribosylnicotinamide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphate ion	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
mercury cation	3	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
GDP-fucose	3	2	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
cadmium cation	3	2	0	0	1	1484	1.0	-	0.0	0.0	NaN
pyrimidine nucleotide	3	2	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphonatoenol pyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuraminase(2-)	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
xanthine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
molybdate	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
sn-glycerol 3-phosphate(2-)	2	1	0	3	1	1481	1.0	0.0	0.0	0.0	0.0
tungstate	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
gluconate	2	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
N-retinylidene phosphatidylethanolamine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
maltodextrin	2	2	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-ascorbate	2	2	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
uridine	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7

Continued on next page

Table 23: Hold-out test results for TooT-Triplet: offline sampling, 90 epochs, Manhattan.

Substrate	Train	Test	TP	FP	FN	TN	Acc	Prec	Recall	F1	MCC
Overall	2970	1485	1351	-	-	-	0.998	0.910	0.756	0.712	0.893

The table shows the hold-out test results on UniProt-SPEC-100 for TooT-Triplet with offline sampling for 90 epochs, with Manhattan distance. The columns Train and Test show the number of samples for each substrate class in the training and test sets. TP, FP, FN, and TN columns show the number of True Positives, False Positives, False Negatives, and True Negatives correspondingly for each class in the results of TooT-Triplet predictions. The overall number for TP is presented as the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric. The significant digits presented for these metrics correspond to the Test set size.

5.7.2 Online Sampling Results

Due to computational limitations for online sampling results are obtained for only one distance metric and only for 10 epochs. Manhattan distance is used for online sampling performance results. Tables 24 and 25 present the results. Overall, the online model achieves the MCC of 0.835 for the hold-out test set classifying 91 out of 96 specific substrates.

Table 24: TooT-Triplet (online) results on hold-out test set.

Distance metric	Accuracy	Precision	Recall	F1-Score	MCC
Manhattan	0.997	0.861	0.699	0.680	0.835

The table presents the results for the model trained with online sampling strategy and Manhattan distance metric. The presented distance metric is used in both the triplet loss function and K-NN algorithm training.

Table 25: Hold-out test results for TooT-Triplet: online sampling, 10 epochs, Manhattan.

Substrate	Train	Test	TP	FP	FN	TN	Acc	Prec	Recall	F1	MCC
proton	817	409	376	32	33	1044	0.956	0.922	0.919	0.920	0.890
calcium(2+)	576	289	261	37	28	1159	0.956	0.876	0.903	0.889	0.862
potassium(1+)	484	242	216	35	26	1208	0.959	0.861	0.893	0.876	0.852
chloride	295	147	117	31	30	1307	0.959	0.791	0.796	0.793	0.770
sodium(1+)	178	88	74	13	14	1384	0.98	0.85	0.84	0.85	0.84
zinc(2+)	53	27	24	2	3	1456	1.00	0.92	0.89	0.91	0.90
sulfate	44	23	10	8	13	1454	0.99	0.56	0.43	0.49	0.48
ammonium	35	17	14	2	3	1466	1.00	0.88	0.82	0.85	0.85
water	29	14	14	1	0	1470	1.00	0.93	1.00	0.97	0.97
riboflavin(1-)	24	13	9	0	4	1472	1.00	1.00	0.69	0.82	0.83
protein	18	10	8	1	2	1474	1.00	0.89	0.80	0.84	0.84

Continued on next page

Table 25: Hold-out test results for TooT-TripleT: online sampling, 10 epochs, Manhattan.

Substrate	Train	Test	TP	FP	FN	TN	Acc	Prec	Recall	F1	MCC
glutathione derivative	18	9	9	0	0	1476	1.0	1.0	1.0	1.0	1.0
retinol	18	9	7	1	2	1475	1.0	0.9	0.8	0.8	0.8
nitrate	18	8	5	1	3	1476	1.0	0.8	0.6	0.7	0.7
iron(2+)	16	8	7	1	1	1476	1.0	0.9	0.9	0.9	0.9
pyruvate	14	6	6	1	0	1478	1.0	0.9	1.0	0.9	0.9
ATP(4-)	12	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
copper(1+)	11	6	6	1	0	1478	1.0	0.9	1.0	0.9	0.9
urea	10	6	4	0	2	1479	1.0	1.0	0.7	0.8	0.8
dehydroascorbide(1-)	9	5	4	0	1	1480	1.0	1.0	0.8	0.9	0.9
choline	9	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
cholesterol	8	4	3	1	1	1480	1.0	0.8	0.8	0.8	0.7
L-ornithinium(1+)	8	4	3	1	1	1480	1.0	0.8	0.8	0.8	0.7
lipid-linked peptido-glycan	8	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
L-cystine zwitterion	8	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
L-glutamate(1-)	7	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
citrate(3-)	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
S-adenosyl-L-methionine zwitterion	6	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
glutathionate(1-)	6	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
1-4-butanediammonium	6	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
UDP-D-glucose(2-)	6	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
bile acid	6	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
heme	6	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
oleate	6	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
cobalamin	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	1	3	2	1479	1.0	0.2	0.3	0.3	0.3
iron(3+)	5	2	0	0	2	1483	1.0	-	0.0	0.0	NaN
maltose	5	2	1	3	1	1480	1.0	0.2	0.5	0.3	0.4
2-oxoglutarate(2-)	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
acetyl-CoA(4-)	5	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
biotinate	5	3	1	3	2	1479	1.0	0.2	0.3	0.3	0.3
urate salt	5	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
copper(2+)	5	3	0	0	3	1482	1.0	-	0.0	0.0	NaN
N-ribosylnicotinamide	4	1	0	3	1	1481	1.0	0.0	0.0	0.0	0.0
phosphate ion	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
mercury cation	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-proline zwitterion	4	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
NAD	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-argininium(1+)	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8

Continued on next page

Table 25: Hold-out test results for TooT-Triplet: online sampling, 10 epochs, Manhattan.

Substrate	Train	Test	TP	FP	FN	TN	Acc	Prec	Recall	F1	MCC
hydrogencarbonate	4	3	2	1	1	1481	1.0	0.7	0.7	0.7	0.7
phosphoglycerate	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
sucrose	4	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
bacteriocin	3	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
maltodextrin	3	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
prostaglandin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-methionine zwitterion	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
L-ascorbate	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-histidine zwitterion	3	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
pyrimidine nucleotide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid zwitterion	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
GDP-fucose	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
cadmium cation	3	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
long-chain fatty acid	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuraminate(2-)	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
arsenate ion	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
xanthine	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
flavin adenine dinucleotide	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
sn-glycerol 3-phosphate(2-)	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
tungstate	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
gluconate	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
N-retinylidene phosphatidylethanolamine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
molybdate	2	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0

Continued on next page

Table 25: Hold-out test results for TooT-Triplet: online sampling, 10 epochs, Manhattan.

Substrate	Train	Test	TP	FP	FN	TN	Acc	Prec	Recall	F1	MCC
D-methionine zwitterion	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
boric acid	2	2	0	0	2	1483	1.0	-	0.0	0.0	NaN
L-arabinose	2	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
uridine	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
Overall	2970	1485	1278	-	-	-	0.997	0.861	0.699	0.680	0.835

The table shows the hold-out test results on UniProt-SPEC-100 for TooT-Triplet with online sampling for 10 epochs, with Manhattan distance. The columns Train and Test show the number of samples for each substrate class in the training and test sets. TP, FP, FN, and TN columns show the number of True Positives, False Positives, False Negatives, and True Negatives correspondingly for each class in the results of TooT-Triplet predictions. The overall number for TP presents the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric. The significant digits presented for these metrics correspond to the Test set size.

5.8 Discussion

This section discusses four points in the TooT-Triplet model results. First, the effect of applying triplet sampling is compared to random sampling in Section 5.8.1. This section also compares the results for models trained with two strategies of online and offline triplet sampling. Section 5.8.2 compares the encoding space of models trained with three distance metrics (Manhattan, Euclidean, and Cosine similarity). This section also investigates the relationship between the reliability of a prediction and the distance of the test sample from the nearest neighbor. The effect of class imbalance on results is explored by an ablation study for the majority class in Section 5.8.3. Finally, the model’s ability to capture information on the binding sites is discussed by an example in Section 5.8.4.

5.8.1 Sampling strategy

TooT-Triplet utilizes the triplet sampling strategy to solve the shortcomings of the TooT-SPEC model. As presented in Section 3.7, this model is trained using a random sampling strategy and predicting the substrate specificity of 83 classes out of 96 using a logistic regression model as the classifier. As discussed in Section 3.7.6, 13 substrate classes remain unclassified with zero True or False positive predictions. Since these classes are the minority classes, the triplet sampling strategy is implemented to increase the consistency in small classes, decreasing their overlap with majority classes.

Table 26 compares the results on the hold-out test set for four model groups: Baseline models, TooT-SPEC models, TooT-Triplet models with offline sampling strategy, and TooT-Triplet model with an online sampling strategy. TooT-SPEC models include the results for the original TooT-SPEC model as described in Section 3.7 which uses a logistic regression classifier. To investigate the effect of Triplet sampling versus random sampling, three additional TooT-SPEC models are presented. These models retain the ProtBERT-BFD language model from the TooT-SPEC fine-tuned with random sampling and replace the logistic regression classifier with a K-NN classifier ($K=1$), implemented with three different distance metrics.

Each of the models in Baseline, TooT-SPEC, and TooT-Triplet offline models is implemented using three different distance metrics: Manhattan, Euclidean, and Cosine. These distance metrics are utilized in both loss function and K-NN implementation for the TooT-Triplet models, while they are only incorporated into the K-NN implementation for the Baseline and TooT-SPEC models. The sampling strategy and loss function are used in the ProtBERT-BFD model parameters updates during the training process (with fine-tuning) but do not apply to the Baseline models with frozen parameters (Section 5.3.4).

According to the table 26, the results on the UniProt-SPEC-100 hold-out test set for all the Baselines with different K-NN models are closely aligned. A pairwise comparison of Baselines and TooT-Triplet models, using the same distance metrics, indicates numerically higher performance for the TooT-Triplet models. This observation suggests that the performance improvement of TooT-Triplet is due to enhancements in sequence embeddings rather than the distance metrics used in K-NN training. TooT-Triplet models also cover more unpredicted classes in the test set, increasing the number of effective classes from 85 to 91. Table 27 shows the TooT-Triplet model covers 10 of 11 unpredicted classes from Baseline 1; however, four classes predicted by Baseline 1 are reported with NaN values in TooT-Triplet.

As the results in Table 26 presents, the online and offline triplet sampling strategies, integrated with a triplet loss function, increase the number of effective classes from 83 (in TooT-SPEC) to 91 for 8 classes. The lower number of unpredicted classes in TooT-Triplet indicates the effectiveness of the triplet sampling and loss function in training minority classes. While the method does not completely solve the problem of unpredicted classes, it still outperforms the TooT-SPEC model, which uses random sampling and a cross-entropy loss function, in terms of covering more effective classes.

According to the Table 26, the performance metrics report higher values for hold-out test set prediction using TooT-SPEC models, specifically TooT-SPEC followed by a Euclidean distance-based K-NN model. However, the overall 3-fold prediction performance of the models is not statistically significant. The pair-wise student test between TooT-SPEC (Euclidean) and TooT-Triplet (offline, Manhattan) reports a p-value of 0.2338 for MCC which does not reject the null hypothesis [Stu08].

Although the definition of the online sampling strategy suggests better performance by sampling *hard* triplets each epoch, the results indicate that offline sampling actually outperforms the online approach (Table 26). The lower performance of TooT-Triplet (online) could be attributed to training constraints (only 10 epochs) imposed by high computational costs. While memory usage for the online model is managed via the LoRa algorithm, one epoch of online training using 4 GPUs takes 26,444.3 seconds, whereas offline sampling on a single GPU takes only 371.42 seconds per epoch.

Table 26: Performance comparison of four model groups for one-shot classification on the hold-out test set of UniProt-SPEC-100 dataset.

Model	Sampling strategy	Loss function	Distance metric	Effective classes F1-score	Effective classes MCC	# Effective /Total classes	Mutual classes F1-score	Mutual classes MCC
Baseline 1 (Frozen Manhattan)	Full dataset	No loss	Manhattan	<u>0.672</u>	0.824	85/96	0.838	0.859
Baseline 2 (Frozen Euclidean)	Full dataset	No loss	Euclidean	<u>0.672</u>	0.824	85/96	0.839	0.861
Baseline 3 (Frozen Cosine)	Full dataset	No loss	Cosine	0.682	<u>0.831</u>	86/96	<u>0.844</u>	<u>0.865</u>
TooT-SPEC (LR)	Random sampling	Cross Entropy	No dist.	0.741	0.908	83/96	0.927	0.936
TooT-SPEC (KNN)	Random sampling	Cross Entropy	Manhattan	0.758	0.914	85/96	0.927	0.936
TooT-SPEC (KNN)	Random sampling	Cross Entropy	Euclidean	<u>0.759</u>	<u>0.915</u>	86/96	0.928	<u>0.937</u>
TooT-SPEC (KNN)	Random sampling	Cross Entropy	Cosine	0.756	0.913	86/96	<u>0.929</u>	<u>0.937</u>
TooT-Triplet (offline)	Offline triplet sampling	Triplet loss	Manhattan	0.712	<u>0.893</u>	91 /96	<u>0.914</u>	<u>0.923</u>
TooT-Triplet (offline)	Offline triplet sampling	Triplet loss	Euclidean	<u>0.715</u>	0.889	91 /96	0.908	0.920
TooT-Triplet (offline)	Offline triplet sampling	Triplet loss	Cosine	0.683	0.875	91 /96	0.897	0.910
TooT-Triplet (online)	Online triplet sampling	Triplet loss	Manhattan	<u>0.680</u>	<u>0.835</u>	91 /96	<u>0.851</u>	<u>0.872</u>

The table shows a performance comparison of four model groups for one-shot classification on the hold-out test set of the UniProt-SPEC-100 dataset. The column “# Effective/# Total classes” shows the ratio of the number of retrieved classes in the test results to the total number of the classes in the test set. For the Baseline and TooT-SPEC models, the distance metrics are only incorporated in K-NN classifier while in TooT-Triplet models these distance metrics are used in the Triplet loss function for updating ProtBERT-BFD model parameters as well as the K-NN model. The F1-score and MCC are reported for the number of effective classes on the hold-out test set. The last two columns show the overall results for 71 mutual classes which are covered in all of the model’s predictions. The bold values indicate the numerically largest entry in each column and the underline presents the numerically highest values in each model group.

Table 27: MCC comparison for unclassified classes in TooT-Triplet and Baseline 1.

Substrate	Baseline 1	TooT-Triplet
iron(3+)	NaN	0.0
cobalamin	0.816	NaN
maltose	0.407	NaN
maltodextrin	NaN	1.0
cadmium cation	0.707	NaN
L-tryptophan zwitterion	NaN	0.0
glycine betaine	NaN	0.0
arsenate ion	NaN	1.0
guanine	NaN	1.0
xanthine	NaN	1.0
folate(2-)	0.0	NaN
sn-glycerol 3-phosphate(2-)	NaN	0.0
tungstate	NaN	NaN
gluconate	NaN	0.6
boric acid	NaN	0.5

The table presents the results for unclassified classes in UniProt-SPEC-100 hold-out test set using the Baseline 1 and TooT-Triplet offline method with Manhattan distance.

5.8.2 Interpreting sample distance

The efficacy of the K-Nearest Neighbors (KNN) algorithm in high-dimensional spaces has been a topic of extensive discussion [Pes13, BDR05, AHK01]. The effectiveness of this algorithm in higher dimensions is highly dependent on the feature qualities, the distance metric, and the dataset in hand. Also, feature qualities captured by BERT models have been presented reliably in previous research. Section 5.8.4 discusses the specific features extracted by Prot-BERT-BFD in our training process.

To determine the most effective distance metric for the training process, it is essential to assess the class representations generated by models trained with various distance metrics. One such evaluation metric is the silhouette score, which measures the consistency of clusters within an encoding space, assigning a number ranging between -1, and 1. For this analysis, samples within each class are treated as clusters in the encoding space. The silhouette score is then computed for each of the 96 classes in the dataset (Figure 26). Figure 27 presents the comparison of silhouette scores of the UniProt-SPEC-100 in the

encoding space trained using each distance metric.

The silhouette score indicates that the L1-norm, or Manhattan distance, is the most effective distance metric for the triplet loss function to build consistent representations of classes. This finding is corroborated by the classification results of the K-Nearest Neighbor classifier, which utilized the same distance metric, as shown in Table 26. The adoption of Manhattan distance as the preferred metric for the TooT-Triplet model, which captures similarities across 1024 dimensions, aligns with prior research aimed at identifying the optimal metric for capturing similarities in high-dimensional spaces [AHK01].

To conduct a more thorough investigation into sample distances in the TooT-Triplet model trained with the Manhattan distance using an offline sampling strategy, the distribution of distances for the nearest positive and nearest negative samples is presented in Figure 28. As shown in the figure, the nearest positive samples exhibit a long-tailed distribution, with a higher frequency of positive samples at distances less than 1. In contrast, the nearest negative neighbors follow a normal distribution, with a majority of the distances ranging between 1.5 and 2.5.

We investigate the frequency of nearest negative and positive distance distributions within each class in Table 28, and Figures 31, 32, and 33 show a box plot of these distances for samples in the head, middle, and tail classes of the imbalanced dataset of UniProt-SPEC-100. Figure 31 suggests a more distinct separation of the nearest positive and negative distances for the five classes with the highest number of samples, with the nearest positive samples holding lower distance values. A similar pattern of separation is noticed in the middle and tail classes (Figures 32 and 33). However, a few classes—such as **L-glutamate(-1)** and **L-cystine zwitterion**—have overlapping distance distributions for negative and positive samples. While not all classes fully adhere to this criterion, the overall model can be considered effectively trained on minimizing the anchor-positive distance and maximizing the anchor-negative distance.

Table 28 presents the average distances of the nearest positive and negative neighbors for the samples in the training set. Generally, a lower average distance is observed for positive samples within a class compared to the negative average distance. However, overlooking exceptions, the prediction distance could serve as a measure of reliability, indicating that a lower calculated distance for a test sample prediction suggests a more reliable prediction. Based on the average nearest negative distance column, distance values exceeding 2 are advised to be interpreted judiciously.

Table 28: Average distance of nearest neighbor after training.

Class	Avg. distance of nearest neighbor with same label	Std. distance of nearest neighbor with same label	Avg. distance of nearest neighbor with different labels	Std. distance of nearest neighbor with different labels	size
proton	0.958	0.595	1.958	0.336	817
calcium(2+)	0.659	0.598	1.666	0.299	577
potassium(1+)	0.727	0.610	1.759	0.317	484
chloride	0.813	0.642	1.756	0.232	295
sodium(1+)	0.712	0.594	1.529	0.254	177
zinc(2+)	1.01	0.62	1.82	0.22	54
sulfate	1.21	0.68	1.84	0.23	45
ammonium	0.67	0.46	1.56	0.25	34
water	0.50	0.45	1.34	0.20	28
riboflavin(1-)	1.12	0.39	1.73	0.25	25
protein	0.89	0.37	1.44	0.44	19
retinol	0.81	0.56	1.89	0.27	18
glutathione derivative	0.58	0.58	1.09	0.23	18
nitrate	1.27	0.63	1.49	0.43	17
iron(2+)	1.17	0.95	1.62	0.33	16
pyruvate	0.72	0.64	1.90	0.25	13
copper(1+)	1.03	0.47	1.39	0.22	12
ATP(4-)	0.37	0.32	1.38	0.13	12
urea	0.68	0.44	1.98	0.38	11
dehydroascorbide(1-)	0.41	0.28	1.13	0.28	10
choline	0.5	0.2	2.9	0.1	9
cholesterol	0.9	0.8	1.7	0.3	8
L-ornithinium(1+)	1.0	0.4	1.4	0.5	8
bile acid	0.2	0.2	1.5	0.1	7
L-glutamate(1-)	0.9	0.6	0.9	0.5	7
L-cystine zwitterion	1.1	0.8	1.8	0.2	7
lipid-linked peptido- glycan	1.6	0.4	1.9	0.1	7
cobalamin	1.3	0.7	1.5	0.3	6
UDP-D-glucose(2-)	1.5	1.4	2.3	0.2	6
citrate(3-)	0.6	0.7	2.0	0.1	6
S-adenosyl-L- methionine zwitterion	1.0	0.3	1.2	0.3	6
3'-phosphonato-5'- adenylyl sulfate(4-)	0.4	0.2	1.4	0.2	6
UDP-N-acetyl-alpha- D-glucosamine(2-)	0.8	0.3	0.9	0.1	6
heme	0.8	0.3	1.1	0.1	6
copper(2+)	1.5	0.2	1.3	0.2	6
urate salt	0.7	0.5	1.5	0.5	6
oleate	0.7	0.9	2.2	0.5	6
phosphoglycerate	0.3	0.2	1.1	0.1	5
glutathionate(1-)	0.6	1.0	2.8	0.4	5
biotinate	0.8	0.4	1.6	0.3	5
hydrogencarbonate	1.0	0.7	1.1	0.5	5

Continued on next page

Table 28: Average distance of nearest neighbor after training.

Class	Avg. distance of nearest neighbor with same label	Std. distance of nearest neighbor with same label	Avg. distance of nearest neighbor with different labels	Std. distance of nearest neighbor with different labels	size
1-4-	1.3	0.8	1.6	0.4	5
butanediammonium					
sucrose	0.5	0.1	3.8	0.1	5
acetylcholine	1.2	0.5	2.0	0.1	4
maltose	1.4	0.0	2.0	0.0	4
cadmium cation	1.2	0.6	1.4	0.1	4
L-argininium(1+)	1.2	0.4	1.3	0.2	4
L-proline zwitterion	1.0	0.7	1.1	0.3	4
NAD	0.5	0.3	1.0	0.1	4
formate	1.0	0.3	2.0	0.1	4
iron-sulfur cluster	0.6	0.1	1.8	0.1	4
iron(3+)	1.8	0.5	1.5	0.2	4
GDP-fucose	0.3	0.3	1.0	0.1	4
acetyl-CoA(4-)	1.2	0.6	2.9	0.1	4
fluoride	0.9	0.3	2.1	0.0	4
gamma-aminobutyric acid zwitterion	0.9	0.7	1.7	0.1	4
2-oxoglutarate(2-)	1.3	0.7	1.8	0.2	4
pyrimidine nucleotide	0.4	0.3	1.1	0.0	4
L-arabinose	2.0	0.1	1.6	0.1	3
N-ribosylnicotinamide	1.7	0.0	1.8	0.1	3
D-methionine zwitterion	1.4	0.8	2.9	0.1	3
maltodextrin	1.1	0.7	1.9	0.1	3
mercury cation	1.3	0.8	2.3	0.1	3
boric acid	0.9	0.4	1.1	0.3	3
phosphate ion	1.5	0.0	2.0	0.1	3
L-ascorbate	1.6	0.1	1.8	0.2	3
fatty acyl-CoA	0.9	0.4	1.8	0.3	3
phosphoenolpyruvate	0.4	0.0	0.9	0.0	2
glycine betaine	2.0	0.0	1.9	0.2	2
arsenate ion	1.6	0.0	1.7	0.1	2
guanine	1.9	0.0	1.5	0.2	2
xanthine	1.6	0.0	2.4	0.1	2
tungstate	1.7	0.0	1.4	0.1	2
bacteriocin	2.0	0.0	1.3	0.3	2
glycine zwitterion	0.4	0.0	2.2	0.0	2
L-tryptophan zwitterion	1.5	0.0	1.3	0.4	2
beta-D-glucan	0.7	0.0	1.9	0.0	2
c-GMP-AMP(2-)	0.2	0.0	1.5	0.0	2
long-chain fatty acid	1.5	0.0	2.1	0.2	2
sn-glycerol 3-phosphate(2-)	2.4	0.0	2.0	0.3	2

Continued on next page

Table 28: Average distance of nearest neighbor after training.

Class	Avg. distance of nearest neighbor with same label	Std. distance of nearest neighbor with same label	Avg. distance of nearest neighbor with different labels	Std. distance of nearest neighbor with different labels	size
uridine	1.7	0.0	1.4	0.3	2
prostaglandin	0.1	0.0	4.1	0.0	2
folate(2-)	2.2	0.0	1.7	0.3	2
L-histidine zwitterion	1.6	0.0	0.8	0.5	2
CMP-N-acetyl-beta-neuraminate(2-)	0.2	0.0	0.8	0.0	2
(R)-carnitine	0.2	0.0	3.4	0.0	2
L-methionine zwitterion	1.2	0.0	1.6	0.0	2
teichoic acid	0.9	0.0	1.8	0.0	2
flavin adenine dinucleotide	1.5	0.0	1.8	0.2	2
N-(4-aminobenzoyl)-L-glutamate	1.3	0.0	2.8	0.1	2
molybdate	1.8	0.0	1.6	0.2	2
N-retinylidene phosphatidylethanolamine	0.6	0.0	3.0	0.0	2
UDP-alpha-D-xylose(2-)	0.4	0.0	1.8	0.1	2
gluconate	1.8	0.0	2.6	0.1	2
lipopolysaccharide	0.9	0.0	1.9	0.0	2
magnesium cation	0.8	0.0	1.6	0.1	2

The table presents the average and standard deviation of the distances for the nearest positive (same label) and negative neighbors (different labels) in the UniProt-SPEC-100 dataset. The classes are sorted by their size in the training set as presented in the last column.

5.8.3 Majority class effect on training

The class `proton` represents the `hydrogen ion` (H^+), which is essential in many cellular functions, particularly in membrane transport tasks. The `hydrogen ion` is involved in the cotransport of various substrates into and out of the cell. This broad involvement accounts for the wide variety of sequences associated with this class in the protein space. To measure the impact of this class as the majority class, an experiment is conducted on `protonclass` ablation.

Two TooT-Triplet models, with the described architecture in Section 5.3 are trained separately with the same network parameters, triplet loss function (euclidean distance) with offline sampling strategy along AdamW as optimizer with learning rate of $1e-5$ for 97 epochs. The difference between the models is in the train and test datasets. The first model, referred to as Proton TooT-Triplet in this section, is trained and tested on

the UniProt-SPEC-100 dataset and the second model, referred to as Proton-Ex TooT-Triplet, is trained and tested on UniProt-SPEC-100-ProtonExcluded, which contains all the samples from the first dataset except for samples from the class **proton**. Both datasets are described in detail in Section 5.4.

Table 29: Comparison of TooT-Triplet models in presence/absence of majority class.

Substrate	Accuracy		Precision		Recall		F1-Score		MCC	
	H+	ex.H+	H+	ex.H+	H+	ex.H+	H+	ex.H+	H+	ex.H+
calcium(2+)	0.970	0.948	0.921	0.879	0.927	0.934	0.924	0.906	0.905	0.871
potassium(1+)	0.975	0.954	0.932	0.925	0.913	0.864	0.923	0.893	0.908	0.864
chloride	0.980	0.958	0.934	0.849	0.865	0.844	0.898	0.846	0.888	0.822
sodium(1+)	0.99	0.97	0.90	0.84	0.94	0.84	0.92	0.84	0.92	0.83
zinc(2+)	1.00	0.99	0.96	0.83	0.85	0.93	0.90	0.88	0.90	0.88
sulfate	0.99	0.98	0.79	0.69	0.68	0.48	0.73	0.56	0.73	0.57
ammonium	1.00	1.00	0.94	0.89	0.94	0.94	0.94	0.91	0.94	0.91
water	1.00	1.00	0.93	1.00	1.00	1.00	0.97	1.00	0.97	1.00
riboflavin(1-)	0.99	0.99	0.83	0.56	0.42	0.69	0.56	0.62	0.59	0.62
protein	1.0	1.0	1.0	0.9	0.9	1.0	0.9	1.0	0.9	1.0
retinol	1.0	1.0	0.9	0.9	1.0	0.9	0.9	0.9	0.9	0.9
glutathione derivative	1.0	1.0	0.7	1.0	0.9	1.0	0.8	1.0	0.8	1.0
nitrate	1.0	1.0	0.6	0.8	0.6	0.6	0.6	0.7	0.6	0.7
iron(2+)	1.0	1.0	0.7	0.7	1.0	0.9	0.8	0.8	0.8	0.8
pyruvate	1.0	1.0	1.0	1.0	0.9	1.0	0.9	1.0	0.9	1.0
ATP(4-)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
copper(1+)	1.0	1.0	0.5	0.7	0.7	1.0	0.6	0.8	0.6	0.8
urea	1.0	1.0	0.8	1.0	0.6	0.8	0.7	0.9	0.7	0.9
dehydroascorbide(1-)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
choline	1.0	1.0	0.8	1.0	0.8	1.0	0.8	1.0	0.8	1.0
cholesterol	1.0	1.0	1.0	0.8	1.0	0.8	1.0	0.8	1.0	0.7
L-ornithinium(1+)	1.0	1.0	1.0	0.5	0.5	0.8	0.7	0.6	0.7	0.6
citrate(3-)	1.0	1.0	0.7	0.8	0.7	1.0	0.7	0.9	0.7	0.9
bile acid	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
L-cystine zwitterion	1.0	1.0	0.6	0.5	0.8	0.7	0.7	0.6	0.7	0.6
lipid-linked peptidoglycan	1.0	1.0	0.7	0.5	1.0	0.7	0.8	0.6	0.8	0.6
cobalamin	1.0	1.0	0.5	1.0	1.0	0.7	0.7	0.8	0.7	0.8
oleate	1.0	1.0	1.0	0.6	1.0	1.0	1.0	0.8	1.0	0.8
heme	1.0	1.0	1.0	1.0	0.7	1.0	0.8	1.0	0.8	1.0
S-adenosyl-L-methionine zwitterion	1.0	1.0	1.0	1.0	0.3	0.3	0.5	0.5	0.6	0.6
UDP-D-glucose(2-)	1.0	1.0	0.3	0.8	0.7	1.0	0.4	0.9	0.5	0.9
L-glutamate(1-)	1.0	1.0	0.6	0.5	0.8	0.3	0.7	0.4	0.7	0.4
biotinate	1.0	1.0	0.5	0.5	0.5	0.3	0.5	0.4	0.5	0.4
urate salt	1.0	1.0	1.0	1.0	1.0	0.7	1.0	0.8	1.0	0.8
acetyl-CoA(4-)	1.0	1.0	1.0	0.5	1.0	1.0	1.0	0.7	1.0	0.7

Continued on next page

Table 29: Comparison of TooT-Triplet models in presence/absence of majority class.

Substrate	Accuracy		Precision		Recall		F1-Score		MCC	
	H+	ex.H+	H+	ex.H+	H+	ex.H+	H+	ex.H+	H+	ex.H+
hydrogencarbonate	1.0	1.0	0.3	1.0	0.5	1.0	0.4	1.0	0.4	1.0
iron(3+)	1.0	1.0	1.0	-	1.0	0.0	1.0	0.0	1.0	NaN
phosphoglycerate	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
maltose	1.0	1.0	0.7	0.5	1.0	1.0	0.8	0.7	0.8	0.7
2-oxoglutarate(2-)	1.0	1.0	1.0	0.7	0.5	1.0	0.7	0.8	0.7	0.8
3'-phosphonato-5'-adenylyl sulfate(4-)	1.0	1.0	1.0	1.0	0.3	1.0	0.5	1.0	0.6	1.0
sucrose	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
glutathionate(1-)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
copper(2+)	1.0	1.0	0.2	1.0	0.3	0.3	0.3	0.5	0.3	0.6
1-4-	1.0	1.0	0.3	0.0	0.3	0.0	0.3	0.0	0.3	0.0
butanediammonium										
UDP-N-acetyl-alpha-D-glucosamine(2-)	1.0	1.0	1.0	1.0	0.3	0.3	0.5	0.5	0.6	0.6
formate	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
acetylcholine	1.0	1.0	1.0	1.0	0.5	0.5	0.7	0.7	0.7	0.7
fatty acyl-CoA	1.0	1.0	-	1.0	0.0	1.0	0.0	1.0	-	1.0
fluoride	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
L-proline zwitterion	1.0	1.0	0.5	0.5	1.0	0.5	0.7	0.5	0.7	0.5
NAD	1.0	1.0	0.4	1.0	1.0	1.0	0.6	1.0	0.6	1.0
L-argininium(1+)	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0
iron-sulfur cluster	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
L-arabinose	1.0	1.0	-	1.0	0.0	0.5	0.0	0.7	-	0.7
prostaglandin	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
D-methionine zwitterion	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
boric acid	1.0	1.0	0.5	-	1.0	0.0	0.7	0.0	0.7	NaN
L-methionine zwitterion	1.0	1.0	0.5	1.0	1.0	1.0	0.7	1.0	0.7	1.0
L-histidine zwitterion	1.0	1.0	-	0.0	0.0	0.0	0.0	0.0	-	0.0
glycine zwitterion	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
bacteriocin	1.0	1.0	-	0.3	0.0	1.0	0.0	0.5	-	0.6
long-chain fatty acid	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid zwitterion	1.0	1.0	-	0.5	0.0	0.5	0.0	0.5	-	0.5
N-ribosylnicotinamide	1.0	1.0	1.0	-	1.0	0.0	1.0	0.0	1.0	NaN
phosphate ion	1.0	1.0	1.0	0.5	1.0	1.0	1.0	0.7	1.0	0.7
mercury cation	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
GDP-fucose	1.0	1.0	1.0	1.0	0.5	1.0	0.7	1.0	0.7	1.0
cadmium cation	1.0	1.0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
pyrimidine nucleotide	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
phosphonateoenolpyruvate	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Continued on next page

Table 29: Comparison of TooT-Triplet models in presence/absence of majority class.

Substrate	Accuracy		Precision		Recall		F1-Score		MCC	
	H+	ex.H+	H+	ex.H+	H+	ex.H+	H+	ex.H+	H+	ex.H+
CMP-N-acetyl-beta-neuraminate(2-)	1.0	1.0	1.0	0.5	1.0	1.0	1.0	0.7	1.0	0.7
c-GMP-AMP(2-)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	1.0	1.0	-	0.0	0.0	0.0	0.0	0.0	-	0.0
N-(4-aminobenzoyl)-L-glutamate	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
glycine betaine	1.0	1.0	-	0.0	0.0	0.0	0.0	0.0	-	0.0
arsenate ion	1.0	1.0	1.0	-	1.0	0.0	1.0	0.0	1.0	NaN
teichoic acid	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
guanine	1.0	1.0	1.0	-	1.0	0.0	1.0	0.0	1.0	NaN
xanthine	1.0	1.0	1.0	-	1.0	0.0	1.0	0.0	1.0	NaN
flavin adenine dinucleotide	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
molybdate	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
magnesium cation	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	1.0	1.0	0.5	1.0	1.0	1.0	0.7	1.0	0.7	1.0
sn-glycerol 3-phosphate(2-)	1.0	1.0	0.0	-	0.0	0.0	0.0	0.0	0.0	NaN
tungstate	1.0	1.0	0.1	-	1.0	0.0	0.2	0.0	0.4	NaN
gluconate	1.0	1.0	1.0	-	1.0	0.0	1.0	0.0	1.0	NaN
beta-D-glucan	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	1.0	1.0	1.0	0.3	1.0	1.0	1.0	0.5	1.0	0.6
N-retinylidene phosphatidylethanolamine	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
folate(2-)	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
maltodextrin	1.0	1.0	0.7	-	1.0	0.0	0.8	0.0	0.8	NaN
L-ascorbate	1.0	1.0	-	0.5	0.0	1.0	0.0	0.7	-	0.7
uridine	1.0	1.0	0.5	-	1.0	0.0	0.7	0.0	0.7	NaN
Overall	0.998	0.997	0.897	0.856	0.754	0.717	0.722	0.688	0.878	0.831

The table presents a comparison of TooT-Triplet (offline) performance in the presence and absence of the majority class (**proton (H+)**) reported in five metrics of Accuracy, Precision, Recall, F1-Score, and MCC. The overall entry shows the macro average over all the classes for each metric. The columns with Proton in their name show the metric values for the TooT-Triplet model trained and tested in the presence of the majority class (**proton (H+)**). The columns with Proton-Ex in their name show the metric values for the TooT-Triplet model trained and tested without the majority class (**proton (H+)**). The training dataset for the TooT-Triplet model in the presence of the majority class is UniProt-SPEC-100 and the the dataset for the TooT-Triplet model trained in the absence of the majority class is UniProt-SPEC-100- ProtonExcluded both described in Section 5.4. The significant digits presented for these metrics correspond to the test set size.

Table 29 presents intriguing results comparing the two models. In the absence of the majority class **proton**, the top seven classes in terms of size have shown lower performance in all presented metrics, while the middle classes numbered 8 to 21 show improvement in results. The comparison of the results for the tail classes is inconclusive, as there are

both improvements and deteriorations in minority class results. The overall evaluation of the dataset indicates a decrease in performance in the absence of the majority class.

However, the evaluation of these two models with the external dataset suggests that better representations for the samples are achieved in the absence of the class **proton**. Table 32 presents the predictions of the Proton-Ex model for external dataset samples where the Proton TooT-Triplet model makes incorrect predictions. Out of the 11 cases of incorrect predictions by Proton TooT-Triplet, the Proton-Ex model correctly predicted 9. Additionally, the distances of the predictions in the Proton-Ex model are notably different from the Proton TooT-Triplet. In the results obtained from the Proton TooT-Triplet model, the distances from a sample to its top three neighbors exhibit close proximity. Conversely, in the Proton-Ex model, the first nearest neighbor, which is also predicted as the model’s output, is notably closer to the test sample compared to the next two nearest neighbors.

For instance, consider the test sequence Q8L7Z0, which was initially predicted as a **proton** transport protein by the Proton TooT-Triplet model but was correctly identified as an **L-glutamate(1-)** transporter by the Proton-Excluded TooT-Triplet model. The distance for this corrected prediction is 0.937, significantly closer than the next nearest neighbors, **L-ascorbate** and **2-oxoglutarate(2-)** which are at distances of 23.654 and 23.868, respectively. However, despite these observations suggesting improved sample representations in the absence of the majority class **proton**, the decline in the test set results precludes a definitive conclusion on this matter.

5.8.4 Binding sites and attention heads

The binding site of a transporter refers to a specific region on the protein surface where a ligand binds. The binding site is formed by a specific arrangement of amino acid residues that create a pocket or cavity that is complementary in shape and chemical properties to the ligand being transported. It plays a crucial role in the substrate specificity of the transport protein, allowing the protein to selectively bind and transport specific molecules across cell membranes [JPN18]. Given its critical role in substrate specificity, the presence of binding site information in the extracted sample encodings is a topic of discussion.

Building upon prior research outlined in the paper titled as “*BERTology meets Biology*”, it is suggested that the BERT language model can capture the folding structure of proteins. This capability allows BERT to connect amino acids that are distant in the sequence but close in three-dimensional structure, particularly targeting binding sites

[VMV⁺21]. To explore the presence of binding site information in TooT-Triplet encodings, a comparison is made between the attention head values and the binding sites of protein samples. The Prot-BERT-BFD model, as detailed in Section 2.3, consists of 16 layers with 30 attention heads each, designed to capture information from a protein sequence.

As an example, a short transport protein sequence of P06703 (S10A6_HUMAN) from UniProt-KB is presented. Binding to the calcium ion Ca(2+) this protein is described to “indirectly play a role in many physiological processes such as the reorganization of the actin cytoskeleton and in cell motility” by UniProt-KB. As Figure 34 presents this protein has 7 amino acids as the binding sites in positions, 28, 33, 61, 63, 65, 67, and 72 in its structure. This 3D structure of this sequence from Alpha Fold2 ¹ focusing on binding site position 65 is presented in Figure 35 [JEP⁺21]. It should be mentioned that this image shows the predicted structure of the protein sequence on its own, independent of the bi-layer lipid membrane. This folding can be different when the protein is placed inside the membrane. However further research on Alpha Fold2’s performance on transmembrane proteins suggests the versatility and robustness of the protein 3D structure prediction as placed inside the membrane, especially in ion channels, and transporters [JTD23].

Figure 36 presents the average values of each attention head over 16 layers for the introduced sequence from TooT-Triplet. The figure indicates that attention head numbers 10 to 16 exhibit higher values at two specific positions. Figure 37 details the average values of attention head number 13, showing elevated values at positions 26 and 65. Notably, position, 65 corresponds to a binding site illustrated in Figure 34. Therefore, the assumption that attention captures this specific binding site is plausible.

In contrast, the first highly valued position, number 26, is not a binding site. To investigate the high values at this position, we examined the sequence analysis provided by UniProt-KB. As shown in Figure 38 this position does not indicate any known specific roles in the sequence. Consequently, the high value at this position remains unexplained based on current knowledge, necessitating further investigation into the sequence. Additionally, the other six binding site positions in this example do not exhibit high values in the attention heads. This observation suggests that, for transmembrane transport proteins, the ProtBERT model in TooT-Triplet captures some binding site information but does not comprehensively capture all binding sites.

¹<https://alphafold.ebi.ac.uk/entry/P06703>

5.9 External validation

For external validation of the TooT-Triplet model, ideally, a dataset containing different samples but the same classes is needed. However, substrate specificity prediction for transport proteins into specific substrates is not a well-studied topic, and there are no other reliable datasets with similarly fine-grained classes. The baseline presented in Chapter 2 is the SOTA study of TooT-SC which predicts the transport protein sequences into 11 substrate classes of `inorganic cations`, `inorganic anions`, `water`, `organic anions`, `organo-oxygens`, `amino acids and derivatives`, `other organonitrogens`, `nucleotides`, `organic heterocyclics`, `miscellaneous`, and `nonselective` [AB22a] (See Section 2.10). Applying the classification of the SVM model with an RBF kernel using a compositional, evolutionary, and positional information approach, this research reports 0.82 MCC over 11 substrate classes.

The TooT-SC dataset contains 1,524 protein sequences across 11 substrate classes. Of these, 470 sequences overlap with the UniProt-SPEC-100 dataset (Figure 39). However, there is no direct one-to-one correspondence between classes in the two datasets. In the TooT-SC dataset, each class encompasses a group of substrates, while in the UniProt-SPEC-100 dataset, each specific substrate is considered a distinct class (Figure 49). The only exception is the class of `water` transport proteins, which is consistently categorized in both datasets, albeit with different samples. As an example of TooT-SC and UniProt-SPEC-100 class mapping, the class `Inorganic cation` in the SC dataset corresponds to 11 classes of `proton`, `calcium(2+)`, `potassium(1+)`, `sodium(1+)`, `zinc(2+)`, `ammonium`, `iron(2+)`, `copper(1+)`, `copper(2+)`, `iron(3+)`, and `magnesium cation`. This indicates a one-to-many relationship between the classes of TooT-SC to UniProt-SPEC-100.

In this work, we use the TooT-SC dataset to externally evaluate the predictions of the TooT-Triplet model. However, the one-to-many mapping between the SC and SPEC datasets complicates the evaluation process. To assess the model’s predictions on the TooT-SC test set, we must verify whether each SPEC prediction can be considered part of the corresponding SC label group. We developed a systematic evaluation method using two techniques, which are detailed in the following sections.

5.9.1 Evaluation using ChEBI terms

The first method is uses the ChEBI ontology tree described in Section 3.2.1. This approach compares the ChEBI terms of SC labels to the TooT-Triplet predictions. It searches for

the TooT-Triplet prediction ChEBI label in the subtree of the SC ChEBI label. If the TooT-Triplet prediction is in the subtree, then the prediction is considered in line with SC labels at a fine-grained level.

The results of this evaluation are presented in the last column of Table E.4.1. The value "True" indicates a match between the SC label and the TooT-Triplet prediction. However, this method has shortcomings. For example, the TooT-Triplet prediction `dehydroascorbide(1-)` contains carbon-oxygen bonds, which are the main criteria for the `organo-oxygens` label in SC classes. Yet, `dehydroascorbide(1-)` is not part of the ChEBI subtree for the ChEBI terms of `organo-oxygens` in the SC dataset (CHEBI:36963).

Another example is the TooT-Triplet predictions of `L-histidine zwitterion` and `L-methionine zwitterion` where SC label is indicated as `amino acids and derivatives` but these terms are not related to the ChEBI terms presented by the TooT-SC research. This issue may arise due to the TooT-SC dataset not encompassing all possible ChEBI terms related to the class `amino acids and derivatives` or due to the ChEBI ontology lacking the necessary relationships between these terms. This confusion in the evaluation underscores the necessity for an alternative evaluation technique.

5.9.2 Evaluation using GPT4

As the second evaluation technique, we used the Generative Pre-trained Transformer 4 or GPT-4 language model, as the state-of-the-art generative language model exceeding in question answering tasks [AAA⁺23]. A part of this evaluation is done to see if GPT-4 could be used as a scientific fact-checker in specific fields of study. This ability has been studied in multiple research before [QB24]. In this experiment, the *Davinci* engine is specifically used as a fact-checker to see if the SPEC prediction belongs to the SC class of substrates. This has been done by prompting GPT-4 with the following prompt:

```
messages=[
  "role": "system", "content": "You are a fact-checking expert chemist. Use Google
  for research. You must respond with one word either 'True' or 'False'. Evaluate the
  truthfulness of the statement.,
  "role": "user", "content": f"{specific} is considered {general} in chemistry ]
```

In the system role, we specified the task as fact-checking in chemistry, demanding the language model to respond with one word of "True" or "False" evaluating the truthfulness of each prompt sentence. In the user role, we ask the question of whether a specific prediction made by the TooT-Triplet is considered a part of the more general label from

the SC dataset. Please note that this fact-checking technique is not a part of our TooT-Triplet model and is utilized in the external validation.

Table E.4.1 presents the results for the evaluation of the SC dataset, where the column “GPT-4” shows the answers provided by GPT-4 to the described prompt for each row of the table. However, this evaluation technique has certain limitations. Firstly, the GPT-4 model produces contradictory results due to the inherent randomness in the model.

For instance, the question “Is ‘proton’ considered ‘inorganic cation’ in chemistry?” has received both “True” and “False” answers (e.g., sequences 075947 and P81449 in Table E.4.1). Moreover, in some cases, GPT-4 fails to answer the question correctly. Table 30 lists the sequences for which the SC label and the Triplet prediction align as fine-grained classes, but GPT-4 has incorrectly categorized them as non-matching classes.

Table 30: SC dataset proteins with matching TooT-Triplet prediction; incorrectly categorized by GPT4.

Sequences	SC label	Prediction	Compound Formula
Q9SZY7, Q9HBG4, Q91V37, P22203, P43457, P38925, O75947	inorganic cation	proton	H(1+)
Q9P5N0	other organonitrogens	heme	C49H56O6N4Fe
P0AAH4	other organonitrogens	1-4-butane diammonium	C4H14I2N2
P9WGP5, P9WGP3, Q9X1R4, O06446, P0DJP3, P10408	other organonitrogens	protein	-
P48751	organic anion	hydrogencarbonate	H2CO3(1-)
Q9JJZ1, Q9UGQ3, P11169, P11168	organo-oxygens	dehydroascorbide(1-)	C6H6O6
P40004, Q76EJ3, Q93890, Q18779	organo-oxygens	UDP-N-acetyl-alpha-D-glucosamine(2-)	C17H25N3O17P2(2-)
P78382	organo-oxygens	CMP-N-acetyl-beta-neuraminate(2-)	C20H31N4O16P
Q96A29, Q84L08, Q9VHT4	organo-oxygens	GDP-fucose	C16H25N5O15P2
Q12520, O64503, Q94B65, Q0D9V6, Q5W7C1	organo-oxygens	UDP-D-glucose(2-)	C15H22N2O17P2(2-)

Continued on next page

Table 30: SC dataset proteins with matching TooT-Triplet prediction; incorrectly categorized by GPT4.

Sequences	SC label	Prediction	Compound Formula
Q8RXL8	organo-oxygens	UDP-alpha-D-xylose(2-)	C14H20N2O16P2(2-)
Q9Y2P4	organo-oxygens	oleate	C18H34O2
P33897	organo-oxygens	fatty acyl-CoA	C23H38N7O17P3S
P0AC96	organic anion	gluconate	C6H11O7
P10384	organo-oxygens	long-chain fatty acid	CH3(CH2) _n COOH
Q12375, Q9Y619, Q93XM7	amino acids and derivatives	L-ornithinium(1+)	C5H12N2O2
P0AAH8, P0AGH5	other organonitrogens	1-4-butanediammonium	C4H14I2N2
P0AF16	organo-oxygens	lipid -linked peptidoglycan	C55-OH
O95342	organo-oxygens	bile acid	C24H40O5

The table presents the sequence IDs in the TooT-SC dataset which have a matching prediction in the TooT-SC and TooT-Triplet models, with TooT-Triplet prediction as a fine-grained specific substrate belonging to the group of substrates in the SC-label. The GPT4 model has failed to detect the match between the SC label and TooT-Triplet prediction.

Since the second automatic method for evaluating the results of the TooT-Triplet predictions for the SC dataset was ineffective, a manual evaluation of the results was conducted. The manual evaluation identified five sequences in the SC dataset that are incorrectly labeled but correctly predicted by TooT-Triplet. The SwissProt descriptions agree with the TooT-Triplet predictions, as presented in Table 31.

Table 32 presents the incorrect predictions of TooT-Triplet for SC data samples. The column ‘Top Neighbors’ presents the labels of the three closest neighbors to the test sequence along with their distances. This table indicates that the samples of the majority class ‘proton’ confuse the model in prediction. The column ‘Proton Excluded Top Neighbors’ presents the results of the same model trained on a similar dataset excluding the proton sequences. The details of the experiment are presented in Section 5.8.3.

Table 31: Proteins with incorrect SC dataset labels predicted correctly by TooT-Triplet.

Sequence	SC label	SwissProt description	Triplet prediction
Q12325	organic anion	Sulfate permease 2 in <i>Saccharomyces cerevisiae</i> (strain ATCC 204508/S288c)	sulfate
P38359	organic anion	Sulfate permease 1 in <i>Saccharomyces cerevisiae</i> (strain ATCC 204508/S288c)	sulfate
P53394	organic anion	Putative sulfate transporter YPR003C in <i>Saccharomyces cerevisiae</i> (strain ATCC 204508/S288c)	sulfate
P53838	inorganic anion	Boron transporter 1 in <i>Saccharomyces cerevisiae</i> (strain ATCC 204508/S288c)	boric acid
Q5T3U5	inorganic anion	ATP-binding cassette sub-family C member 10 in <i>Homo sapiens</i> (Human)	glutathione derivative

The table presents the sequence IDs in the TooT-SC dataset which have an incorrect label as their substrate class (column SC label). These sequences are predicted to correct fine-grained specific substrates using the TooT-Triplet model (column Triplet Prediction). The reference for the label accuracy is the SwissProt-reviewed description (column SwissProt description).

Table 32: Sequences of SC dataset with the incorrect prediction using TooT-Triplet.

Sequence ID	Top Neighbors	Top Neighbor Distance	ProtonEx Top Neighbor	Top Neighbor Distance in ProtonEx model	True label
Q8BWY7	citrate(3-) calcium(2+) calcium(2+)	2.20 2.22 2.22	zinc(2+) zinc(2+) sodium(1+)	1.47 16.80 30.56	zinc(2+)
Q9FGS5	potassium(1+) nitrate chloride	1.36 1.40 1.62	nitrate nitrate potassium(1+)	1.55 19.73 21.58	nitrate
Q56ZZ7	proton proton proton	0.08 1.23 1.28	dehydroascorbide(1-) dehydroascorbide(1-) dehydroascorbide(1-)	21.79 23.07 23.68	Plastidic glucose (unknown)
Q8L7Z9	proton 2-oxoglutarate(2-) proton	1.73 1.77 1.79	L-glutamate(1-) L-ascorbate 2-oxoglutarate(2-)	0.94 23.65 23.87	L-glutamate(1-)
P64550	proton urea proton	0.15 1.95 2.00	urea urea urea	28.18 28.45 28.56	L-alanine

Continued on next page

Table 32: Sequences of SC dataset with the incorrect prediction using TooT-Triplet.

Sequence ID	Top Neighbors	Top Neighbor Distance	ProtonEx Top Neighbor	Top Neighbor Distance in ProtonEx model	True label
P0AGH3	glycine betaine	2.08	1-4-butane diammonium	1.91	1-4-butane di-ammonium
	1-4-butane diammonium	2.08	nitrate	28.39	
	potassium(1+)	2.09	1-4-butane diammonium	29.41	
P77328	L-ascorbate	1.74	xanthine	2.63	xanthine
	L-methionine zwitterion	1.80	boric acid	27.58	
	L-ascorbate	1.82	boric acid	27.60	
P39719	calcium(2+)	1.69	flavin adenine dinucleotide	2.40	flavin adenine dinucleotide
	potassium(1+)	1.70	flavin adenine dinucleotide	10.88	
	potassium(1+)	1.71	flavin adenine dinucleotide	11.63	
P12428	cholesterol	0.96	guanine	0.58	guanine
	cholesterol	1.01	cholesterol	13.03	
	cholesterol	1.08	cholesterol	13.70	
	ammonium	0.99	boric acid	1.11	
Q03263	proton	1.60	boric acid	8.27	boric acid
	proton	1.62	ammonium	16.82	
Q93646	potassium(1+)	1.64	acetylcholine	1.59	acetylcholine
	proton	1.83	potassium(1+)	25.90	
	potassium(1+)	1.92	calcium(2+)	26.47	

The table shows the sequence IDs with their ‘True labels’ (last column) and the incorrect predictions of the TooT-Triplet model. The ‘Top Neighbors’ and ‘Top Neighbor Distance’ columns show the top three nearest neighbor labels for each sequence using TooT-Triplet encodings along with their distances calculated using the Euclidean metric. The table also compares the predictions of the TooT-Triplet to the TooT-Triplet-ProtonEx model which is trained and tested in the absence of the majority class **proton** (dataset UniProt-SPEC-100-ProtonExcluded). The columns ‘ProtonEx Top Neighbor’ and ‘Top Neighbor Distance in ProtonEx’ show the top three nearest neighbor labels for each sequence using TooT-Triplet-ProtonEx encodings along with their distances trained and tested the absence of the majority class (distances calculated using Euclidean metric).

Additionally, Table 33 presents the cases where both SC label and TooT-Triplet prediction could be correct. Most of these protein sequences are predicted as **proton**, for which the ‘Additional Information’ column indicates proteins interacting with more than one substrate. Same as the previous table, the ‘Proton Excluded TooT-Triplet Prediction’ column is added to analyze the performance of the TooT-Triplet model in the ‘proton’ class absence. Both Tables 32 and 33 show prediction improvement in the absence of the majority class ‘proton’. Figure 48 in Appendix E.4.1 presents each SC class with its corresponding TooT-Triplet fine-grained substrate predictions.

Table 33: SC dataset entries with unmatching prediction where both could apply.

Sequence ID	SC label	TooT-Triplet Prediction	Proton Excluded TooT-Triplet Prediction	Additional Information
Q9P3U6	organo-oxygens	proton	dehydroascorbide(1-)	GO:0005351 (carbohydrate:proton symporter activity), GO:0005358 (high-affinity glucose: proton symporter activity)
Q10PW9	organo-oxygens	proton	boric acid	alternative name: Sugar:proton symporter MST4
A0PJK1	organo-oxygens	sodium(1+)	sodium(1+)	Sodium/mannose cotransporter SLC5A10
P42833	organo-oxygens	proton	dehydroascorbide(1-)	GO:0005351 (carbohydrate:proton symporter activity), GO:0005354 (galactose transmembrane transporter activity)
P54854	organo-oxygens	proton	dehydroascorbide(1-)	GO:0005351 (carbohydrate:proton symporter activity)
P39932	organo-oxygens	proton	dehydroascorbide(1-)	GO:0005351 (carbohydrate:proton symporter activity)
P41036	organo-oxygens	proton	sodium(1+)	GO:0015538 (sialic acid:proton symporter activity)
Q2M3M2	organo-oxygens	sodium(1+)	sodium(1+)	Sodium/glucose cotransporter 4
P31639	organo-oxygens	sodium(1+)	sodium(1+)	Sodium/glucose cotransporter 2, GO:0005412 (glucose:sodium symporter activity)
P0AEP1	organo-oxygens	proton	nitrate	Galactose-proton symporter
P23585	organo-oxygens	proton	dehydroascorbide(1-)	GO:0005351 (carbohydrate:proton symporter activity)
P13866	organo-oxygens	sodium(1+)	sodium(1+)	name: Sodium/glucose cotransporter 1
P02920	organo-oxygens	proton	potassium(1+)	GO:0005351 (carbohydrate:proton symporter activity)
Q9NRA2	organo-oxygens	proton	sodium(1+)	GO:0015538 (sialic acid:proton symporter activity)
P11551	organo-oxygens	proton	nitrate	name: L-fucose-proton symporter
O23492	organo-oxygens	proton	dehydroascorbide(1-)	Alternative names: Myo-inositol-proton symporter INT4

Continued on next page

Table 33: SC dataset entries with unmatching prediction where both could apply.

Sequence ID	SC label	TooT-Triplet Prediction	Proton Excluded TooT-Triplet Prediction	Additional Information
P15325	organo-oxygens	proton	urate salt	GO:0005351 (carbohydrate:proton symporter activity)
Q46839	organic anion	proton	gamma-aminobutyric acid zwitterion	GO:0015129 (lactate transmembrane transporter activity) GO:0015295 (solute:proton symporter activity)
P0AGE4	amino acids and derivatives	sodium(1+)	sodium(1+)	GO:0005295 (neutral L-amino acid:sodium symporter activity)
P0AER8	organic anion	sodium(1+)	sodium(1+)	name: Sodium/glutamate symporter
P07117	amino acids and derivatives	sodium(1+)	sodium(1+)	name: Sodium/proline symporter
P39276	other organonitrogens	proton	nitrate	GO:0015333 (peptide:proton symporter activity)
Q16348	other organonitrogens	proton	nitrate	GO:0015333 (peptide:proton symporter activity)
O01840	other organonitrogens	proton	nitrate	GO:0071916 (dipeptide transmembrane transporter activity), GO:0015252 (proton channel activity)
Q9JIP7	other organonitrogens	proton	nitrate	GO:0015333 (peptide:proton symporter activity)
O95255	inorganic anion	glutathione derivative	glutathione derivative	GO:0015431 (ABC-type glutathione S-conjugate transporter activity)
Q12908	organo-oxygens	sodium(1+)	sodium(1+)	name: Ileal sodium/bile acid cotransporter
Q96EP9	organo-oxygens	sodium(1+)	sodium(1+)	name: Sodium/bile acid cotransporter 4
O08705	organo-oxygens	sodium(1+)	sodium(1+)	name: Hepatic sodium/bile acid cotransporter
Q12407	organo-oxygens	proton	biotinate	GO:0005351 (carbohydrate:proton symporter activity)

The table presents the protein sequence IDs where the TooT-Triplet prediction and SC label do not match. However, both labels could apply to the sequence according to the additional information from SwissProt (last column). Also, the prediction for TooT-Triplet ProtonEx is included to compare.

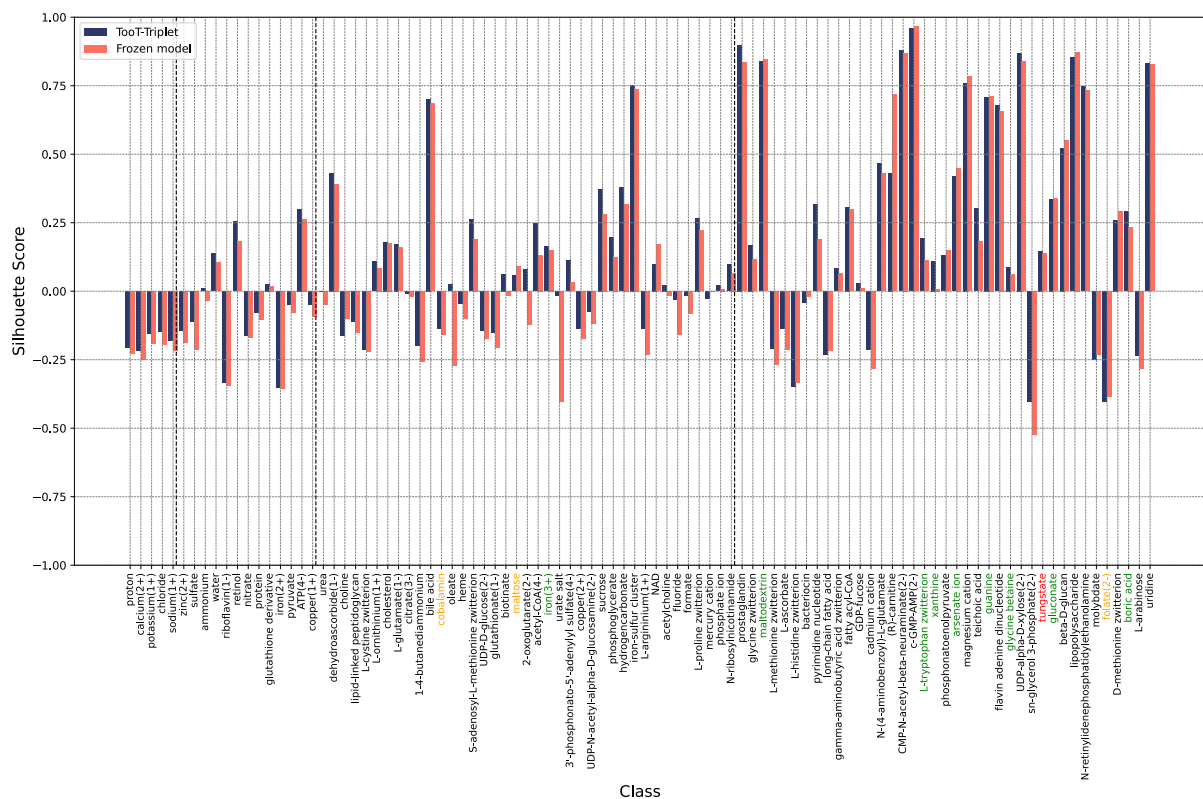


Figure 26: Silhouette score comparison of TooT-Triplet (offline, Manhattan) with Baseline embedding space (frozen) for each class in UniProt-SPEC-100.

The silhouette score calculated for each class using both Baseline with frozen weights and the TooT-Triplet model trained with the Manhattan distance metric in loss function using offline sampling strategy. The silhouette score ranges between -1 and 1, presenting the lowest cluster consistency to the highest. The x-axis presents the classes in the UniProt-SPEC-100 dataset sorted by their size, from the largest class, left, to the smallest class, right. The majority classes, middle classes, and minority classes with less than 10 and less than 3 samples are separated by dashed vertical lines. The green color in the x-axis labels indicates the change of the class from unpredicted in Baseline 1 to predicted in the TooT-Triplet (offline, Manhattan) model. The color orange in the x-axis labels indicates a change of the class from predicted in Baseline 1 to unpredicted in the TooT-Triplet (offline, Manhattan) model. The color red indicates the unpredicted classes in both models. The y-axis presents the Silhouette score for each class based on the model.

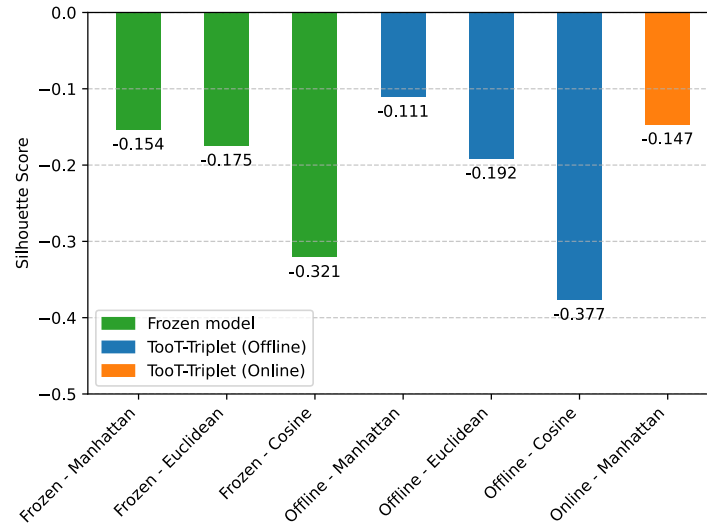


Figure 27: Overall silhouette score for three model groups.

The average silhouette score over all classes of UniProt-SPEC-100 is presented for the encodings of the samples extracted from three model groups: baseline (frozen), TooT-Triplet offline, and online models trained with different distance metrics. The silhouette score ranges between -1 and 1, presenting the lowest cluster consistency to the highest.

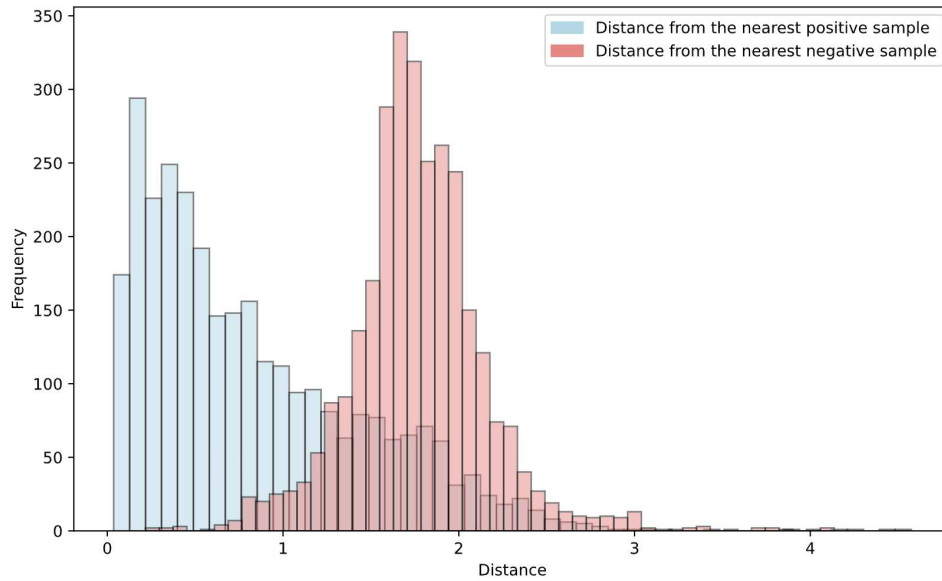


Figure 28: (Offline) Distance distributions for negative and positive nearest neighbor. The graph shows the nearest neighbor distances on TooT-Triplet (offline) encodings. The y-axis shows the frequency of samples with a specific distance (in the x-axis) from their nearest neighbor with the same label or positive (blue) and their nearest neighbor with different labels or negative (red).

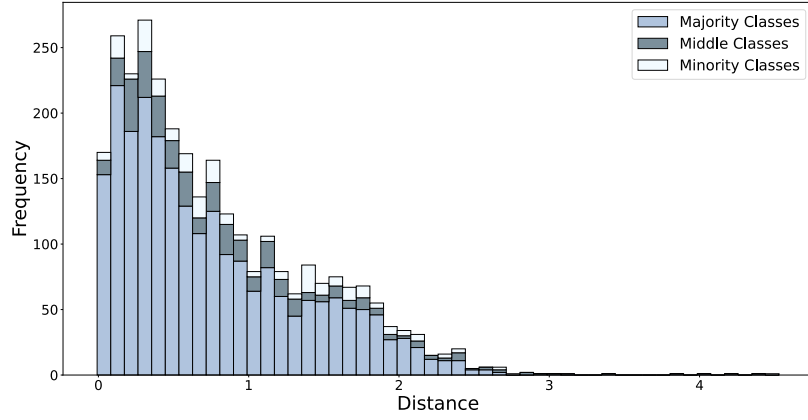


Figure 29: (Offline) Distance distributions for positive nearest neighbor.

The graph shows the nearest neighbor distances on TooT-Triplet (offline) encodings. The positive distribution plot shows the frequency of samples (y-axis) with a specific distance (in the x-axis) from their nearest neighbor with the same label or positive samples. The samples from the majority, minority, and middle classes are distinct in different shades of blue.

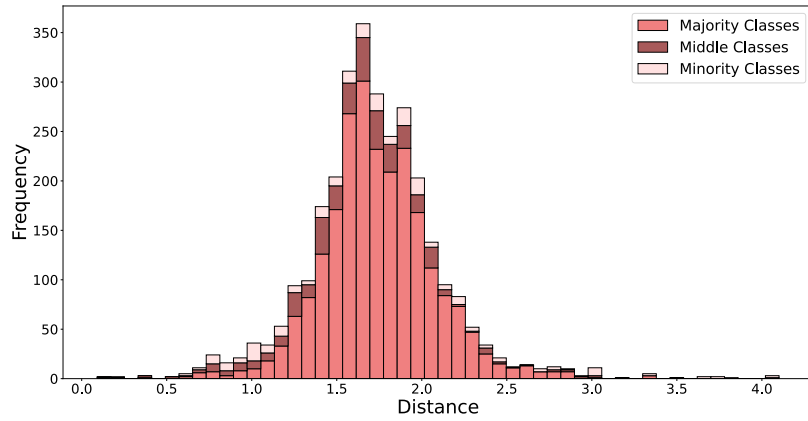


Figure 30: (Offline) Distance distributions for negative nearest neighbor.

The graph shows the nearest neighbor distances on TooT-Triplet (offline) encodings. The negative distribution plot shows the frequency of samples (y-axis) with a specific distance (in the x-axis) from their nearest neighbor with a different label or negative samples. The samples from the majority, minority, and middle classes are distinct in different shades of red.

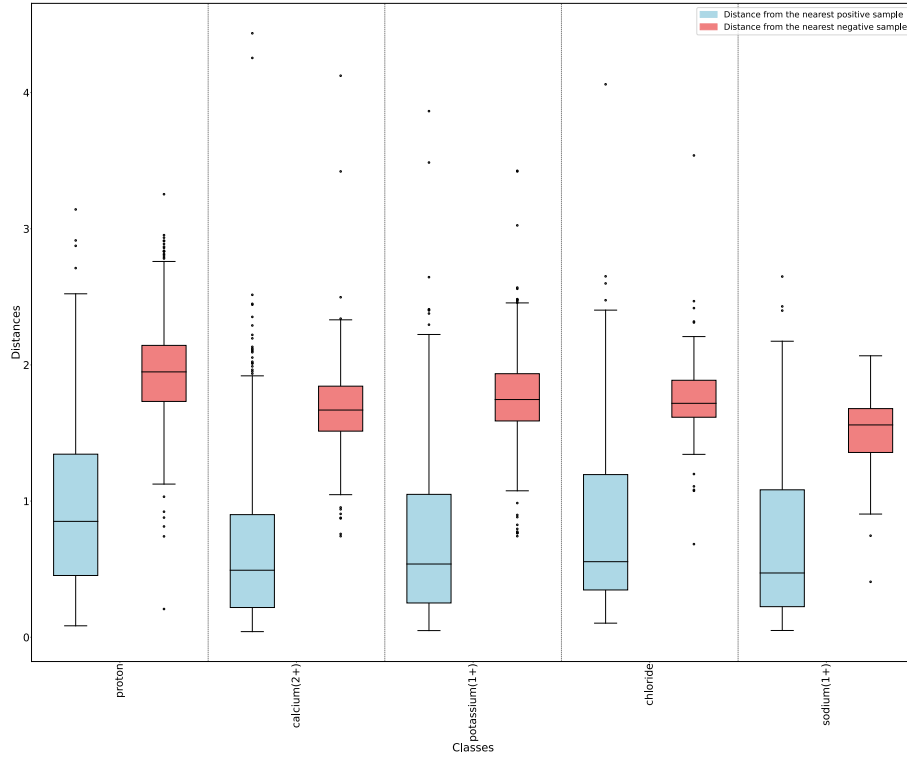


Figure 31: Class separated distributions of distances for the negative and positive nearest neighbor in majority classes encodings extracted from TooT-Triplet (offline).

The distance distribution of positive (same label) and negative (different labels) nearest neighbors in the UniProt-SPEC-100 dataset is presented separately for each majority class. The distance distribution of the nearest neighbors with the same label is presented in blue and the distance distribution of the nearest neighbor with different labels is presented in red.

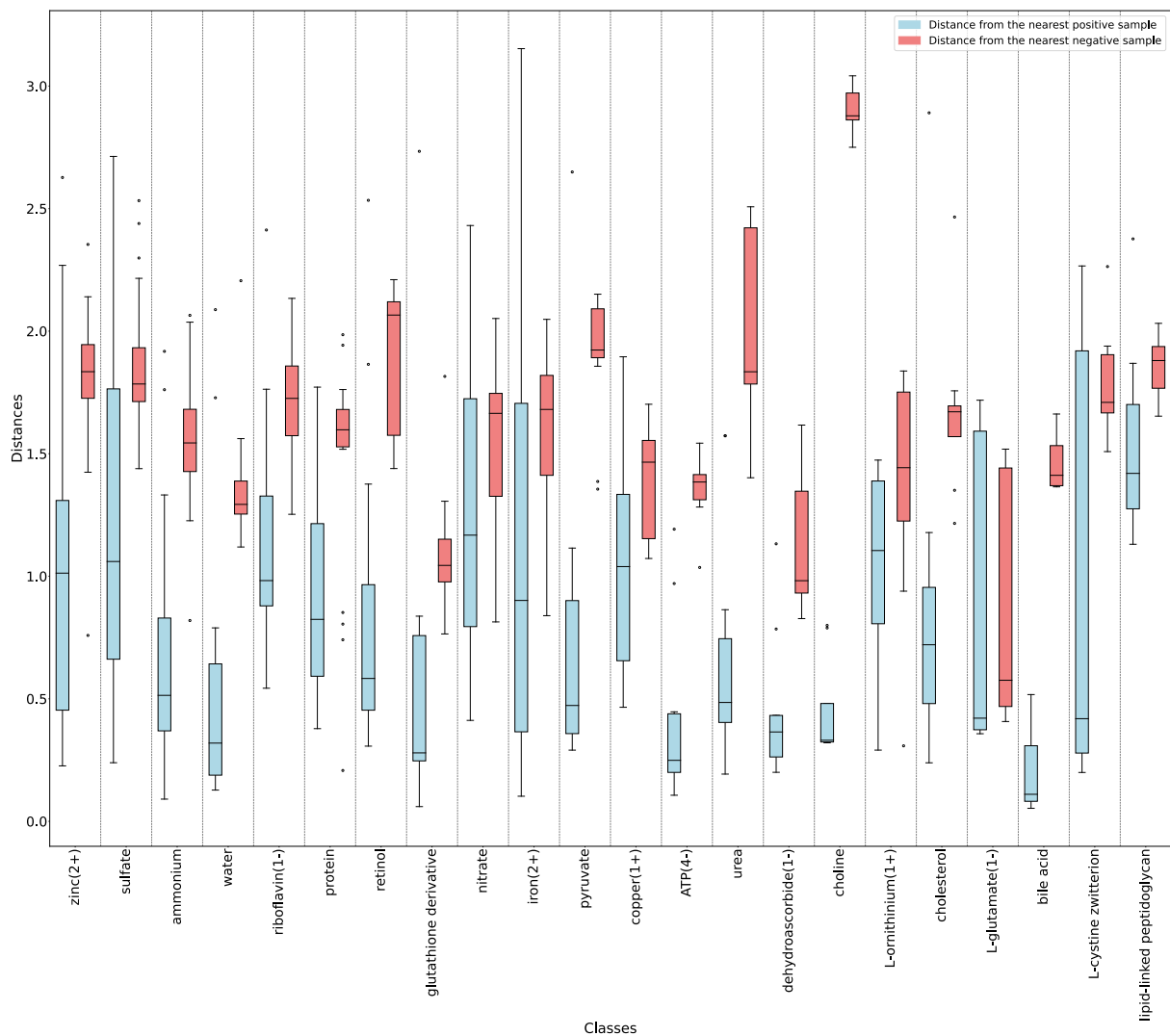


Figure 32: Class separated distributions of distances for the negative and positive nearest neighbor in middle classes encodings extracted from TooT-Triplet (offline).

The distance distribution of positive (same label) and negative (different labels) nearest neighbors in the UniProt-SPEC-100 dataset is presented separately for each middle class. The distance distribution of the nearest neighbors with the same label is presented in blue and the distance distribution of the nearest neighbor with different labels is presented in red.

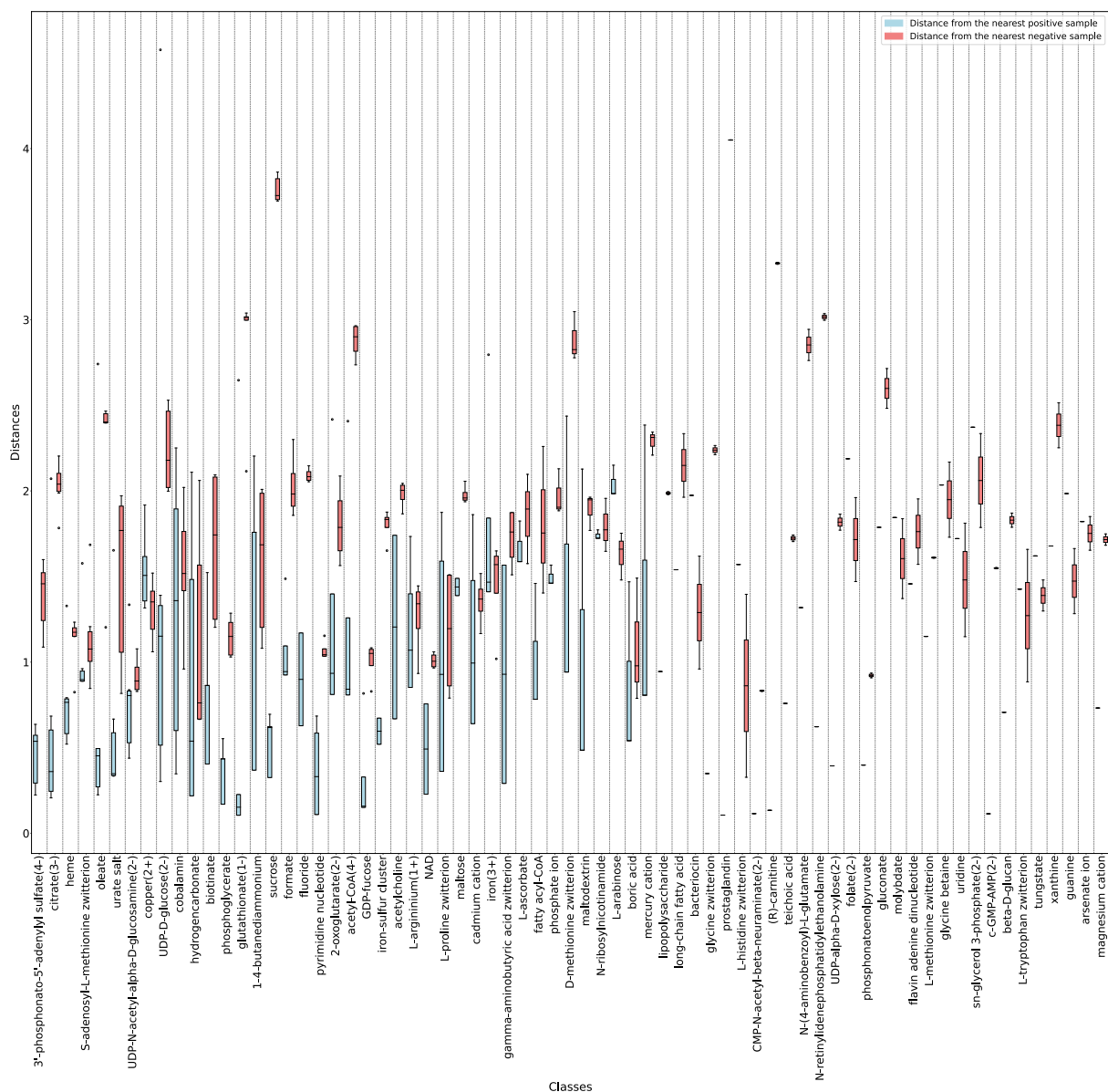


Figure 33: Class separated distributions of distances for the negative and positive nearest neighbor in minority classes encodings extracted from TooT-Triplet (offline).

The distance distribution of positive (same label) and negative (different labels) nearest neighbors in the UniProt-SPEC-100 dataset is presented separately for each minority class. The distance distribution of the nearest neighbors with the same label is presented in blue and the distance distribution of the nearest neighbor with different labels is presented in red.

1 10 20 30 40 50 60 70 80 90			
M A C P L D Q A I G L L V A I F H K Y S G R E G D K H T L S K K E L K E L I Q K E L T I G S K L Q D A E I A R L M E D L D R N K D Q E V N F Q E Y V T F L G A L A L I Y N E A L K G			
TYPE	ID	POSITION(S)	DESCRIPTION
Binding site			
► Binding site	28		Ca ²⁺ 1 (UniProtKB ChEBI) Curated
► Binding site	33		Ca ²⁺ 1 (UniProtKB ChEBI) Curated
► Binding site	61		Ca ²⁺ 2 (UniProtKB ChEBI) PROSITE-ProRule Annotation
► Binding site	63		Ca ²⁺ 2 (UniProtKB ChEBI) PROSITE-ProRule Annotation
► Binding site	65		Ca ²⁺ 2 (UniProtKB ChEBI) PROSITE-ProRule Annotation
► Binding site	67		Ca ²⁺ 2 (UniProtKB ChEBI) PROSITE-ProRule Annotation
► Binding site	72		Ca ²⁺ 2 (UniProtKB ChEBI) PROSITE-ProRule Annotation

Figure 34: Binding sites of protein sequence P06703.

The figure shows seven different positions in the amino acid chain as the binding sites of the transport protein sequence P06703 binding to calcium ion Ca(2+).

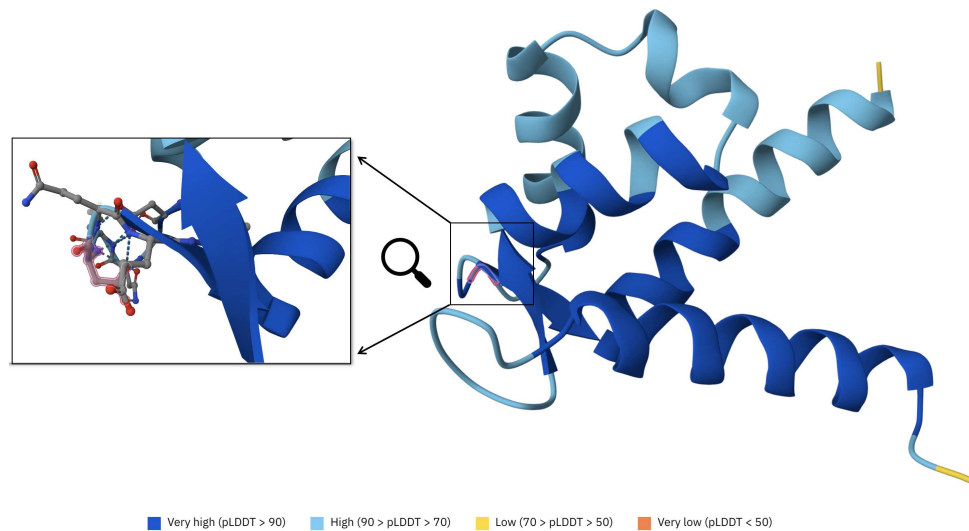


Figure 35: AlphaFold 3D structure and binding site of P06703 on position 65.

The 3-D structure of the transport protein P06703 is presented with a focus on the amino acid “Aspartic Acid” in position 65. This amino acid is the only binding site of this sequence with its information captured as a high level of attention in TooT-Triplet (offline) model encodings. (AlphaFold produces a per-residue model confidence score (pLDDT) for 3-D structure prediction between 0 and 100, presented by color codes in the predicted structure from dark blue as very high confidence to orange as very low confidence)

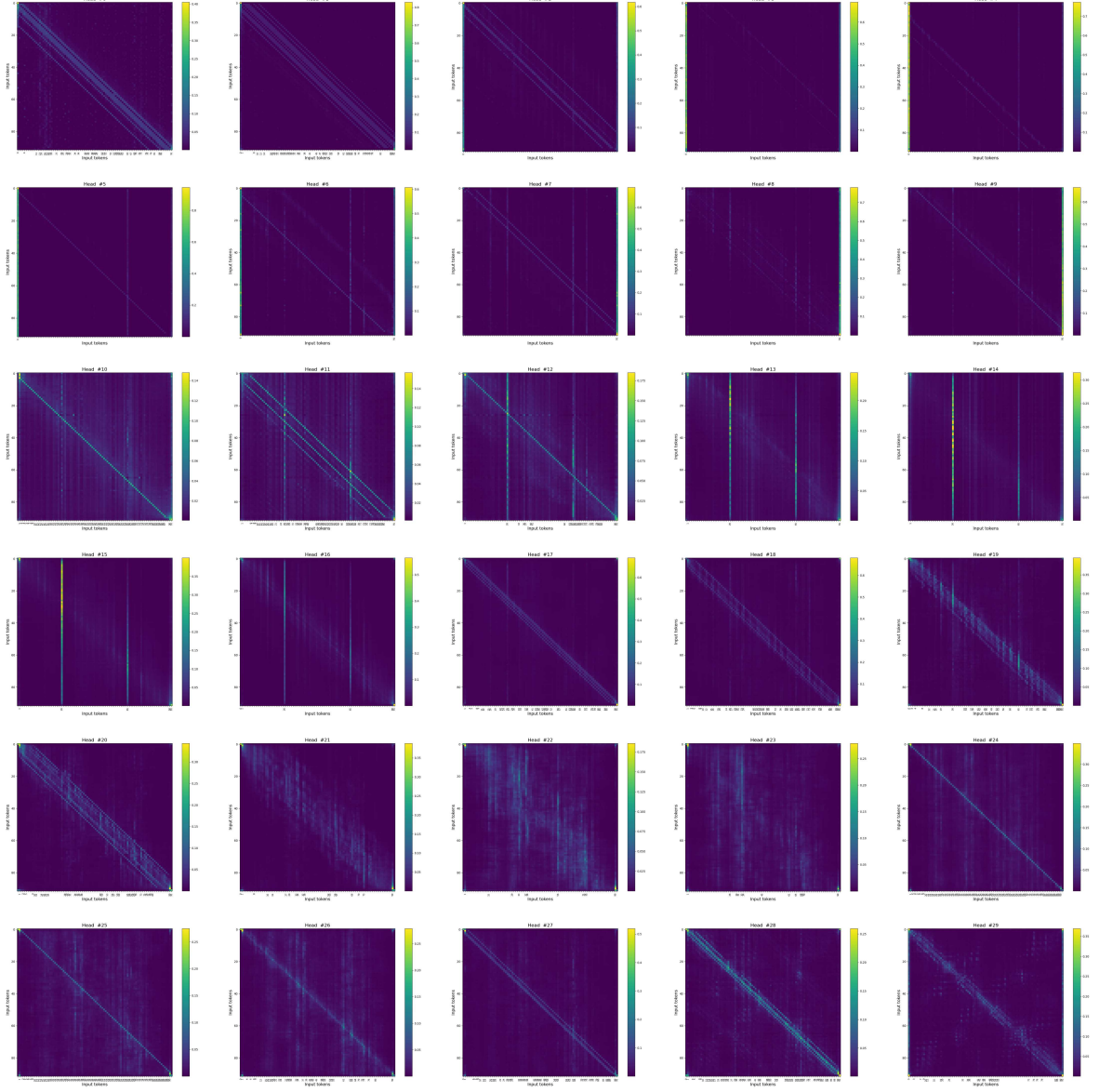


Figure 36: Heatmaps of 30 attention heads in TooT-Triplet for P06703 over 16 layers. Each matrix shows one attention head with the average attention values over 16 layers for each token in the full sequence. The heads are sorted by head number, starting with head 0 top, left ending with head number 29 bottom right. The x and y axes in each matrix present the chain of amino acids as the input token. Attention values are presented by a range of colors from low to high values, darker to lighter.

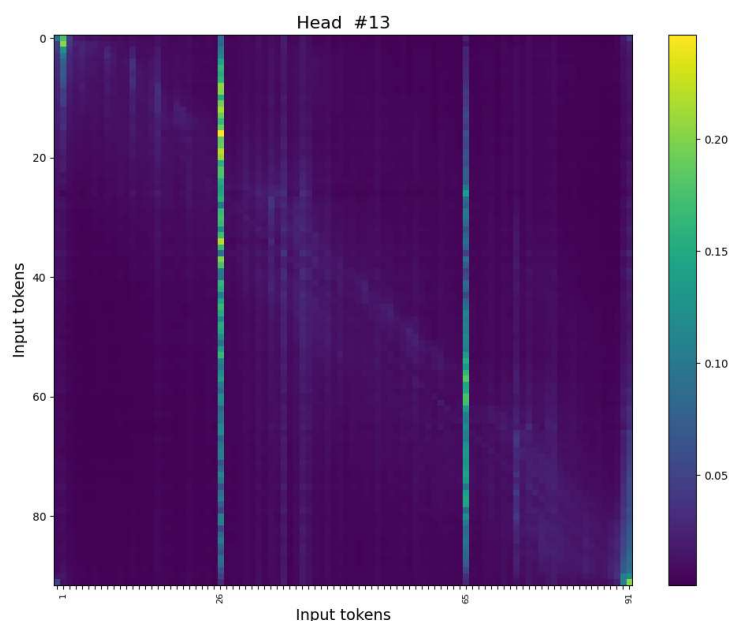


Figure 37: Heatmap of head 13 in TooT-Triplet for P06703 over 16 layers. The figure shows the heatmap of attention head number 13 with the average attention values over 16 layers for each token in the full sequence. This head shows high values of average attention in amino acid positions 26 and 65 in the sequence. Amino acid number 65 is a binding site (Figure 34) while amino acid number 26 is not a binding site.

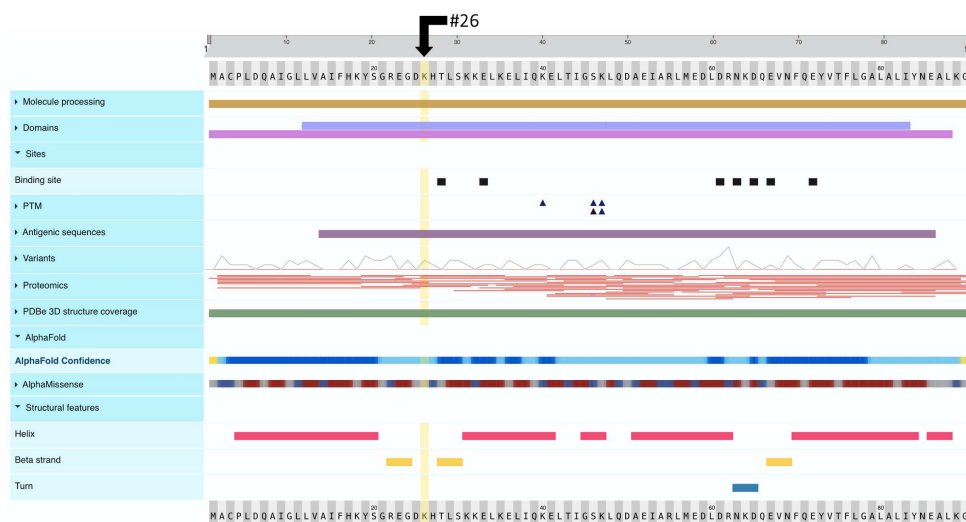


Figure 38: Analysis of the sequence P06703 from UniProt-KB. The beige column shows the selected amino acid number 26, Lysine (K), with the highest attention score. The analysis shows no specific values in this position while it captures the highest values in attention heads (Figures 37 and 36).

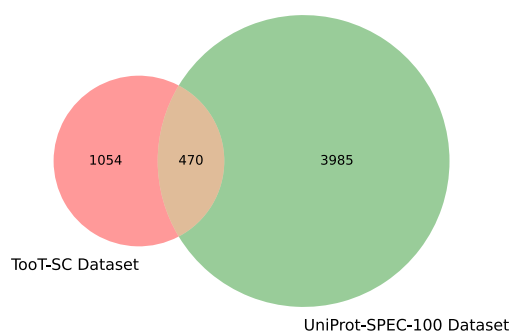


Figure 39: Venn diagram of TooT-SC and UniProt-SPEC-100 dataset sequences. The diagram shows the samples in the UniProt-SPEC-100 dataset in green and the TooT-SC dataset in red. There are 470 samples in common in the two datasets.

5.10 Conclusion

The main question this research project follows is:

“Given a transmembrane transport protein from a longtailed imbalanced dataset can a model predict the specific substrate it carries?”

The TooT-SPEC model proposes a solution for the prediction of 83 from 96 specific substrate classes (Section 3.7). However, according to the results, the TooT-SPEC model is unable to provide predictions for 13 out of 96 classes, mostly in the minority classes in the dataset with sample sizes less than 10 (Table 13).

The TooT-Triplet model is implemented to overcome the limitations of the TooT-SPEC model. It explores the effect of triplet sampling and loss function on the one-shot classification of transmembrane transport proteins using the Prot-BERT-BFD protein language model. Two strategies of offline and online triplet sampling along with three distance metrics of Manhattan, Euclidean, and Cosine similarity are implemented (Sections 5.3.1, 5.3.2, 5.7.1, and 5.7.2).

The results suggest the effectiveness of the offline sampling strategy over both random sampling and online triplet sampling (Table 26). The silhouette score indicates higher consistency in minority classes compared to the majority (Figure 26). Additionally, it identifies Manhattan distance as the most efficient distance metric for high-dimensional feature space classifications (Figure 27). However, the performance difference was not statistically significant according to the pairwise Student’s t-test (Table 21) among 3-fold cross-validation results for the models with different distance metrics.

The TooT-Triplet offline model outperforms the online model (Table 26) in all metrics. The comparison of the TooT-Triplet offline model to the TooT-SPEC model (trained using random sampling) does not show a significant difference in MCC (Table 26). The results show an improvement in effective class coverage of both online and offline strategies, decreasing the number of unpredicted classes from 13 to 5 compared to TooT-SPEC. TooT-Triplet demonstrates the effectiveness of reducing intra-class distances while expanding inter-class distances in the one-shot classification of transmembrane transport proteins. The distance of a test sample and its nearest neighbor is proposed as a metric of reliability (Section 5.8.2). Furthermore, the binding site analysis shows attention partially captures binding site information (Section 5.8.4).

Chapter 6

TooTranslator: Zero-shot classification of *de novo* transport proteins

6.1 Introduction

The research question which this project follows is that “*Given a de novo transmembrane transport protein can a model predict the specific substrate it carries?*” This question has two aspects: is the protein *de novo* protein and what is the predicted specific substrate. By definition, *de novo* protein sequences are not structurally similar to known protein sequences. This does not necessarily result in different roles; meaning that *de novo* transport protein might carry either previously known or new substrates.

When predicting the specific substrate carried by a transport protein, the model has two possibilities: (A) the substrate exist in the training set; the *seen* classes; and (B) the substrate is new, not included in the training set; the *unseen* classes.

In a standard classification scenario, the model is trained to predict that test samples belong to one of the classes included in the training set. However, due to the nature of this problem, where *de novo* proteins may transport substrates not in the training set, we adopt a zero-shot classification approach. Zero-shot classification allows the model to predict that the class of a sample belongs to an unseen category. This method suits our task, as *de novo* test samples might carry entirely new substrates, making traditional classification approaches insufficient.

In this study, the UniProt-SPEC-100 dataset is the primary dataset for training the model. The first group of labels, group (A), includes 96 specific substrate classes. However, it is noted that transmembrane transport proteins may facilitate more substrates

than these 96, some of which were not identified during the construction of the UniProt-SPEC dataset. This could be due to missing annotations (in ChEBI or Gene Ontology) or inconsistencies between annotation systems. As discussed in Section 3.5.2, a threshold of 3 samples per class was applied, excluding 59 classes with fewer than 3 samples from the dataset. Additionally, new molecular entities continue to be discovered, meaning that more substrates will be identified and added through future research. This second group of substrates, not included in UniProt-SPEC training, is referred to as group (B).

Definition 6.1.1. *A **regression** task in machine learning is a supervised learning problem where a model is trained to predict a continuous numerical value based on input features.*

The core objective of this project is to evaluate whether training the model on a limited set of protein-substrate pairs can lead to improved representations for unseen classes, enabling an accurate zero-shot classification. Specifically, we transform the problem by replacing the class labels with their detailed descriptions (ChEBI definitions and SMILES strings), which are then vectorized using two BERT encoders. This shifts the task from a traditional classification problem to a regression problem, where the model learns to translate protein vectors into label vectors. By training the model on a restricted dataset and updating the network weights for seen classes, we aim to determine if it can generalize to unseen substrates.

The hypothesis to investigate is that samples from unseen classes will be closer to their label vectors in the learned embedding space, even without direct training. This experiment seeks to explore if incorporating language models, such as ProtBERT-BFD, SciBERT, and ChemBERTa, into a unified embedding space facilitates zero-shot learning. Additionally, understanding if mapping the embedding spaces of separate BERT language models into one mutual space could offer insights into the broader potential of integrating multiple LMs for effective zero-shot classification in biological tasks.

The initial design of the TooTranslator model aimed to map two embedding spaces: ProtBERT-BFD (for protein sequences) and SciBERT (for label definitions from ChEBI) into a shared embedding space, combining both protein and label vectors. However, testing this approach yielded unsatisfactory results. To address this, we incorporated SMILES strings for each label into the label vector encodings. Simplified Molecular Input Line Entry System (SMILES), a line notation for describing the structure of chemical compounds, is also sourced from ChEBI . We used ChemBERTa-2 to encode the SMILES strings and combined this with the definition vector to form a single, unified label vector.

This manuscript presents the results of mapping ProtBERT-BFD protein representations to a combination of SciBERT and ChemBERTa-2 label representations.

The rest of this chapter is designed as follows. First, the problem of *de novo* sequence prediction is defined in Section 6.2. The methodology section (Section 6.3) describes two phases of vectorization/projection and classification in Sections 6.3.1 and 6.3.4. The results and discussion points are explored in Sections 6.4 and 6.5. Finally, the chapter conclusion is presented in Section 6.6. The dataset and code for this project are publicly available on GitHub at the following link (<https://github.com/simaataei/TooTranslator>)

6.2 Problem statement: *de novo* substrate prediction

Given a labeled dataset of n transmembrane transport protein sequences as:

$$D = \{(p_1, l_1), (p_2, l_2), (p_3, l_3), \dots, (p_n, l_n)\},$$

p_i representing the protein sequences, as a chain of amino acids, associated to a label, l_i , in the label set L of seen classes; build a model, $f(p)$, to classify samples of the test set, D_{test} , associated with a label set as:

$$S = \{l_1, l_2, l_3, \dots, l_n, s_1, s_2, s_3, \dots, s_z\}$$

with l_i as labels in the training set and s_i as other specific substrate classes not included in the training set. The model $f(p)$ is used to learn the features and generalize them to the unseen classes in the test set. The only condition applied on protein samples in the D_{test} accepted by the model is that the sample is a membrane transport protein transporting a substrate in ChEBI. In other words, a transportome of any organism could be presented to the model and the model predicts the specific substrate carried by each protein.

6.3 Methodology

To solve the problem of *de novo* sequence substrate prediction, we propose the TooTranslator model. This model “translates” the primary structure of transmembrane transport proteins to their carried substrate. The novelty of TooTranslator is in jointly modeling protein sequences and their substrate class information resulting in a translation machine to calculate the nearest label for a given protein sequence. This model modifies the protein sequence encoder weights based on seen <protein sequence, label> pairs aiming to produce reliable encoding vectors for the unseen classes of *de novo* proteins.

Definition 6.3.1. A *ChEBI definition* is a short verbal definition of the entry provided by ChEBI ontology for all new entries annotated after June 2009 [Eur].

Definition 6.3.2. *SMILES* is a standardized ASCII-based notation for representing chemical structures, widely used in cheminformatics and recognized by IUPAC [Wei88].

TooTranslator relies on the sequence of transmembrane transport proteins and the ChEBI definition, and SMILES chemical strings of the label. It has two main components: *vectorization* and *projection*. This model vectorizes the protein sequence and the label information separately, forming two encoding spaces: protein and label spaces. In the projection component, the model learns to project the protein vector spaces into the label space, translating the primary structure of the protein to the carried substrate.

In zero-shot classification, the model is expected to predict new classes without prior training on them. In other words, the test set could include samples from classes unseen in the training set. In the testing process, the test sequences could be predicted into labels of groups of seen and unseen substrate classes or groups (A) and (B). In this study, Group (A) are the classes from the UniProt-SPEC-100 dataset called TS1, and Group (B) are classes in TS2 (details in Section 6.3.4). Figure 40 describes the process of training and testing for the TooTranslator model. This project includes two phases: (1) the vectorization and projection phase and (2) the classification phase. Each of the steps is described in detail in Sections 6.3.1 and 6.3.4.

6.3.1 Phase 1: Vectorization and projection

In the training phase, each input sample is a pair of a protein sequence and its corresponding label: <protein sequence, label>. In the vectorization component, the model extracts efficient encodings separately for both the protein sequence and the label. Three language models of Prot-BERT-BFD [EHD⁺22], SciBERT [BLC19], and ChemBERTa-2 [ASC⁺22] are utilized to vectorize protein sequences and their corresponding labels. This process is demonstrated in Figure 40, vectorization component.

Each input sample consists of a protein sequence represented as a chain of amino acids. To convert this sequence into an encoded vector representation, we use protein language models (PLMs). Section 2.3 introduces PLMs as NLP-based models that are pre-trained on protein sequences and have been successfully applied to various tasks. Several protein language models can be utilized to vectorize protein sequences (as discussed in Chapter 2). In this study, we use the Prot-BERT-BFD language model [EHD⁺22] to generate

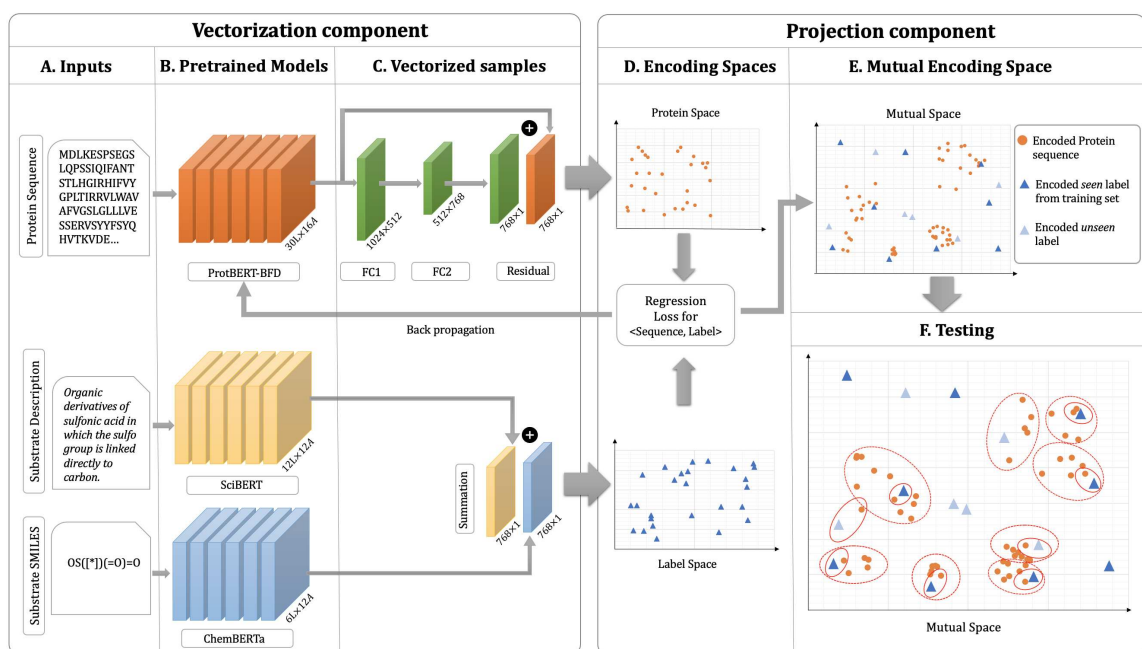
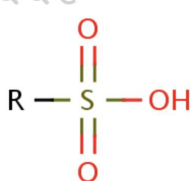


Figure 40: TooTranslator model with two components: vectorization and projection. The figure presents the training and testing process for the TooTranslator model. In the training phase, the samples are a pair of $\langle \text{protein sequence, label} \rangle$. Each label includes a substrate definition and a SMILES string (A). Protein sequences are encoded using ProtBERT-BFD PLM. The label definitions and SMILES strings are encoded using SciBERT and ChemBERTa-2 respectively (B). L and A show the number of layers and the number of attention heads in each Language model and the size of the layers and output vectors is presented under each. Two fully connected layers and one residual layer are applied to the protein vectors resulting in a vector with the same size as labels. The label definition and SMILES vectors are summed element-wise into one vector representing the label (C). Vectorizing all of the samples in the training set results in two encoding spaces: protein space and label space (D). The protein space is projected into the label space calculating a regression loss for sequence, and label vector pairs; back-propagating the loss into the ProtBERT-BFD PLM. The mutual space includes both protein and label vectors (E). This mutual space also includes unseen labels from ChEBI (light blue triangles) vectorized without any proteins associated with them. In the testing phase, each test sample is a protein sequence in the mutual space. The nearest label vector to the protein in this space is selected as the prediction (F).

input protein sequence vectors, as preliminary studies indicate that it surpasses other vectorization methods in terms of both efficiency and accessibility, making it the optimal choice for this task. The output vectors are configured to a size of 1024 by preprocessing the sequences with padding and truncation to this length. The CLS token from the output of Prot-BERT-BFD is extracted to form the final representation. Vectorizing protein sequences of all samples in the training set results in a 1024-dimensional protein space.

CHEBI:33551 - organosulfonic acid

Main | ChEBI Ontology | Automatic Xrefs | Reactions | Pathways | Models



ChEBI Name	organosulfonic acid
ChEBI ID	CHEBI:33551
Definition	An organic derivative of sulfonic acid in which the sulfo group is linked directly to carbon.
Stars	☆☆☆ This entity has been manually annotated by the ChEBI Team.
Download	Molfile XML SDF

- Find compounds which contain this structure
- Find compounds which resemble this structure
- Take structure to the Advanced Search

Formula	HO3SR
Net Charge	0
Average Mass (excl. R groups)	81.07100
Monoisotopic Mass (excl. R groups)	80.96464
SMILES	<chem>OS([*])(=O)=O</chem>

Figure 41: An example ChEBI entry with definition and SMILES string. The figure shows the ChEBI term CHEBI:33551 or class of organosulfonic acid in the UniProt-SPEC dataset. The Definition and SMILES string for this term are identified with green boxes.

Next, the labels of the input entries must be encoded. Each class in UniProt-SPEC-100 represents a chemical compound from the ChEBI ontology (details in Section 3.4). As presented in Figure 41, each ChEBI entry includes a definition of the chemical compound, written in a few sentences, and a SMILES string that describes the compound’s chemical structure. Both of these pieces of information contribute to constructing the label vectors, which are detailed below, along with the corresponding language models used for encoding them. An element-wise summation of these two vectors is the label encoding vector.

The ChEBI ontology provides a “Definition” field for each entry, explaining the meaning of class names in human-readable language. For example, the definition of CHEBI:33551,

organosulfonic acids is “*Organic derivatives of sulfonic acid in which the sulfo group is linked directly to carbon*” (Figure 41). This textual description is vectorized using the SciBERT language model [BLC19], a variant of BERT pre-trained on 1.14 million academic papers from the corpus of `semanticscholar.org`

Additionally, the ChEBI ontology includes the chemical formula for each compound in SMILES (Simplified Molecular Input Line Entry System) format [Wei88]. SMILES is a standardized method for representing chemical structures as ASCII strings, recognized by organizations such as the IUPAC (International Union of Pure and Applied Chemistry). For example, the SMILES string for CHEBI:33551, **organosulfonic acids** is `OS([*])(=O)=O`. These SMILES strings are encoded using ChemBERTa-2 [ASC⁺22], a BERT-based model designed for molecular representation, pre-trained on the PubChem dataset, which contains over 111 million chemical compounds in SMILES format.

For each label, two encoding vectors of size 768 are generated through padding and truncation: one representing the definition (output from SciBERT) and the other corresponding to the SMILES string (output from ChemBERTa-2). The CLS token from each output is extracted, and these two vectors are then combined through element-wise summation to create a single label vector. Consequently, encoding all labels from the training set results in a representation of the seen classes in a 768-dimensional space.

Extracting encodings for both protein sequences and substrate labels yields two vector spaces: the protein space in 1024 dimensions and the label space in 768 dimensions. During training, the protein space is projected into the label space using two fully connected layers and one residual layer (see Section 6.3.2). At the end of this process, a regression loss function is applied to bring the protein vectors closer to their corresponding label vectors (see Section 6.3.3). This projection results in a mutual space with both protein and label representation vectors (Figure 40, projection component).

6.3.2 Network structure

As presented in Figure 40, the TooTranslator model employs three language models as encoders to vectorize protein sequences and their corresponding substrates. These models include ProtBERT-BFD with 30 layers and 16 attention heads for protein encoding, SciBERT with 12 layers and 12 attention heads for encoding substrate descriptions, and ChemBERTa with 6 layers and 12 attention heads for encoding the substrate’s SMILES strings (details in Section 2.3) [EHD⁺22, BLC19, ASC⁺22].

The projection process involves mapping the protein-encoding space to the label-encoding space, to minimize the distance between the protein-encoding vector and its corresponding substrate-encoding vector. To identify the best architecture for this mapping task, various combinations of fully connected layers, residual networks, and multi-head attention layers are explored. Ultimately, the optimal network architecture consists of two fully connected layers followed by a residual layer [HZRS16] (Figure 40). The first fully connected layer (FC1) maps the 1024-dimensional output vectors from ProtBERT-BFD to a 512-dimensional space. The second fully connected layer (FC2) maps this vector from 512 to 768 dimensions. In the residual layer, the initial output of the ProtBERT-BFD model (truncated to 768 dimensions) is added to the final output of the second fully connected layer (FC2).

6.3.3 Training

Training in TooTranslator occurs during the projection process, where the protein space is mapped to the label space. In this phase, for each input sequence, the protein-encoding vector is extracted from the ProtBERT-BFD language model, followed by two fully connected layers and a residual layer (Figure 40). Simultaneously, the corresponding label description is encoded by SciBERT, and the SMILES string is encoded by ChemBERTa. The final label encoding vector is the summation of the two vectors.

The goal of the training process is to minimize the loss between the protein-encoding vector and the label-encoding vector (Figure 40D). The loss is calculated and back-propagated through the ProtBERT-BFD network. Updating the weights of the ProtBERT-BFD model improves the protein encodings, bringing the protein vectors in the substrate space closer to their corresponding substrate labels (Figure 40E). To identify the optimal loss function for this process, four separate models are implemented, each with a different loss function: mean squared error (MSE loss), weighted mean squared error (WMSE loss), mean absolute error (MAE loss), and Huber loss (defined in Section 2.12. [The15, WM05, Hub92]).

The training of the TooTranslator model is implemented using samples from the training set of the UniProt-SPEC-100 dataset as described in Section 3.5.2. This dataset includes 4,455 samples from 96 specific substrate classes. During the test process, the number of classes could be increased to 113 specific substrates including unseen classes of UniProt-SPEC-100 from TS2 as described in Section 6.3.4.

The models are developed in `Python`, version 3.9.1, using `PyTorch` and `Transformers`

packages [PGM⁺19, WDS⁺20]. To compare the loss function effects on the results the training parameters are set as follows for all four models: 100 epochs of training with batch size 1, AdamW optimizer, and learning rate equal to 5e-6 [KB15]. To find the best hyperparameters for the TooTranslator model trained with Huber loss, the following search space is explored: number of epochs from 1 to 200, batch size: 1, 2, 10, and 40, learning rate between 1e-6 to 1e-2. All models are trained and tested on Concordia University computing cluster High-Performance Computing (HPC) Facility (Speed) using one NVIDIA Tesla P6 GPU, with 16 GB of memory.

6.3.4 Phase 2: Classification

In the classification phase, a protein sequence is presented to the TooTranslator as a test sample. The ProtBERT-BFD model encodes this sample into a vector, which is then passed through a network structure with fully connected and residual layers, resulting in a mutual space with 768 dimensions. In this space, the pairwise distance between the protein vector and existing label vectors is calculated, and the nearest label vector to the protein vector is selected as the prediction for the test sample.

This classification method differs slightly from K-Nearest Neighbors (KNN), where the labels of the K nearest protein samples are retrieved as the prediction for the test sample. Using this algorithm, the model can include label vectors from unseen classes, which may represent any substrates characterized by ChEBI ontology. The unseen classes are vectorized by SciBERT and ChemBERTa based on their definition and SMILES strings and could be included in the predictions of *de novo* proteins. Additionally, the model retrieves a list of the 10 nearest labels as top substrate candidates for the test sample, serving as actual suggested substrates for characterization in a wet lab.

This model is tested on two separate test sets. The first test set is the hold-out set derived from the original train/test split of the dataset, with a proportion of 80:20 here referred to as TS1. This test set includes substrates from the same classes as the training set (seen classes). The goal of the second test set is to evaluate the model’s performance on unseen classes. This test set, referred to as TS2 is obtained from the UniProt-SPEC dataset series and includes classes that were excluded from the original datasets due to not meeting the minimum sample size threshold (sample size ≥ 3). As noted in Appendix B Table B, this test set contains 70 samples from 52 classes.

6.3.5 TooTranslator: Validity

In order to investigate the validity of the core idea for the TooTranslator, the following experiment has been implemented on the UniProt-ICAT-100 dataset as a smaller prototype of the implementation (details in Section 3.5). First, in the vectorization phase, the protein sequences have been vectorized by the Prot-BERT-BFD. The label definitions and SMILES strings vectorized by SciBERT and ChemBERTa-2 respectively (see Table 34). To explore label vector representation options, four different label vectors are investigated: definition only, SMILES only, summation of definition and SMILES vectors, and concatenation of the definition and SMILES vectors. Each of the label vectors is utilized to train and test four separate prototypes, referred to as prototype-1 to prototype-4 respectively. The model is trained on the UniProt-ICAT training set and tested by finding the nearest (label) neighbor for test sequences in the hold-out test set of UniProt-ICAT-100.

Definition 6.3.3. A *t-SNE plot* is a visualization created using the *t-Distributed Stochastic Neighbor Embedding (t-SNE)* algorithm to project high-dimensional data into a lower-dimensional space while preserving the local structure of the data [vdMH08].

Figure 42 shows the t-SNE analysis of the mutual space for both protein sequences and labels in the TooTranslator prototype. The figure includes the mutual spaces for each prototype trained with different label vectors. The figure depicts the mutual spaces for each prototype, trained with different label vectors. The consistent patterns across all subplots highlight two key aspects of the model’s validity. First, the t-SNE visualizations for all prototype models exhibit well-defined clusters for substrate classes. For example, `proton`, `potassium(1+)`, and `calcium(2+)` classes form distinct, well-separated clusters.

Second, the plot demonstrates the alignment between the protein vectors and their corresponding label vectors. As shown in Figure 42, protein vectors tend to cluster around their corresponding label vectors (diamonds) in most classes. For example, in the `chloride` class, the protein samples, represented as green circles, cluster around the green diamond, which represents the encoded `chloride` label vector in all four prototype models.

Although all four prototype models show reliable representations in the mutual space, comparing them reveals valuable information. Figure 42-(A) presents the t-SNE visualization of the TooTranslator prototype-1 trained and tested using label definition only. As the figure presents the representations of the label vectors for the classes `iron(2+)`,

copper(1+), and zinc(2+) are encoded close to each other due to similar ChEBI definitions (Table 34), also similar characteristics of these three metals according to SciBERT. On the other hand, subplot (B) which presents the prototype-2 mutual space, trained and tested using SMILES strings as label vectors, differentiates these three vectors as they have different SMILES strings (see Table 34).

Table 34: CHEBI terms and definitions for UniProt-ICAT-100 classes.

Class	CHEBI term	Label Name	Definition	SMILES string
0	CHEBI:24636	proton	Nuclear particle of charge number +1, spin 1/2 and rest mass of 1.007276470(12) u.	[1H+]
1	CHEBI:29108	calcium(2+)	A calcium cation that has formula Ca.	[Ca++]
2	CHEBI:29103	potassium(1+)	A monoatomic monocation obtained from potassium.	[K+]
3	CHEBI:17996	chloride	A halide anion formed when chlorine picks up an electron to form an anion.	[Cl-]
4	CHEBI:29101	sodium(1+)	A monoatomic monocation obtained from sodium.	[Na+]
5	CHEBI:16189	sulfate	A sulfur oxoanion obtained by deprotonation of both OH groups of sulfuric acid.	[O-]S([O-])(=O)=O
6	CHEBI:29105	zinc(2+)	A zinc cation that has formula Zn.	[Zn++]
7	CHEBI:28938	ammonium	An onium cation obtained by protonation of ammonia.	[H][N+](H)(H)[H]
8	CHEBI:17632	nitrate	A nitrogen oxoanion formed by loss of a proton from nitric acid. Principal species present at pH 7.3.	[O-][N+](O)=O
9	CHEBI:29033	iron(2+)	An iron cation that has formula Fe.	[Fe++]
10	CHEBI:35780	phosphate ion	A phosphorus oxoanion that is the conjugate base of phosphoric acid.	[O-]P(=O)([O-])[O-]

The table shows class number, ChEBI term, substrate name, definition in ChEBI ontology, and its SMILES string for substrate classes in UniProt-ICAT-100.

6.4 Results

As discussed in Section 6.3.4, the TooTranslator model is tested using two test sets: the UniProt-SPEC-100 hold-out test set or TS1 and the TS2 dataset, which includes classes with fewer than three samples (Section 6.3.4). The first test set consists of samples from the same classes that are used during training, referred to as seen classes in the context of open-world classification (Section 2.4). The classes of the second dataset, on the other hand, have not been included in the training process (unseen classes) In contrast, the TS2 dataset contains unseen classes, which are not included in the training process, allowing for the evaluation of the model’s ability to detect these novel classes.

Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported for this task (definitions in Section 2.11). F1-score and MCC are the main metrics of the evaluation but precision and recall can be important in different scenarios. For example, when identifying a specific substrate transporter in an organism’s transportome, recall is critical to avoid missing any transporters (False Negatives). On the other hand, if identifying a substrate for a newly sequenced protein, high precision is necessary to reduce costs in later wet lab experiments.

6.4.1 Results for Group A: seen classes

This section presents the results of the TooTranslator model on the samples from the test set of the UniProt-SPEC-100 dataset. Since this dataset is split to train and test using the stratified strategy for random sampling, all of the classes are present in both training and the test set, meaning that in this experiment, the model classifies the test samples of the seen classes.

As described in Section 6.3.3, the TooTranslator is implemented using four different loss functions separately. The results for Huber loss, MSE loss, wMSE loss, and MAE loss functions are presented in Tables 37, 67, 65, and 66. Table 35 compares the overall results for these four models in F1-score and MCC metrics. According to Table 35 and Figure 43, the TooTranslator model trained with the MAE loss function presents numerically largest value for MCC metric with a competitive F1-score (second largest number). However, taking the number of unpredicted classes into account, the wMSE and Huber loss function cover 4 more minority classes in their results.

Table 36 presents the 3-fold cross-validation results for each model on the UniProt-SPEC-100 dataset. The one-way ANOVA test is performed on each performance metric. The p-values indicate that there is no statistically significant difference in model performance, as all values exceed 0.05, failing to reject the null hypothesis. However, the TooTranslator trained with the Huber loss function is selected as a compromised model for a lower number of unpredicted classes and high-performance results (Table 35).

Table 37: TooTranslator Huber loss function results for the hold-out test set.

Substrate	Train	Test	TP	FP	FN	TN	Acc	Prec	Rec	F1	MCC
proton	817	409	393	14	16	1062	0.980	0.966	0.961	0.963	0.949
calcium(2+)	577	288	283	14	5	1183	0.987	0.953	0.983	0.968	0.960
potassium(1+)	484	242	232	4	10	1239	0.991	0.983	0.959	0.971	0.965
chloride	295	147	138	2	9	1336	0.993	0.986	0.939	0.962	0.958
sodium(1+)	177	89	79	1	10	1395	0.99	0.99	0.89	0.94	0.93
zinc(2+)	54	26	26	0	0	1459	1.00	1.00	1.00	1.00	1.00

Continued on next page

Table 37: TooTranslator Huber loss function results for the hold-out test set.

Substrate	Train	Test	TP	FP	FN	TN	Acc	Prec	Rec	F1	MCC
sulfate	45	22	20	1	2	1462	1.00	0.95	0.91	0.93	0.93
ammonium	34	18	18	0	0	1467	1.00	1.00	1.00	1.00	1.00
water	28	15	15	0	0	1470	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	25	12	10	0	2	1473	1.00	1.00	0.83	0.91	0.91
protein	19	9	8	3	1	1473	1.0	0.7	0.9	0.8	0.8
glutathione derivative	18	9	9	2	0	1474	1.0	0.8	1.0	0.9	0.9
retinol	18	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
nitrate	17	9	8	2	1	1474	1.0	0.8	0.9	0.8	0.8
iron(2+)	16	8	8	5	0	1472	1.0	0.6	1.0	0.8	0.8
pyruvate	13	7	6	0	1	1478	1.0	1.0	0.9	0.9	0.9
ATP(4-)	12	6	5	0	1	1479	1.0	1.0	0.8	0.9	0.9
copper(1+)	12	5	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
urea	11	5	5	3	0	1477	1.0	0.6	1.0	0.8	0.8
dehydroascorbide(1-)	10	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
choline	9	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
L-ornithinium(1+)	8	4	3	0	1	1481	1.0	1.0	0.8	0.9	0.9
cholesterol	8	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
bile acid	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
L-glutamate(1-)	7	3	3	3	0	1479	1.0	0.5	1.0	0.7	0.7
L-cystine zwitterion	7	4	1	1	3	1480	1.0	0.5	0.2	0.3	0.4
lipid-linked peptidoglycan	7	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
3'-phosphonato-5'-adenylyl sulfate(4-)	6	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
copper(2+)	6	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
UDP-N-acetyl-alpha-D-glucosamine(2-)	6	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
S-adenosyl-L-methionine zwitterion	6	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
citrate(3-)	6	4	2	1	2	1480	1.0	0.7	0.5	0.6	0.6
heme	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
oleate	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
cobalamin	6	3	1	2	2	1480	1.0	0.3	0.3	0.3	0.3
UDP-D-glucose(2-)	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
urate salt	6	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
hydrogencarbonate	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
sucrose	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphoglycerate	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
biotinate	5	3	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
1-4-butanediammonium	5	3	0	1	3	1481	1.0	0.0	0.0	0.0	-0.0
glutathionate(1-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
2-oxoglutarate(2-)	4	3	2	1	1	1481	1.0	0.7	0.7	0.7	0.7
maltose	4	3	0	1	3	1481	1.0	0.0	0.0	0.0	-0.0
iron(3+)	4	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
acetyl-CoA(4-)	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
pyrimidine nucleotide	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
cadmium cation	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
NAD	4	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
L-argininium(1+)	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
L-proline zwitterion	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
gamma-aminobutyric acid zwitterion	4	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
GDP-fucose	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-ascorbate	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-arabinose	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
mercury cation	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
N-ribosylnicotinamide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphate ion	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
D-methionine zwitterion	3	1	1	3	0	1481	1.0	0.2	1.0	0.4	0.5
boric acid	3	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
maltodextrin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
N-(4-aminobenzoyl)-L-glutamate	2	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0

Continued on next page

Table 37: TooTranslator Huber loss function results for the hold-out test set.

Substrate	Train	Test	TP	FP	FN	TN	Acc	Prec	Rec	F1	MCC
guanine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
xanthine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
uridine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	2	1	1482	1.0	0.0	0.0	0.0	-0.0
long-chain fatty acid	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-histidine zwitterion	2	2	0	2	2	1481	1.0	0.0	0.0	0.0	-0.0
glycine zwitterion	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-methionine zwitterion	2	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
bacteriocin	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
prostaglandin	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	0	1	1484	1.0	Nan	0.0	0.0	Nan
molybdate	2	1	0	0	1	1484	1.0	Nan	0.0	0.0	Nan
N-retinylidene	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phatidylethanolamine											
lipopolysaccharide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gluconate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
tungstate	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
sn-glycerol 3-phosphate(2-)	2	1	0	3	1	1481	1.0	0.0	0.0	0.0	-0.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuraminate(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	0	1	1484	1.0	Nan	0.0	0.0	Nan
Overall	2970	1485	1392	-	-	-	0.999	0.937	0.832	0.790	0.926

The table presents the results of the hold-out test set of UniProt-SPEC-100 using TooTranslator trained using the Huber loss function, predicted by a K-NN classifier (K=1). The overall number for TP presents the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric.

6.4.2 Results for Group B: unseen classes

This section presents the results of the TooTranslator model on the TS2 test set. The labels of this data set are not included in the classes of the training set and the model is evaluated on the ability of detecting samples from unseen classes. Table 38 presents the TooTranslator predictions for each test sample in *TS2* along with the sample’s correct label. This table also presents the distances of the protein sequence from the true labels and the prediction. According to this table, none of the samples of unseen classes are correctly predicted to their correct label. However, as Table 70 shows for 9 test samples the correct label is in the 10 nearest neighbors of the test sample.

Table 35: Results of TooTranslator trained with four loss functions on TS1.

Model name	Loss function	F1-score	MCC	# Unpredicted
TooTranslator-MSE	Mean Square Error	0.740	0.908	7
TooTranslator-wMSE	Weighted Mean Square Error	0.748	0.918	3
TooTranslator-MAE	Mean Absolute Error	0.762	0.928	7
TooTranslator-Huber	Huber	0.790	0.926	3

The TooTranslator models share the same structure (Figure 40). All training parameters are held consistent, with the exception of the loss function, to facilitate a comparative analysis of the results. The evaluation metrics presented are F1-score, MCC accompanied by the number of unpredicated classes (zero True or False positives) on the last column. Bold shows the numerically largest value.

Table 36: 3-Fold Cross-validation results for TooTranslator with four loss functions.

Model name	Accuracy	Precision	Recall	F1-Score	MCC
TooTranslator-Huber	0.998 ± 0.000	0.921 ± 0.034	0.699 ± 0.092	0.627 ± 0.092	0.907 ± 0.040
TooTranslator-wMSE	0.997 ± 0.001	0.861 ± 0.034	0.629 ± 0.018	0.539 ± 0.014	0.837 ± 0.039
TooTranslator-MSE	0.998 ± 0.001	0.920 ± 0.035	0.700 ± 0.080	0.627 ± 0.080	0.906 ± 0.041
TooTranslator-MAE	0.997 ± 0.000	0.881 ± 0.023	0.591 ± 0.039	0.526 ± 0.035	0.860 ± 0.027
p-value	2.503e-01	2.370e-01	3.093e-01	2.368e-01	2.893e-01

The average values and the standard deviation for results of 3 fold of UniProt-SPEC-100 are presented for five evaluation metrics. The last row presents the p-value calculated using the one-way ANOVA algorithm to assess the statistical significance of the metric values over 3-fold results.

Table 38: Results of the TooTranslator with Huber loss for TS2 dataset.

Sample #	Predicted	Dist. to Predic- tion	Actual	Dist. to Actual	Match
0	L-serine zwitterion	0.35	L-aspartate(1-)	0.48	False
1	silver cation	0.46	poly-beta-1-6-N-acetyl-D-glucosamine	0.66	False
2	L-serine zwitterion	0.36	nicotinate	0.56	False
3	sulfite	0.40	nitrite	0.41	False
4	L-serine zwitterion	0.28	succinate(2-)	0.30	False
5	O-acylcarnitine	0.34	spermine(4+)	0.68	False
6	O-acylcarnitine	0.33	spermine(4+)	0.70	False
7	L-tyrosine zwitterion	0.32	L-glutamine zwitterion	0.34	False
8	L-serine zwitterion	0.26	L-threonine zwitterion	0.43	False
9	O-acylcarnitine	0.37	fluconazole	1.09	False
10	L-glutamine zwitterion	0.28	agmatinium(2+)	0.70	False
11	oxalate(2-)	0.34	creatine zwitterion	0.49	False
12	oxalate(2-)	0.33	creatine zwitterion	0.49	False

Continued on next page

Table 38: Results of the TooTranslator with Huber loss for TS2 dataset.

Sample #	Predicted	Dist. to Prediction	Actual	Dist. to Actual	Match
13	acadesine	0.42	acetate	1.10	False
14	oxalate(2-)	0.29	mevalonate	0.92	False
15	ferrienterobactin(3-)	0.36	ferrienterobactin(3-)	0.36	True
16	L-aspartate(1-)	0.19	hexuronic acid	0.95	False
17	L-tyrosine zwitterion	0.31	S-methyl-L-methionine zwitterion	0.56	False
18	L-serine zwitterion	0.33	L-leucine zwitterion	0.51	False
19	silicic acid	0.29	arsenite(1-)	0.45	False
20	coenzyme A(4-)	0.26	GDP-mannose	0.53	False
21	O-acylcarnitine	0.42	acetate	1.06	False
22	O-acylcarnitine	0.50	manganese cation	1.07	False
23	L-glutamine zwitterion	0.24	O-acylcarnitine	0.59	False
24	coenzyme A(4-)	0.29	coenzyme A(4-)	0.29	True
25	benzoate	0.44	enterobactin(1-)	1.79	False
26	L-aspartate(1-)	0.22	(R)-pantothenate	0.71	False
27	sulfite	0.52	glycerol	0.74	False
28	cytidine	0.30	uracil	1.23	False
29	benzoate	0.46	glycerol 1-phosphodiester(1-)	0.72	False
30	benzoate	0.46	glycerol 1-phosphodiester(1-)	0.72	False
31	O-acylcarnitine	0.38	fluconazole	1.12	False
32	microcin	0.34	bilirubin(2-)	0.55	False
33	tetracycline zwitterion	0.37	nickel cation	1.89	False
34	O-acylcarnitine	0.37	methyl beta-D-galactoside	0.53	False
35	O-acylcarnitine	0.40	cytidine	0.43	False
36	cytidine	0.34	cytosine	0.99	False
37	nitrite	0.33	2-cis-abscisate	1.54	False
38	L-tyrosine zwitterion	0.32	L-glutamine zwitterion	0.34	False
39	nitrite	0.36	silver cation	0.47	False
40	benzoate	0.46	sulfite	0.52	False
41	L-aspartate(1-)	0.19	allantoate	0.52	False
42	acadesine	0.49	oxalate(2-)	1.33	False
43	O-acylcarnitine	0.36	nicotinate	0.60	False
44	O-acylcarnitine	0.50	microcin	0.53	False
45	silver cation	0.46	L-serine zwitterion	0.67	False
46	sulfite	0.52	silicic acid	0.62	False
47	mevalonate	0.13	D-galactofuranose	1.05	False
48	coenzyme A(4-)	0.34	thiamine(1+) diphosphate(3-)	1.10	False
49	oxalate(2-)	0.42	organic phosphonate	0.61	False
50	L-serine zwitterion	0.39	benzoate	0.54	False
51	cytidine	0.30	acadesine	0.41	False
52	mevalonate	0.13	methyl beta-D-galactoside	1.19	False
53	oxalate(2-)	0.38	tetracycline zwitterion	1.55	False
54	coenzyme A(4-)	0.35	thiamine(1+) diphosphate(3-)	1.12	False
55	ferrienterobactin(3-)	0.50	nickel cation	1.50	False
56	nitrite	0.38	galactitol	1.00	False
57	L-serine zwitterion	0.31	arsenite(1-)	0.52	False
58	O-acylcarnitine	0.38	uracil	0.90	False

Continued on next page

Table 38: Results of the TooTranslator with Huber loss for TS2 dataset.

Sample #	Predicted	Dist. to Prediction	Actual	Dist. to Actual	Match
59	oxalate(2-)	0.31	L-tyrosine zwitterion	0.50	False
60	mevalonate	0.13	D-ribose	0.18	False
61	L-serine zwitterion	0.32	L-serine zwitterion	0.32	True
62	microcin	0.35	bilirubin(2-)	0.65	False
63	L-glutamine zwitterion	0.42	L-tyrosine zwitterion	0.54	False
64	silver cation	0.46	manganese cation	0.56	False
65	tetracycline zwitterion	0.41	phytochelatin	1.77	False
66	L-tyrosine zwitterion	0.31	L-phenylalanine zwitterion	0.40	False
67	allantoate	0.60	adenine	0.72	False
68	microcin	0.35	(+)-abscisic acid D-glucopyranosyl ester	0.51	False
69	L-aspartate(1-)	0.30	allantoate	0.58	False

The table presents the TooTranslator prediction for each sample in test set TS2 in columns Predicted. The column Dist. to Prediction presents the distance of the test sample from the predicted label in the mutual space. The column Actual is the true label of the test sample and the column Dist. to Actual shows the distance of the test sample from its true label. The last column indicates if the prediction matches the actual label.

6.5 Discussion

This section expands on three discussion points according to the results of the TooTranslator model on group (A) seen classes and group (B) unseen classes. Section 6.5.3 clarifies the change of the task from classification to a regression task and the impacts of this change on the model performance. Section 6.5.1 discusses the distances of samples in seen and unseen classes to their correct label. Section 6.5.2 discusses the effect of the dataset size and the composition’s effect on seen and unseen class prediction.

6.5.1 Training process positively affects unseen classes distances

The training process of TooTranslator aims to map protein sequence vectors to positions near their corresponding label vectors. Figure 44 displays a 2D t-SNE representation of protein sequences and label vectors in TS1 without any training (Frozen weights in Figure 44a) and after training with Huber loss function (Figure 44b). This figure highlights two key observations post-training. Firstly, as shown in Figure 44b protein sequence vectors (gray circles) are clustered into multiple groups, ideally corresponding to substrate classes, in contrast to their random distribution in 2D space before training 44a. Secondly,

the distribution of protein representations within the substrate space forms clusters around substrate vectors, whereas before training, protein and substrate representations were completely separated (Figure 44a). Additionally, some clusters of proteins are observed with a single substrate in their center, representing the class label.

Figure 44 illustrates the representations of test samples from the seen classes during the training process (TS1). The objective of the study is to evaluate whether the training process updates the model weights to produce reliable representations of protein sequences from unseen classes (TS2), positioning them close to their corresponding label vectors.

Table 39 compares the distance between protein vectors from unseen classes (TS2) and their corresponding label vectors before and after training using both UniProt-SPEC-100 and UniProt-ICAT-100 datasets. Figure 45 presents the distance distribution of samples in both TS1 and TS2 before and after training using the UniProt-SPEC-100 dataset. They both show the distance of unseen classes from TS2 has decreased even though these classes are not included in the training process. The training of the TooTranslator has modified the representations for unseen class samples closer to their corresponding label even though it does not provide accurate results for classification.

Table 39: Distances from actual label on TS2 with three TooTranslator models.

Sample #	Actual Label	Frozen Distance	Trained with UniProt-SPEC-100	Trained with UniProt-ICAT-100
0	L-aspartate(1-)	1.59	0.70	1.49
1	poly-beta-1-6-N-acetyl-D-glucosamine	1.63	0.70	1.01
2	nicotinate	1.64	0.60	1.00
3	nitrite	1.63	0.45	1.03
4	succinate(2-)	1.65	0.26	1.14
5	spermine(4+)	1.62	0.71	1.56
6	spermine(4+)	1.62	0.71	1.56
7	L-glutamine zwitterion	1.60	0.31	1.56
8	L-threonine zwitterion	1.60	0.42	1.41
9	fluconazole	1.53	1.19	2.28
10	agmatinium(2+)	1.61	0.70	1.49
11	creatine zwitterion	1.59	0.46	1.23
12	creatine zwitterion	1.59	0.46	1.23
13	acetate	1.64	0.64	1.58
14	mevalonate	1.66	0.93	1.01
15	ferrienterobactin(3-)	1.66	0.36	2.09
16	hexuronic acid	1.43	0.91	1.57
17	S-methyl-L-methionine zwitterion	1.65	0.52	1.40
18	L-leucine zwitterion	1.62	0.44	1.28
19	arsenite(1-)	1.64	0.45	1.25
20	GDP-mannose	1.69	0.53	2.30

Continued on next page

Table 39: Distances from actual label on TS2 with three TooTranslator models.

Sample #	Actual Label	Frozen Distance	Dis- UniProt-SPEC-100	Trained with UniProt-ICAT-100
21	acetate	1.64	0.59	1.58
22	manganese cation	1.65	0.88	1.01
23	O-acylcarnitine	1.35	0.61	1.36
24	coenzyme A(4-)	1.72	0.31	2.27
25	enterobactin(1-)	1.71	1.75	2.43
26	(R)-pantothenate	1.65	0.79	1.86
27	glycerol	1.60	0.74	1.49
28	uracil	1.56	0.97	1.39
29	glycerol 1-phosphodiester(1-)	1.61	0.70	1.35
30	glycerol 1-phosphodiester(1-)	1.61	0.70	1.35
31	fluconazole	1.53	1.28	2.28
32	bilirubin(2-)	1.75	0.64	2.37
33	nickel cation	1.66	0.91	1.01
34	methyl beta-D-galactoside	1.69	0.76	1.69
35	cytidine	1.68	0.83	1.99
36	cytosine	1.62	0.90	1.34
37	2-cis-abscisate	1.70	1.52	2.23
38	L-glutamine zwitterion	1.60	0.33	1.56
39	silver cation	1.65	0.52	0.68
40	sulfite	1.65	0.48	1.05
41	allantoate	1.59	0.58	1.49
42	oxalate(2-)	1.65	0.45	1.12
43	nicotinate	1.64	0.64	1.00
44	microcin	1.56	0.33	2.23
45	L-serine zwitterion	1.57	0.54	1.17
46	silicic acid	1.59	0.62	1.39
47	D-galactofuranose	1.69	1.05	1.60
48	thiamine(1+) diphosphate(3-)	1.74	1.12	1.89
49	organic phosphonate	1.19	0.70	1.21
50	benzoate	1.63	0.65	1.20
51	acadesine	1.71	0.47	1.88
52	methyl beta-D-galactoside	1.69	1.19	1.63
53	tetracycline zwitterion	1.64	1.70	2.29
54	thiamine(1+) diphosphate(3-)	1.74	1.13	1.89
55	nickel cation	1.66	0.92	1.01
56	galactitol	1.64	1.05	1.66
57	arsenite(1-)	1.64	0.52	1.32
58	uracil	1.56	1.00	1.47
59	L-tyrosine zwitterion	1.55	0.41	1.72
60	D-ribose	1.66	0.18	1.0
61	L-serine zwitterion	1.57	0.27	1.18
62	bilirubin(2-)	1.75	0.66	2.36
63	L-tyrosine zwitterion	1.55	0.51	1.57
64	manganese cation	1.65	1.00	1.03
65	phytochelatin	1.65	1.72	1.04
66	L-phenylalanine zwitterion	1.60	0.35	1.24

Continued on next page

Table 39: Distances from actual label on TS2 with three TooTranslator models.

Sample #	Actual Label	Frozen Distance	Trained with UniProt-SPEC-100	Trained with UniProt-ICAT-100
67	adenine	1.66	0.73	1.35
68	(+)-abscisic acid D-glucopyranosyl ester	1.73	0.50	2.22
69	allantoate	1.59	0.53	1.50

The table presents the distances of the test samples of unseen classes in TS2 from their actual label in the mutual spaces of three TooTranslator models: frozen weights (before training), trained with UniProt-SPEC-100 and trained with UniProt-ICAT-100 (or prototype-3).

6.5.2 Training dataset composition is critical

One important aspect to explore in understanding the model’s performance on unseen classes is the size and composition of the training dataset. The UniProt-SPEC-100 training dataset includes 96 classes and 2,970 protein sequences. A key question is how representative these classes are of the full substrate space involved in the metabolism of any organism. Additionally, whether including more classes, even with a low number of samples, could improve the model’s performance on unseen classes. For example, does a model trained only on inorganic ion transporters perform worse than a model trained on both inorganic and organic ion transporters when predicting the unseen classes?

To investigate this hypothesis, we compare the performance of the TooTranslator model trained on the UniProt-SPEC-100 and UniProt-ICAT-100 (prototype-3 in Section 6.3.5 datasets. UniProt-ICAT-100 is a subset of UniProt-SPEC-100, containing only inorganic ion transporters with more than 10 samples per class. Comparing these results can help determine whether dataset size and composition are important factors in predicting unseen classes. Table 39 presents the distances of samples in the TS2 dataset from their actual label in the mutual space. Both models are trained and tested with the same hyperparameters with the wMSE loss function.

As the table shows, the distances of unseen class data samples for prototype-3 are significantly higher for all test samples in TS2 compared to the TooTranslator model trained on UniProt-SPEC-100. Although the prototype-3 model shows lower average distances compared to the frozen TooTranslator model, some test samples exhibit higher distances in the trained prototype-3 mutual space. The comparison of results suggests that the composition of the training dataset is a critical factor in this study, aligning with the initial hypothesis.

It is important to note that this experiment investigates only three BERT-based language models (LMs), so the conclusions are specific to these models and may not be generalizable to other LMs as different models may yield different results. We selected three BERT-based models to evaluate whether their embedding spaces align for de novo predictions, ensuring compatibility by choosing models from the same family. However, many other pre-trained language models could be explored for this task. This study makes no assumptions about how other models might perform.

6.5.3 Regression versus Classification

The TooTranslator model approaches substrate specificity prediction by redefining the training process as a regression task rather than a classification task. This section explores this task shift and its impact on the results for both seen and unseen classes. By switching from classification to regression, the model can expand the classification framework to include unseen classes, potentially predicting de novo protein sequences.

As Figure 40 illustrates, the TooTranslator network does not incorporate a classifier layer, and the prediction process is independently executed using a KNN classifier. As detailed in Section 6.3.3, the training procedure for TooTranslator involves vectorizing the sequence and its label, followed by the projection of the protein space to the substrate space. The objective of the training is to minimize the distance between two vectors: the protein-encoding vector and the label encoding vector. This reformulates the training process as a regression problem, which is addressed using regression loss functions such as MSE and Huber loss.

The TooT-SPEC and TooT-Triplet-SPEC (Sections 3.7.3 and 5.7) address substrate specificity prediction as a classification task. Comparative analysis of the TooTranslator model against these previous models demonstrates improved performance, particularly for the minority classes within the long tail of the UniProt-SPEC-100 sample distribution (Table 37). Tables 26, E.2.1, and 37 show that shifting from classification to regression reduces the number of unpredicted minority classes. Specifically, the number of unpredicted classes decreases from 13 in 1TooT-SPEC and 5 in TooT-Triplet (using classification) to just 3 in TooTranslator through regression. According to Table 37, the only unpredicted classes in the hold-out test set of UniProt-SPEC-100 are `folate(-2)`, `N-retinylidene\textasciitilde`, and `L-tryptophan zwitterion`. Each class has only three samples in the dataset, with one sample in each training, validation, and test set.

6.5.4 Unseen class prediction analysis

To evaluate the performance of TooTranslator on unseen-class test samples in TS2, we train the model on the entire UniProt-SPEC-100 dataset—comprising its training, validation, and test subsets—for 35 epochs using the same configurations described in Section 6.3.3. During testing, we vectorize both the UniProt-SPEC-100 and TS2 label sets using TooTranslator’s vectorization component. We then provide these label embeddings, along with the vectorized TS2 test proteins, to the trained K-Nearest Neighbors (KNN) model for prediction. Importantly, while the KNN model is trained only on seen classes, it is configured to identify the nearest label embeddings from the combined set of both seen (UniProt-SPEC-100) and unseen (TS2) classes for each test protein.

Similar to the experiment described in Section 6.3.3, none of the unseen class test samples is correctly predicted as the top (rank-1) candidate. Among the 70 test samples, the top predictions for 64 samples are incorrectly predicted with labels from UniProt-SPEC-100 (seen classes). Table 40 presents the most frequent incorrect predictions made by TooTranslator for samples from TS2. The table shows that the model misclassifies these samples into 31 different labels from UniProt-SPEC-100. Notably, only 2 labels appear as the top candidate predictions from TS2, although they are not the correct labels.

According to these results, unseen labels may form a second cluster separate from the seen labels and proteins’ cluster. This separation prevents them from appearing in the 1-nearest neighbor list of protein samples and, consequently, from being candidate predictions. This separation may result from the training strategy. As shown in Figure 40, the loss computed for protein and label encodings is only back-propagated to the protein sequence encoder, while the label encoder remains fixed with its pre-trained weights. This constraint arises from memory limitations, as retaining all three language models in memory to back-propagate the loss to both the label and protein encoders is infeasible. Allocating additional resources would allow us to investigate this approach further and determine whether updating the label encoder improves its representation of unseen classes.

To investigate the positions of the correct predictions in the ranked candidate substrate list, their rank distributions are examined. Figure 46 shows the number of correct predictions in rank intervals from 1 to 150, with 10-step intervals. As illustrated, the

largest number of correct predictions (12 test samples) occurs within the top 10 candidates. An additional 7 test samples have their correct predictions within ranks 11–20, while 8 test samples have correct predictions in the 21–30 range.

Table 41 shows the distribution for the top 3, 5, 10, 50, and 100 candidates. it indicates the number of correct predictions within the top n candidates. According to these results, 12 test samples from TS2 have their correct label among the top 10 substrate candidates identified by TooTranslator, and this number increases to 56 within the top 100 candidates. We define the hit rate as the fraction of samples for which the correct prediction appears within the specified candidate range. Table 41 shows the hit rate for substrates in the top 100 candidates is 0.800, indicating that 80% of the test samples in the unseen TS2 dataset have the correct prediction within their top 100 candidates.

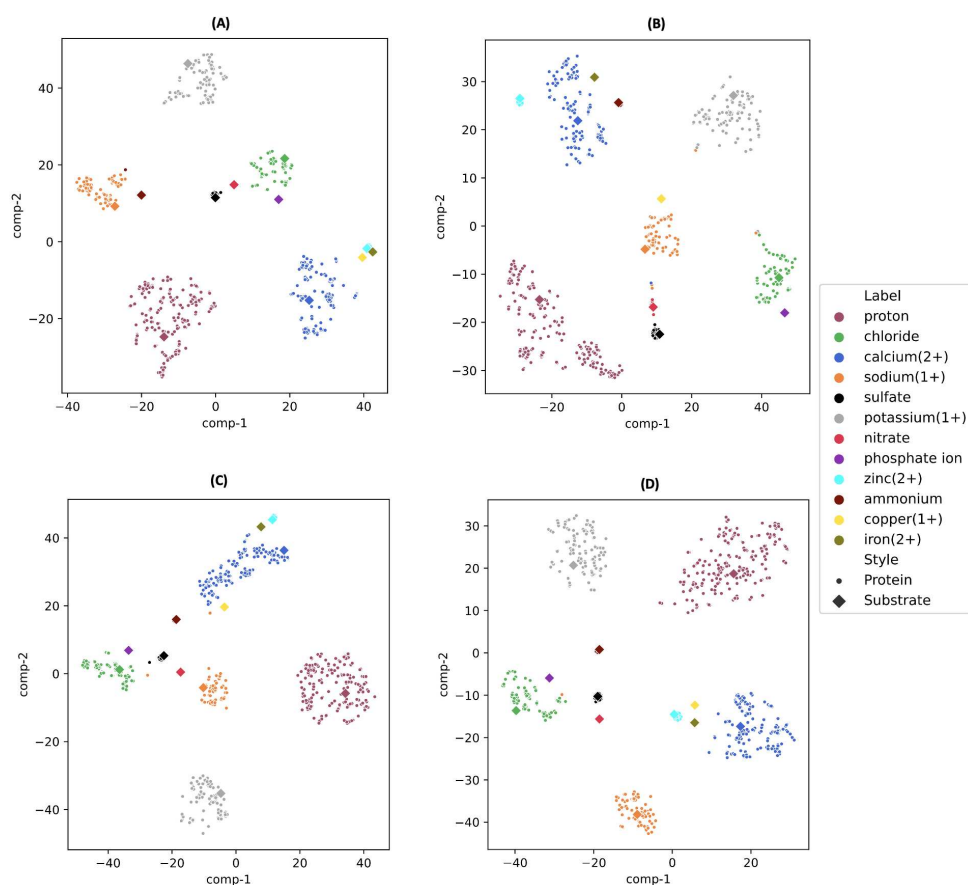


Figure 42: t-SNE analysis of TooTranslator prototype using UniProt-ICAT-100 dataset. The figure shows the 2-d t-SNE analysis of representations for test sequences in UniProt-ICAT-100 generated by TooTranslator trained on the same dataset. The substrate classes are presented in different colors as indicated in the legend (right). The protein sample and label description representations are depicted with circles and diamonds respectively. Four subplots present the representations for the model trained using four different label vectors. (A) presents the sample vectors from TooTranslator trained on the label definition only. (B) presents the sample vectors from TooTranslator trained on the label SMILES strings only. (C) presents the sample vectors from TooTranslator trained on the summation vectors for label definition and SMILES strings. (D) presents the sample vectors from TooTranslator trained on the label vector concatenated from label definition and SMILES strings.

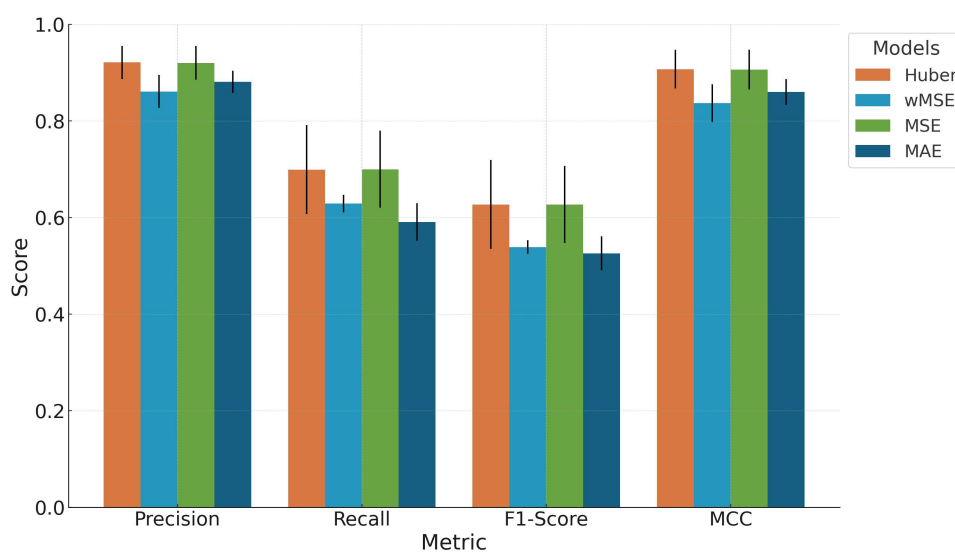
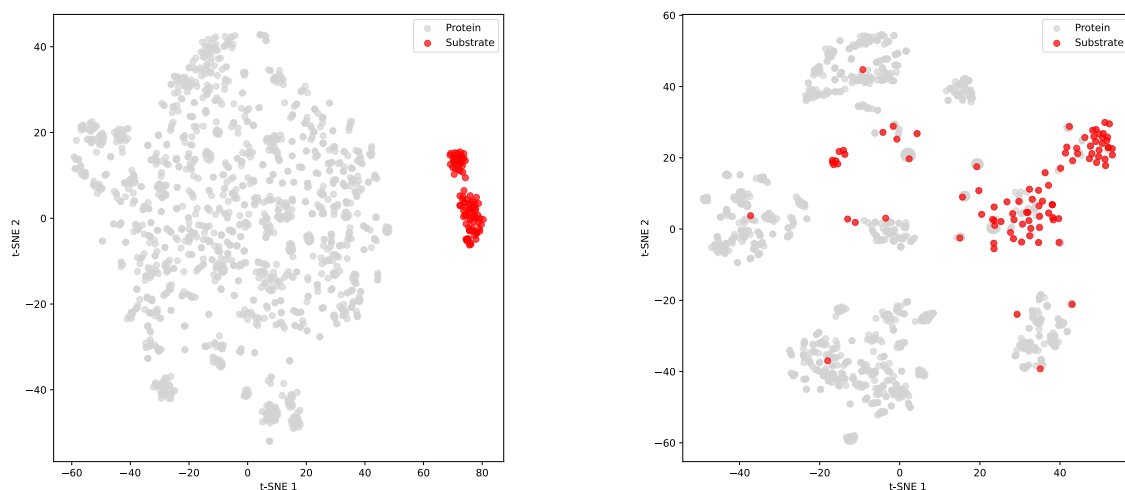


Figure 43: Results of 3-fold cross-validation for TooTranslator using four loss functions.

Results evaluation for four TooTranslator models trained and tested separately with the same training parameters and different loss functions of Huber, weighted Mean Square Error (wMSE), Mean Square Error (MSE), and Mean Absolute Error (MAE).



(a) TooTranslator representations with Frozen weights.

(b) TooTranslator representations trained with Huber Loss.

Figure 44: t-SNE plot for protein and label representations.

The figure presents the 2-D t-SNE plot of protein representations (gray) and label representations (red) before and after training in TooTranslator with Huber loss for samples in UniProt-SPEC-100 test set.

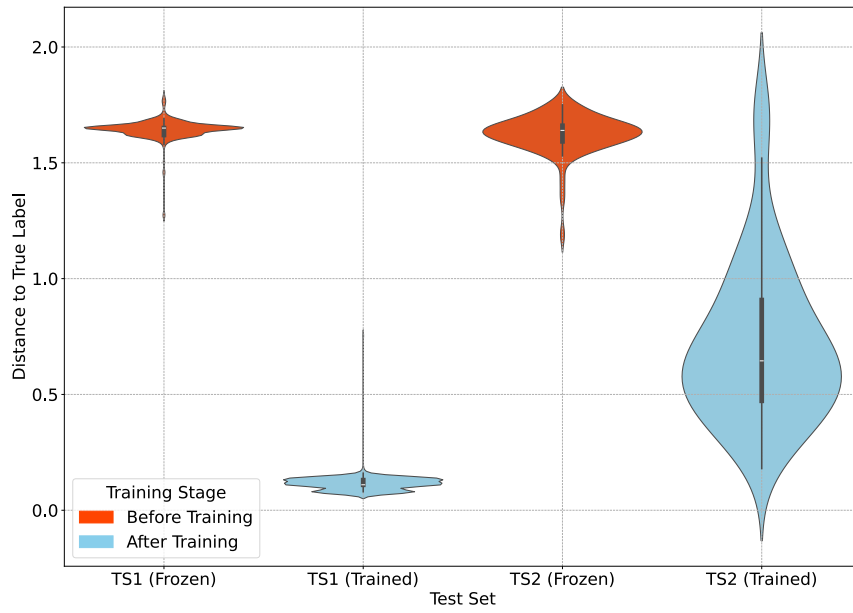


Figure 45: Distribution of distances to the true label for TooTranslator representations before and after training.

TooTranslator is tested with two test sets of TS1 and TS2 (see Section 6.3.4). The figure shows the distribution of each test sample from its true label before (orange plot) and after (blue plot) training for TS1 (left) and TS2 (right).

Table 40: TooTranslator predicted labels for unseen class samples from TS2

#	Name	Frequency	Dataset
0	(R)-carnitine	6	UniProt-SPEC-100
1	L-serine zwitterion	5	TS2
2	D-methionine zwitterion	4	UniProt-SPEC-100
3	pyruvate	4	UniProt-SPEC-100
4	proton	3	UniProt-SPEC-100
5	L-arabinose	3	UniProt-SPEC-100
6	folate(2-)	3	UniProt-SPEC-100
7	L-argininium(1+)	3	UniProt-SPEC-100
8	ATP(4-)	3	UniProt-SPEC-100
9	L-cystine zwitterion	2	UniProt-SPEC-100
10	sn-glycerol 3-phosphate(2-)	2	UniProt-SPEC-100
11	biotinate	2	UniProt-SPEC-100
12	1-4-butanediammonium	2	UniProt-SPEC-100
13	urate salt	2	UniProt-SPEC-100
14	cobalamin	2	UniProt-SPEC-100
15	water	2	UniProt-SPEC-100
16	L-ornithinium(1+)	2	UniProt-SPEC-100
17	tungstate	2	UniProt-SPEC-100
18	L-glutamate(1-)	2	UniProt-SPEC-100
19	gamma-aminobutyric acid zwitterion	2	UniProt-SPEC-100
20	L-proline zwitterion	2	UniProt-SPEC-100
21	nitrate	1	UniProt-SPEC-100
22	copper(1+)	1	UniProt-SPEC-100
23	CMP-N-acetyl-beta-neuraminate(2-)	1	UniProt-SPEC-100
24	calcium(2+)	1	UniProt-SPEC-100
25	L-methionine zwitterion	1	UniProt-SPEC-100
26	urea	1	UniProt-SPEC-100
27	glutathione derivative	1	UniProt-SPEC-100
28	iron(2+)	1	UniProt-SPEC-100
29	acetyl-CoA(4-)	1	UniProt-SPEC-100
30	guanine	1	UniProt-SPEC-100
31	UDP-D-glucose(2-)	1	UniProt-SPEC-100
32	silicic acid	1	TS2

The table presents the label names as the incorrect substrate predictions of samples in TS2 with the frequency of each label in misclassifications. The table also covers the dataset to which each predicted label belongs to.

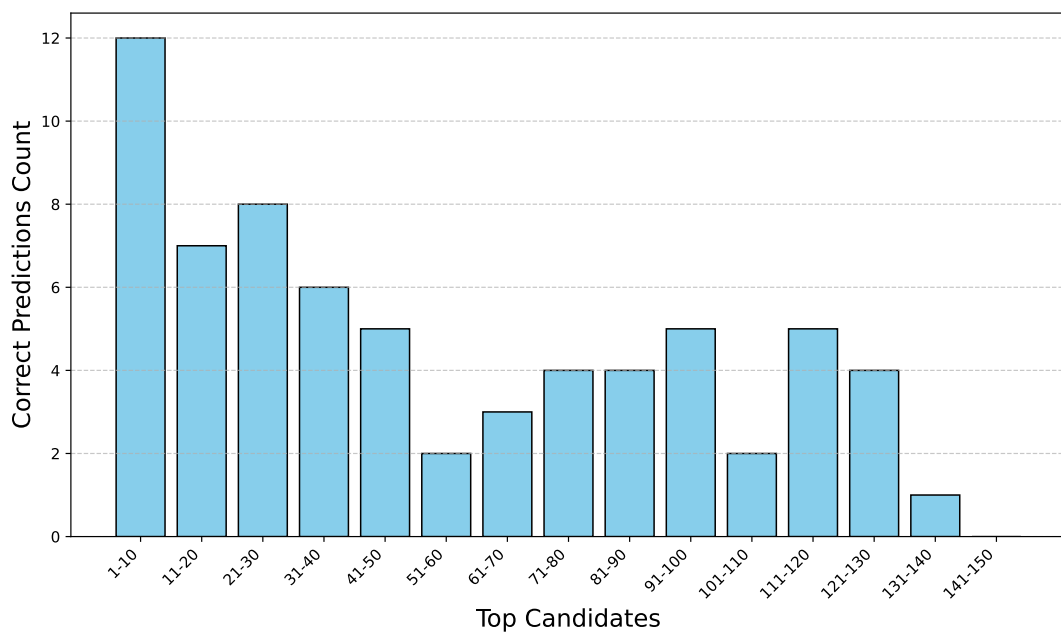


Figure 46: Distribution of correct prediction range for unseen class samples in TS2. The figure shows the number of correct predictions in top candidates for each test sample in TS2 ranging from 1 to 150 top candidates with 10-step intervals.

Table 41: The correct substrate prediction in TS2 for each top candidates range

Candidates	# Correct Predictions	Hit rate
Top 3	3	0.043
Top 5	7	0.100
Top 10	12	0.171
Top 50	38	0.543
Top 100	56	0.800

The table shows the correct prediction hit rate and the number of correct predictions in a range of top candidates for each substrate. The hit rate is calculated as the number of correct predictions in each range/total number of samples in test set TS2.

6.6 Conclusion

The main research question of this project is: *“Given a de novo transmembrane transport protein, can a model predict the specific substrate it carries?”*. The project aims to address this problem within an open-world system model that predicts the substrate for any transmembrane transport protein. The TooTranslator model is implemented to predict the specific substrate for an input protein, trained on the primary structure of the protein, the definition, and the chemical SMILES string of the substrate.

By redefining the classification problem as a regression task, TooTranslator aims to leverage the zero-shot learning process to predict both groups of seen and unseen classes during training (group A and group B in Section 6.3). This model outperforms previous classification-based models for substrate specificity prediction on seen classes (TooT-SPEC and TooT-Triplet-SPEC in Section 3.7.3 and Chapter 4) achieving broader class coverage while maintaining high reported performance values. However, the model’s performance on unseen classes remains suboptimal, with only 9 out of 70 samples from unseen classes correctly labeled among the 10 nearest labels.

As detailed in Section 6.3.3, the model successfully maps the protein-encoding space to the label-encoding space for seen classes. Two fully connected layers and one residual layer implement the mapping. A Huber loss function and the AdamW optimizer train the network. Loss back-propagation to the protein encoder produces accurate representations for samples in seen classes (Section 6.4) and improves protein-label distances for unseen classes (Section 6.5.1).

Although the model does not fully resolve the unseen class substrate prediction in the first substrate candidate, 80% of unseen test samples have the correct prediction within the top 100 candidates (Section 6.5.4). A key direction for future work is adjusting the training strategy to enable gradient back-propagation to both the label and protein encoders. If sufficient resources allow storing all three language models in memory during training, the representation of unseen classes may improve.

Chapter 7

Conclusion

7.1 Introduction

This research addresses the challenge of predicting the substrate specificity of transmembrane transport proteins, a critical yet underexplored problem in molecular biology using computational methods. Experimental complexities, such as the need for membrane systems and protein-ligand binding assays, constrain traditional wet lab approaches to understanding substrate specificity. These limitations, coupled with the scarcity of experimentally characterized transmembrane transport proteins, highlight the importance of computational methods. This research employed large language models to navigate these complexities, leveraging their capabilities in protein understanding and prediction tasks.

The study utilizes machine learning, particularly low-shot learning, to predict specific substrates for these proteins. While previous studies successfully predicted general substrate classes, this work extends the models to the more granular level of specific substrate prediction. We also addressed the challenge of *de novo* sequence prediction, an area where current models fall short. While the proposed model did not fully resolve the issue, it yielded promising results worth exploring further.

The applications of this study are broad, as substrate specificity is fundamental to understanding many biological processes. For instance, predicting substrate specificity can aid in drug development by identifying potential targets for transport inhibitors or enhancers. Additionally, substrate transport mechanisms are integral to neural signaling, where they influence the synaptic uptake of neurotransmitters. These examples underscore the importance of accurate predictions, which may advance these fields.

7.2 Contributions

This study is conducted within a defined scope and resource constraints, which should be considered when interpreting its contributions. It focuses exclusively on predicting substrate specificity in transmembrane transport proteins, excluding multi-label transport proteins from the dataset. The study employs only one protein language model, ProtBERT-BFD, and the contributions are not generalizable to other protein language models. Additionally, it is constrained by limited computational resources and time. The contributions presented in this thesis are derived from results within this specific context. Extending these findings to multi-label transport proteins, other characterization tasks, or different protein language models requires further research. In the following, we expand on contributions mentioned in Section 1.7 providing details for each point.

Contribution 1: Transfer learning using ProtBERT-BFD as the protein language model is effective for classifying specific substrates of transmembrane transport proteins with small labeled datasets.

Background: This research enhances substrate prediction by leveraging transfer learning. Traditional models trained from scratch require extensive data per class. For example, TrSSP, TranCEP, and TooT-SC require a minimum of 60, 70, and 26 sequences per class, respectively [MCZ14, AAB20, AB22a].

Evidence: The models developed in this thesis utilize pre-trained language models, transferring knowledge from pretraining to downstream tasks. This approach significantly reduces the number of samples required per class. The TooT-ICAT model predicts specific substrates with as few as 10 samples per class. This model achieves an MCC of 0.808 ± 0.02 in 5-fold cross-validation (Section 3.6, Table 7).

Contribution 1a: Transfer learning using ProtBERT-BFD as the protein language model is effective in classifying specific substrates of transmembrane transport proteins with long-tailed imbalanced labeled datasets

Evidence: TooT-SPEC further reduces the minimum class size to three samples while maintaining high predictive performance. This model classifies 83 out of 96 specific substrates in an imbalanced long-tailed dataset, UniProt-SPEC-100 (Figure 15) and achieves an MCC of 0.906 ± 0.02 in 3-fold cross-validation (Section 3.7.6, Table 12). However, it should be mentioned that the model leaves 13 specific substrates unclassified with zero True or False positive predictions (Table 13).

Contribution 2: Metric learning using the triplet strategy effectively extends the application of transfer learning for classifying specific substrates of transmembrane transport proteins using ProtBERT-BFD as the protein language model, particularly with long-tailed and imbalanced labeled datasets.

Background: Transmembrane transport protein datasets often exhibit a long-tailed class imbalance, where majority classes have many samples, while minority classes have very few (see Figure 15 for UniProt-SPEC-100 distribution). Conventional fine-tuning methods fail to address this imbalance, often resulting in minority classes with no true or false positive detections. For instance, the TooT-SPEC model (Section 3.7.3), a fine-tuned Prot-BERT-BFD network for predicting 96 substrates, leaves 13 minority classes without any assigned test samples (Appendix C.2, Table 45). These minority classes (listed in Table 13) remain uncovered, with their test samples misclassified into majority classes.

Evidence: TooT-Triplet-SPEC applies metric learning with triplets (anchor, positive, and negative samples) to enhance effective class coverage while maintaining overall performance. According to Table 26 in Section 5.8.1, it identifies 91 out of 96 specific substrate classes, covering eight more minority classes than TooT-SPEC. The model achieves an MCC of 0.893, comparable to 0.908 in TooT-SPEC for the hold-out test set of UniProt-SPEC-100, with no statistically significant difference. The results demonstrate improved minority class coverage while maintaining model performance through metric learning.

Contribution 3: Open-world classification using thresholded Prot-BERT-BFD detects novel specific substrate classes that are not included in the training process.

Background: Datasets for transmembrane transport proteins are often incomplete due to limited data and labor-intensive manual annotation. Many substrate classes are excluded from the training set because they lack sufficient known proteins for ML model training. However, in real-world applications such as transportome analysis, new samples from these unseen classes frequently appear, requiring the model to make predictions.

State-of-the-art (SOTA) models handle unseen classes by assigning them to a class as `miscellaneous` or `others`, but this solution reduces performance as `miscellaneous` class merges small, diverse classes into one class. In a 10-fold cross-validation of TooT-SC, the MCC for `miscellaneous` is 0.5858 ± 0.0136 , while other classes range from 0.6830 ± 0.0084 to 0.9376 ± 0.0000 . The 0.1 MCC gap between `miscellaneous` and the next worst-performing class, `Other organonitrogens`, significantly impacts the model’s overall performance.

Evidence: The TooT-Open-ICAT model employs an open-world classification approach by setting class-specific thresholds based on cross-validation. It determines the threshold τ for each class as the minimum softmax value from validation sets. During testing, if a sample’s softmax value falls below τ , it is flagged as a novel (unseen) class. However, the model’s performance depends on the validation set distribution, making it prone to errors. While the model did not fully resolve the issue, it yielded promising results worth further exploration.

According to Table 17, a closed-world classifier with an additional `miscellaneous` class achieves an MCC of 0.620 on the UniProt-ICAT-100 hold-out test set. In contrast, the TooT-Open-ICAT model, which applies thresholds calculated in cross-validation, improves performance to 0.944 MCC. This model detects novel class samples by flagging them as unseen but does not assign them to a specific substrate class.

Contribution 4: A regression-based model developed with ProtBERT-BFD, SciBERT, and ChemBERTa-2 provides a foundation for identifying novel specific substrate classes.

Evidence: The TooTranslator model extends open-world classification by predicting 10 candidate substrate classes for test samples. Reformulating the problem as regression, it first encodes proteins and substrates using three PLMs and LLMs. The regression model then maps protein encodings to substrate encodings, creating a mutual space where protein sequences are positioned near their associated substrates.

During testing, the model encodes a test protein and ranks the top ten closest chemical substrates in the shared space. Its flexibility allows new substrate classes to be introduced without retraining, as new chemicals can be added using their description and SMILES string encoded by the trained model.

Since this is the first attempt in the field to predict specific novel substrates for new proteins, no existing models serve as direct baselines. While it does not fully solve the novel substrate identification problem, Table 39 and Figure 45 show that training reduces the distance between unseen protein sequences and their true labels, despite these classes being absent during training. This highlights the model’s potential to generalize from known to novel substrate classes.

7.3 Developed Tools

The following tools were developed during the thesis’s implementation process, contributing to the field of bioinformatics through either automated processes or machine learning-based models to characterize transmembrane transport proteins.

- **SGI Dataset Generator** tool is an ontology-based pipeline for the production of preprocessed datasets of protein sequences integrating data from multiple resources. This pipeline addresses the absence of protein sequence datasets labeled with interacting substrates. Given a user-defined substrate group, it automatically constructs a dataset of protein sequences labeled with their corresponding substrate class. This automation reduces time, cost, and human error compared to manual dataset creation, enabling a shift from general to fine-grained labeling—such as expanding from 11 general classes in the TooT-SC dataset to 96 substrate classes in UniProt-SPEC-100 for transmembrane transport proteins. The pipeline is flexible and can be applied to other protein groups interacting with substrates in the ChEBI database, such as enzymes.
- **TooT-ICAT** and **TooT-SPEC** study the applicability of protein language models, specifically ProtBERT-BFD, to substrate specificity prediction. Sections 3.6 and 3.7.3 demonstrate the reliability of PLMs in substrate prediction tasks. Two models were fine-tuned on inorganic ion transporter prediction and transmembrane transport prediction, improving the state-of-the-art models’ class granularity from general classes to specific substrate prediction.
- **TooT-Open-ICAT** is an open classification approach on predicting substrate specificity for inorganic ion transporters (Chapter 4). This model redefines substrate specificity prediction as an open classification problem, allowing it to differentiate between known transporters and novel unseen classes. Built on a transformer-based architecture using two ProtBERT-BFD networks, the model first identifies inorganic ion transport proteins and then predicts their specific substrates. A threshold is set based on training data to flag unseen class samples, enabling the model to handle real-world datasets with diverse specific substrates, overcoming the limitation of being confined to the training classes (Section 4.4.5).
- **TooT-Triplet-SPEC** is a model for one-shot learning of substrate specificity in imbalanced datasets of transport proteins (Chapter 5). This model employs a triplet

sampling strategy to train a ProtBERT-BFD model, followed by a K-NN classifier to optimize sample distances in the encoding space. By using a triplet loss function, the model minimizes intra-class distances and maximizes inter-class distances, effectively tackling class imbalance issues. The approach incorporates both online and offline sampling strategies, along with various distance metrics (Manhattan, Euclidean, and Cosine similarity), to improve the representation of minority classes. The results highlight the effectiveness of the TooT-Triplet-SPEC model in classifying transmembrane transport proteins, particularly in minority classes with as few as three samples (Section 5.8).

- **TooTranslator** investigates *de novo* protein prediction using mappings between large language model encoding spaces. TooTranslator is a deep learning model based on three Large Language Models designed to “translate” a protein sequence into its corresponding substrate (Chapter 6). This study explores whether switching from a classification to a regression task would enable the model to predict unseen classes. In this approach, labels are replaced by encoding vectors constructed from label definitions and compound SMILES strings to create a vector space. While TooTranslator outperforms previous models in predicting known classes, it still faces challenges with unseen class predictions. However, the model shows improvement in reducing the distance between unseen class samples and their actual labels, even without training on these classes (Section 6.5.1), offering the potential for *de novo* prediction in future research.

7.4 Addressing Research Questions

In this section, the research questions presented in Section 1.5 are revisited to highlight how each has been addressed throughout the thesis. We also discuss the key lessons learned for each question and examine any remaining limitations.

The first research question addressed in this thesis is: “*Given a transportome of a specific organism, can a model predict the specific inorganic ions each protein transports?*” The TooT-Open-ICAT (Chapter 4) framework revisits this classic classification problem by introducing a rejection threshold for unseen substrate transporters within a transportome. However, determining this threshold proved to be an intricate.

The key lesson learned was that the threshold is highly dependent on the training and validation datasets and tends to perform poorly on test samples that fall outside

the distribution of these sets, as is often the case in transportome analysis. This study is limited to thresholding at two layers—the penultimate layer and the softmax layer. Future work could explore expanding this approach to incorporate thresholds within the inner layers of the network for learning out of distribution samples to improve performance.

The second research question explored in this thesis is: “*Given a transmembrane transport protein from a long-tailed imbalanced dataset, can a model predict the specific substrate it carries?*” This challenge is addressed using metric learning, implemented in the TooT-Triplet model (Chapter 5). The approach demonstrated effectiveness by successfully predicting 91 out of 96 substrate classes. However, five minority classes remained as unpredicted, with zero true or false positive samples for these classes. A key limitation of this model is the computational constraint associated with the online sampling strategy, which restricted a more extensive evaluation of the approach. Future directions could involve investigating cost-effective methods to implement the online sampling strategy to improve both scalability and performance.

The final research question addressed in this thesis is: “*Given a de novo transmembrane transport protein, can a model predict the specific substrate it carries?*” This challenge is tackled using TooTranslator (Chapter 6), the first effort in novel substrate identification. The model ranks the top 10 candidate substrates for a given protein based on their distance from the encoded protein in the mutual embedding space. A key insight from this project is that replacing categorical labels with descriptive label representations shifts the task from classification to regression. This transition introduces greater flexibility, enabling the model to generalize to unseen protein or substrate classes during testing. However, a limitation of TooTranslator is that it modifies only a small portion of the embedding space, constrained by the limited size of the training set. Future work could expand the training data to enhance the embedding space coverage and improve model generalization.

7.5 Research publications

This section presents the publications resulting from this research, which include two conference papers (one published), three journal papers (in process), and two posters presented at the listed conferences.

1. Predicting the specific substrate for transmembrane transport proteins using BERT language model [AB22b]

- Sima Ataei, Gregory Butler, 2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB); pages 1-8
2. Performance of a fine-tuned ProtBERT model on the classification of imbalanced transporter protein datasets
Sima Ataei, Gregory Butler, ISCB-Latin America SoIBio BioNetMX Conference on Bioinformatics, 2022
 3. Substrate specificity prediction of transporters using triplet loss and protein language models
Sima Ataei, Gregory Butler, 32nd Conference on Intelligent Systems for Molecular Biology (ISMB), 2024
 4. TooT-SS: Transfer Learning using ProtBERT-BFD Language Model for Predicting Specific Substrates of Transport Proteins
Sima Ataei, Gregory Butler, Accepted for 15th International Conference on Bio-science, Biochemistry and Bioinformatics (ICBBB), 2025
 5. TooT-Open-ICAT: Open classification of inorganic ion transport proteins
Sima Ataei, Gregory Butler, journal paper in progress
 6. TooT-Triplet-SPEC: One-shot classification of transmembrane transport proteins
Sima Ataei, Gregory Butler, journal paper in progress
 7. TooTranslator: Zero-shot classification of *de novo* transport proteins
Sima Ataei, Gregory Butler, journal paper in progress

7.6 Future work

One potential direction for improving the TooT-Open-ICAT model is enhancing its ability to handle out-of-distribution (OOD) samples, particularly in low-data scenarios where the available training data may not cover the test set distributions. Investigating models that are better equipped for OOD prediction, using techniques such as uncertainty quantification, domain adaptation, or adversarial learning, could improve the model's generalization. This would be especially beneficial in real-world applications where novel or previously unseen classes may appear.

While TooTranslator shows promise in reducing distances between unseen class samples and their actual labels, further improvements are needed for accurate *de novo* prediction. Key areas for exploration include refining the projection process to better map protein sequences and substrate encoding vectors, experimenting with alternative neural architectures such as graph-based networks or transformer variations, and incorporating multi-task learning to simultaneously learn substrate-specific and sequence-specific information. Additionally, fine-tuning the model using larger and more diverse datasets could help it generalize better to unseen classes.

Future work could also investigate integrating multi-modal data sources, such as protein structures, ligand binding sites, and evolutionary data, alongside sequence information. This integration could enhance the prediction of substrate specificity by capturing more intricate biological relationships. Techniques like multi-modal deep learning and attention mechanisms across modalities could be explored to address this.

Additionally, the tools developed in this research — TooT-Open-ICAT, TooT-Triplet-SPEC, and TooTranslator — could benefit from testing with biological data. While computational validation provides reliable insights and predictions, biological validation would enhance the robustness and applicability of the models, showcasing their potential to generate new biological knowledge. Collaborations with experimental biologists could open opportunities to validate the predicted substrate specificities in real-world systems, reinforcing the findings and unlocking the transformative potential of these tools.

Bibliography

- [AAA⁺23] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [AAB20] Munira Alballa, Faizah Aplop, and Gregory Butler. TranCEP: Predicting the substrate class of transmembrane transport proteins using compositional, evolutionary, and positional information. *PloS One*, 15(1):e0227683, 2020.
- [AB15] Faizah Aplop and Greg Butler. On predicting transport proteins and their substrates for the reconstruction of metabolic networks. In *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–9. IEEE, 2015.
- [AB19] Munira Alballa and Gregory Butler. Ontology-based transporter substrate annotation for benchmark datasets. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2613–2619. IEEE, 2019.
- [AB22a] Munira Alballa and Gregory Butler. TooT-SC: Predicting eleven substrate classes of transmembrane transport proteins. *bioRxiv*, 2022.
- [AB22b] Sima Ataei and Gregory Butler. Predicting the specific substrate for transmembrane transport proteins using BERT language model. In *2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–8. IEEE, 2022.
- [ABB⁺00] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, and Janan T. Eppig. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

- [ABH⁺15] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander D Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Essential Cell Biology*. Garland Science, 2015.
- [AHK01] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory - Proceedings of ICDT 2001: 8th International Conference, London, UK, January 4–6, 2001*, pages 420–434. Springer, 2001.
- [AJL⁺03] Bruce Alberts, Bruce Johnson, Julian Lewis, Keith Raff, Martin Roberts, and Peter Walter. *Molecular Biology of the Cell*. Oxford University Press, 4 edition, 2003.
- [AKEG09] Yalini Arinaminpathy, Ekta Khurana, Donald M Engelman, and Mark B Gerstein. Computational analysis of membrane proteins: the largest class of drug targets. *Drug discovery today*, 14(23-24):1130–1135, 2009.
- [ASC⁺22] Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERT-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022.
- [ASY⁺19] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631, 2019.
- [BB15] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1893–1902, 2015.
- [BDA13] Mohamed Bekkar, Hassiba Kheliouane Djemaa, and Taklit Akrouf Ali-touche. Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*, 3:27–38, 2013.
- [BDR05] Yoshua Bengio, Olivier Delalleau, and Nicolas Roux. The curse of highly variable functions for local kernel machines. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18, pages 107–114. MIT Press, 2005.

- [Ber13] James O Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media, 2013.
- [BGL⁺93] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a “siamese” time delay neural network. In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6, pages 737–744. Morgan-Kaufmann, 1993.
- [BGRS99] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is “nearest neighbor” meaningful? In *Database Theory - Proceedings of ICDT’99: 7th International Conference Jerusalem, Israel, January 10–12, 1999*, pages 217–235. Springer, 1999.
- [BHH⁺78] George EP Box, William H Hunter, Stuart Hunter, et al. *Statistics for Experimenters*, volume 664. John Wiley and sons New York, 1978.
- [BHS15] Aurélien Bellet, Amaury Habrard, and Marc Sebban. *Metric Learning*. Morgan & Claypool Publishers, 2015.
- [BLC19] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620. Association for Computational Linguistics, 2019.
- [BMR⁺20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.
- [BN06] Christopher M Bishop and Nasser M Nasrabadi. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006.

- [BPPK24] Preston Billion Polak, Joseph D Prusa, and Taghi M Khoshgoftaar. Low-shot learning and class imbalance: a survey. *Journal of Big Data*, 11(1):1, 2024.
- [Brö20] Stefan Bröer. Amino acid transporters as targets for cancer therapy: why, where, when, and how. *International Journal of Molecular Sciences*, 21(17):6156, 2020.
- [BS18] Doreen Braun and Ulrich Schweizer. Thyroid hormone transport and transporters. *Vitamins and Hormones*, 106:19–44, 2018.
- [BS20] Pabitra Krishna Bhattacharya and Prakash B Samnani. *Metal Ions in Biochemistry*. CRC press, 2020.
- [CDTTN12] Jia-Ming Chang, Paolo Di Tommaso, Jean-François Taly, and Cedric Notredame. Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. *BMC Bioinformatics*, 13(4):1–7, 2012.
- [DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.
- [DDME⁺07] Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(suppl_1):D344–D350, 2007.
- [Dia14] George Dhallinas. Understanding transporter specificity and the discrete appearance of channel-like gating domains in transporters. *Frontiers in Pharmacology*, 5:207, 2014.
- [Din11] Zejin Ding. *Diversified Ensemble Classifiers for Highly Imbalanced Data Learning and its Application in Bioinformatics*. PhD thesis, Georgia State University, 2011.

- [DL15] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 3079–3087. Curran Associates, Inc., 2015.
- [DOM02] Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, 35(5-6):352–359, 2002.
- [DP18] Laura J Dunphy and Jason A Papin. Biomedical applications of genome-scale metabolic network reconstructions of human pathogens. *Current opinion in biotechnology*, 51:70–79, 2018.
- [DRFR10] Oscar Dias, Miguel Rocha, Eugénio C Ferreira, and Isabel Rocha. merlin: Metabolic models reconstruction using genome-scale information. *IFAC Proceedings Volumes*, 43(6):120–125, 2010.
- [DTM19] Enrico Drioli, Elena Tocci, and Francesca Macedonio. Membrane science and membrane engineering: A successful story. *General Chemistry*, 5(2):180027, 2019.
- [EGZ95] Bruce E Eaton, Larry Gold, and Dominic A Zichi. Let’s get specific: the relationship between specificity and affinity. *Chemistry & Biology*, 2(10):633–638, 1995.
- [EHD⁺22] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehaw, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022.
- [EMT07] Sean Ekins, Jordi Mestres, and Bernard Testa. *In silico* pharmacology for drug discovery: methods for virtual ligand screening and profiling. *British Journal of Pharmacology*, 152(1):9–20, 2007.
- [Eur] European Bioinformatics Institute. *ChEBI User Manual* . Retrieved on January 3, 2025.

- [FGHC18] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61:863–905, 2018.
- [FH22] Noelia Ferruz and Birte Höcker. Controllable protein design with language models. *Nature Machine Intelligence*, 4(6):521–532, 2022.
- [FNZ⁺12] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.
- [FS89] Stanley Fields and Ok-kyu Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, 1989.
- [FZC18] Yi Fu, Ji Zhao, and Zhiguo Chen. Insights into the molecular mechanisms of protein-ligand interactions by molecular docking and molecular dynamics simulation: a case of oligopeptide binding protein. *Computational and mathematical methods in medicine*, 2018(1):3502514, 2018.
- [GAL⁺06] David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. A closer look at skip-gram modelling. In *LREC*, volume 6, pages 1222–1225, 2006.
- [GME87] Michael Gribskov, Andrew D. McLachlan, and David Eisenberg. Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences*, 84(13):4355–4358, 1987.
- [Gol87] Ira D Goldfine. The insulin receptor: molecular biology and transmembrane signaling. *Endocrine Reviews*, 8(3):235–255, 1987.
- [Gor04] Jan Gorodkin. Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, 28(5):367–374, 2004.
- [GTC⁺21] Yu Gu, Robert Tinn, Hao Cheng, et al. PubMedBERT: A pre-trained biomedical language model for biomedical text mining. *Bioinformatics*, 37(3):367–376, 2021.

- [HAB⁺13] David P. Hill, Nico Adams, Mike Bada, Colin Batchelor, Tanya Z. Berardini, Heiko Dietze, Harold J. Drabkin, Marcus Ennis, Rebecca E. Foulger, Midori A. Harris, Janna Hastings, Namrata S. Kale, Paula de Matos, Christopher J. Mungall, Gareth Owen, Paola Roncaglia, Christoph Steinbeck, Steve Turner, and Jane Lomax. Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology. *BMC Genomics*, 14(1):513, 2013.
- [HBR19] Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. Aspect-based sentiment analysis using BERT. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196, 2019.
- [HG09] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.
- [HLS⁺22] Michael Heinzinger, Maria Littmann, Ian Sillitoe, Nicola Bordin, Christine Orengo, and Burkhard Rost. Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genomics and Bioinformatics*, 4(2):lqac043, 2022.
- [HLTZ18] Anfeng He, Chong Luo, Xinmei Tian, and Wenjun Zeng. A twofold siamese network for real-time object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4834–4843, 2018.
- [HNG⁺10] Ying Huang, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, 2010.
- [HOD⁺16] Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, 44(D1):D1214–D1219, 2016.
- [HSW⁺22] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

- [HTF09] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, volume 2. Springer, 2009.
- [Hub92] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 492–518. Springer, 1992.
- [HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [JDHL19] Xiaoyang Jing, Qiwen Dong, Daocheng Hong, and Ruqian Lu. Amino acid encoding methods for protein sequences: A comprehensive review and assessment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(6):1918–1931, 2019.
- [JEP⁺21] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.
- [JPN18] Katharina EJ Jungnickel, Joanne L Parker, and Simon Newstead. Structural basis for amino acid transport by the CAT family of SLC7 transporters. *Nature communications*, 9(1):550, 2018.
- [JTD23] Márton A Jambrich, Gabor E Tusnady, and Laszlo Dobson. How AlphaFold2 shaped the structural coverage of the human transmembrane proteome. *Scientific Reports*, 13(1):20283, 2023.
- [KB15] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

- [KKP06] Sotiris Kotsiantis, Dimitris Kanellopoulos, and Panayiotis Pintelas. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36, 2006.
- [KMPG08] Gregory J. Kaczorowski, Owen B. McManus, Birgit T. Priest, and Maria L. Garcia. Ion channels as drug targets: the next GPCRs. *The Journal of General Physiology*, 131(5):399–405, 2008.
- [Kos01] Timo Koski. *Hidden Markov Models for Bioinformatics*, volume 2. Springer Science & Business Media, 2001.
- [Kra16] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.
- [Küh14] Werner Kühlbrandt. The resolution revolution. *Science*, 343(6178):1443–1444, 2014.
- [KWH15] Bartosz Krawczyk, Michał Woźniak, and Francisco Herrera. On the usefulness of one-class classifier ensembles for decomposition of multi-class problems. *Pattern Recognition*, 48(12):3969–3982, 2015.
- [KZS⁺15] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, pages 1–30. Lille, 2015.
- [Lan19] Zhenzhong Lan. Albert: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.
- [LB94] Stephen J Lippard and Jeremy Mark Berg. *Principles of Bioinorganic Chemistry*. University Science Books, 1994.
- [LBK⁺21] Harvey F Lodish, Arnold Berk, Chris Kaiser, Monty Krieger, Anthony Bretscher, Hidde L Ploegh, Kelsey C Martin, Michael B Yaffe, and Angelika Amon. *Molecular Cell Biology*, volume 1. WH Freeman New York, 2021.
- [LEB08] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, volume 1, page 3, 2008.

- [LFL⁺23] Davide Lagoa, José P. Faria, Filipe Liu, Emanuel Cunha, Christopher S. Henry, and Oscar Dias. TranSyT, an innovative framework for identifying transport systems. *Bioinformatics*, 39(8):btad466, 08 2023.
- [LG06] Weizhong Li and Adam Godzik. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.
- [LGG⁺17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.
- [Liu19] Yinhan Liu. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019.
- [LRD21] Mark Lennox, Neil Robertson, and Barry Devereux. Deep learning proteins using a Triplet-BERT network. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4341–4347. IEEE, 2021.
- [LS08] Charles X. Ling and Victor S. Sheng. Cost-sensitive learning and the class imbalance problem. *Encyclopedia of Machine Learning*, 2011:231–235, 2008.
- [LWZ08] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008.
- [LXHNE17] Bo Larsen, Deyang Xu, Barbara Ann Halkier, and Hussam Hassan Nour-Eldin. Advances in methods for identification and characterization of plant transporter function. *Journal of Experimental Botany*, 68(15):4045–4056, 2017.
- [MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [MCZ14] Nitish K Mishra, Junil Chang, and Patrick X Zhao. Prediction of membrane transport proteins and their substrate specificities using primary sequence information. *PloS One*, 9(6):e100278, 2014.

- [MD19] Emmanuel Mikros and George Diallinas. Tales of tails in transporters. *Open Biology*, 9(6):190083, 2019.
- [MMB21] Yassine Marrakchi, Osama Makansi, and Thomas Brox. Fighting class imbalance with contrastive learning. In *Medical Image Computing and Computer Assisted Intervention - Proceedings, Part III, of MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021*, pages 466–476. Springer, 2021.
- [MRA20] Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 243–248. IEEE, 2020.
- [MRIE24] Lasse Middendorf, Bharat Ravi Iyengar, and Lars A Eicholt. Sequence, structure, and functional space of drosophila *de novo* proteins. *Genome Biology and Evolution*, 16(8):evae176, 2024.
- [MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [MT21] Andrew G. McDonald and Keith F. Tipton. Enzyme nomenclature and classification: The state of the art. *The FEBS Journal*, pages 2214–2231, 2021.
- [MVPC12] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *ECCV 2012-12th European Conference on Computer Vision*, volume 7573, pages 488–501. Springer, 2012.
- [OBL21] Dan Ofer, Nadav Brandes, and Michal Linial. The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19:1750–1758, 2021.
- [Pes13] Vladimir Pestov. Is the k-NN classifier in high dimensions affected by the curse of dimensionality? *Computers & Mathematics with Applications*, 65(10):1427–1437, 2013.

- [PGM⁺19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 8026–8037. Curran Associates, Inc., 2019.
- [PL22] Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*, 2022.
- [Pol10] Thomas D. Pollard. A guide to simple and informative binding assays. *Molecular Biology of the Cell*, 21(23):4061–4067, 2010.
- [PSM14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
- [PVG⁺11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Perrot Matthieu, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [PZA07] Sona Pandey, Wei Zhang, and Sarah M. Assmann. Roles of ion channels and transporters in guard cell signal transduction. *FEBS letters*, 581(12):2325–2336, 2007.
- [QB24] Dorian Quelle and Alexandre Bovet. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7:1341697, 2024.

- [RAAS11] Mathias Rask-Andersen, Markus Sällman Almén, and Helgi B Schiöth. Trends in the exploitation of novel drug targets. *Nature Reviews Drug Discovery*, 10(8):579–590, 2011.
- [RBT⁺19] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with TAPE. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 9689–9701. Curran Associates, Inc., 2019.
- [RG19] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics, 2019.
- [ROM19] Tharindu Ranasinghe, Constantin Orăsan, and Ruslan Mitkov. Semantic textual similarity with siamese neural networks. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1004–1011, 2019.
- [Rou87] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.
- [RWC16] Sanjeev Jagannatha Rao, Yufei Wang, and Garrison W Cottrell. A deep siamese neural network learns the human-perceived similarity structure of facial expressions without explicit categories. In *CogSci*, 2016.
- [San19] V Sanh. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- [SG12] Mahendra Sahare and Hitesh Gupta. A review of multi-class classification for imbalanced data. *International Journal of Advanced Computer Research*, 2(3):160, 2012.
- [SJRT⁺16] Milton H. Saier Jr., Vamsee S. Reddy, Brian V. Tsu, Muhammad Saad Ahmed, Chun Li, and Gabriel Moreno-Hagelsieb. The transporter classification database (TCDB): recent advances. *Nucleic Acids Research*, 44(D1):D372–D379, 2016.

- [SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.
- [SLCS19] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019.
- [SLJ⁺15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.
- [SSC19] Sailu Sarvagalla, Safiulla Basha Syed, and Mohane Selvaraj Coumar. An overview of computational methods, tools, servers, and databases for drug repurposing. In *Silico Drug Design*, pages 743–780, 2019.
- [SSS24] Mina Seifi, Ali Soltanmanesh, and Afshin Shafiee. Mimicking classical noise in ion channels by quantum decoherence. *Scientific Reports*, 14(1):16130, 2024.
- [SSZ17] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4077–4087. Curran Associates, Inc., 2017.
- [Ste12] Wes Stein. *The Movement of Molecules across Cell Membranes*, volume 6. Elsevier, 2012.
- [Stu08] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.
- [SUG⁺17] Rita Santos, Oleg Ursu, Anna Gaulton, A. Patrícia Bento, Ramesh S. Donadi, Cristian G. Bologna, Anneli Karlsson, Bissan Al-Lazikani, Anne Hersey, Tudor I. Oprea, and John P. Overington. A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, 16(1):19–34, 2017.

- [SW89] Lars Stahle and Svante Wold. Analysis of variance (ANOVA). *Chemometrics and Intelligent Laboratory Systems*, 6(4):259–272, 1989.
- [SXL17] Lei Shu, Hu Xu, and Bing Liu. DOC: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2911–2916. Association for Computational Linguistics, 2017.
- [SXL18] Lei Shu, Hu Xu, and Bing Liu. Unseen class discovery in open-world classification. *arXiv preprint arXiv:1801.05609*, 2018.
- [SYZ⁺18] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.
- [The15] Sergios Theodoridis. *Machine Learning: A Bayesian and Optimization Perspective*. Academic press, 2015.
- [The21a] The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2021.
- [The21b] The UniProt Consortium. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2021.
- [The23] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023.
- [TP10] Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5(1):93, 2010.
- [TS10] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 242–264. IGI global, 2010.
- [VBL⁺16] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 3630–3638. Curran Associates, Inc., 2016.

- [vdMH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.
- [VMV⁺21] Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Rajani. BERTology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations*, 2021.
- [VSP⁺17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017.
- [WBR⁺10] Christoph J Wienken, Philipp Baaske, Ulrich Rothbauer, Dieter Braun, and Stefan Duhr. Protein-binding assays in biological liquids using microscale thermophoresis. *Nature communications*, 1(1):100, 2010.
- [WDS⁺20] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, 2020.
- [Wei88] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.
- [WGC14] Michał Woźniak, Manuel Grana, and Emilio Corchado. A survey of multiple classifier systems as hybrid systems. *Information Fusion*, 16:3–17, 2014.
- [WHSC14] Ling-Gang Wu, Edaeni Hamid, Wonchul Shin, and Hsueh-Cheng Chiang. Exocytosis and endocytosis: modes, functions, and coupling mechanisms. *Annual Review of Physiology*, 76(1):301–331, 2014.

- [Wil72] Dennis L Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2(3):408–421, 1972.
- [WM05] Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1):79–82, 2005.
- [Wol20] Thomas Wolf. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2020.
- [WP03a] M.A. Webber and L.J.V. Piddock. The importance of efflux pumps in bacterial antibiotic resistance. *Journal of Antimicrobial Chemotherapy*, 51(1):9–11, 2003.
- [WP03b] Gary M Weiss and Foster Provost. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19:315–354, 2003.
- [Wüt89] Kurt Wüthrich. Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *Science*, 243:4887, 1989.
- [WYKN20] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (csur)*, 53(3):1–34, 2020.
- [XW22] Hanwen Xu and Sheng Wang. ProTranslator: zero-shot protein function prediction using textual description. In *International Conference on Research in Computational Molecular Biology*, pages 279–294. Springer, 2022.
- [YH15] Nicole J Yang and Marlon J Hinner. Getting across the cell membrane: an overview for small molecules, peptides, and proteins. *Site-Specific Protein Labeling: Methods and Protocols*, pages 29–53, 2015.
- [Yin22] Jie Yin. Research on question answering system based on BERT model. In *2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)*, pages 68–71. IEEE, 2022.

- [YZLL21] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.
- [ZE02] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 694–699, 2002.
- [ZF14] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.
- [ZKH⁺23] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep Long-Tailed Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023.
- [ZL10] Zhi-Hua Zhou and Xu-Ying Liu. On multi-class cost-sensitive learning. *Computational Intelligence*, 26(3):232–257, 2010.
- [ZWGC15] Yun-Kai Zhang, Yi-Jun Wang, Pranav Gupta, and Zhe-Sheng Chen. Multidrug resistance proteins (MRPs) and cancer therapy. *The AAPS journal*, 17:802–812, 2015.
- [ZWZ⁺20] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. Semantics-aware BERT for language understanding. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 34, pages 9628–9635, 2020.
- [ZXL⁺16] Quan Zou, Sifa Xie, Ziyu Lin, Meihong Wu, and Ying Ju. Finding the best classification threshold in imbalanced classification. *Big Data Research*, 5:2–8, 2016.

Appendix A

C2GO Mapping

A.1 Introduction

This chapter is dedicated to presenting the C2GO mapping. As described in Section 3.4, this map has been extracted from both ChEBI and Gene Ontology to relate these two annotations. Each specific substrate in ChEBI ontology has been mapped to one or multiple GO terms related to the transmembrane transport of the substrate. This map has been used in the research to label the datasets.

Table 42: C2GO map with ChEBI terms and their corresponding GO terms

ChEBI term	GO terms
CHEBI:17843 transfer RNA	[GO:0051034 tRNA transmembrane transporter activity]
CHEBI:57603 ethanolaminium(1+)	[GO:0034228 ethanolamine transmembrane transporter activity]
CHEBI:63534 monoamine	[GO:0015311 monoamine:proton antiporter activity]
CHEBI:35237 cysteine zwitterion	[GO:0033230 ABC-type cysteine transporter activity]
CHEBI:22918 branched-chain amino acid	[GO:0015657 branched-chain amino acid:sodium symporter activity]
CHEBI:35418 N-acetylneuraminate	[GO:0015306 sialate:cation symporter activity]
CHEBI:29034 iron(3+)	[GO:0015090 low-affinity ferric iron ion transmembrane transporter activity, GO:0015408 ABC-type ferric iron transporter activity, GO:0015092 high-affinity, ferric iron transmembrane transporter activity, GO:0015091 ferric iron, transmembrane transporter activity]
CHEBI:38290 L-ascorbate	[GO:0015229 L-ascorbic acid transmembrane transporter activity, GO:0008520 L-ascorbate:sodium symporter activity]
CHEBI:57977 bilirubin(2-)	[GO:0015351 bilirubin secondary active transmembrane transporter activity]
CHEBI:456216 ADP(3-)	[GO:0005471 ATP:ADP antiporter activity]
CHEBI:30584 silicon(4+)	[GO:0032523 silicon efflux transmembrane transporter activity]
CHEBI:67063 poly-beta-1,6-N-acetyl-D-glucosamine	[GO:1901515 poly-beta-1,6-N-acetyl-D-glucosamine transmembrane transporter activity]
CHEBI:35172 vanadium cation	[GO:0015100 vanadium ion transmembrane transporter activity]
CHEBI:5019 ferrichrome	[GO:0042929 ferrichrome transmembrane transporter activity]
CHEBI:17544 hydrogencarbonate	[GO:0140413 zinc:bicarbonate:selenite symporter activity, GO:0015383 sulfate:bicarbonate antiporter activity, GO:0008510 sodium:bicarbonate symporter activity]
CHEBI:24838 inorganic phosphate	[GO:0015319 sodium:inorganic phosphate symporter activity]
CHEBI:35391 aspartate(1-)	[GO:0070906 aspartate:alanine antiporter activity]
CHEBI:17568 uracil	[GO:0015505 uracil:cation symporter activity]
CHEBI:58095 L-phenylalanine zwitterion	[GO:0015192 L-phenylalanine transmembrane transporter activity]

Continued on next page

Table 42: C2GO map with ChEBI terms and their corresponding GO terms

ChEBI term	GO terms
CHEBI:49713 lithium(1+)	[GO:0010348 lithium:proton antiporter activity]
CHEBI:16988 D-ribose	[GO:0015611 ABC-type D-ribose transporter activity]
CHEBI:57762 L-valine zwitterion	[GO:0005304 L-valine transmembrane transporter activity]
CHEBI:350546 serotonin(1+)	[GO:0005335 serotonin:sodium symporter activity]
CHEBI:46081 fluconazole	[GO:0015313 fluconazole:proton antiporter activity]
CHEBI:37565 GTP(4-)	[GO:0010292 GTP:GDP antiporter activity]
CHEBI:60940 D-mannosyl-D-glyceric acid	[GO:0051477 mannosylglycerate transmembrane transporter activity]
CHEBI:26948 vitamin B1	[GO:0048502 ABC-type thiamine transporter activity]
CHEBI:24163 galactoside	[GO:0044669 sodium:galactoside symporter activity]
CHEBI:57972 L-alanine zwitterion	[GO:0140407 L-alanine:proton antiporter activity]
CHEBI:26546 rhamnose	[GO:0015561 rhamnose:proton symporter activity]
CHEBI:15354 choline	[GO:0033266 ABC-type choline transporter activity, GO:0005307 choline:sodium symporter activity]
CHEBI:24996 lactate	[GO:0015650 lactate:proton symporter activity]
CHEBI:13389 NAD	[GO:0051724 NAD transmembrane transporter activity]
CHEBI:27570 histidine	[GO:0070907 histidine:histamine antiporter activity]
CHEBI:29036 copper(2+)	[GO:0043682 P-type divalent copper transporter activity]
CHEBI:63056 zinc cation	[GO:0016463 P-type zinc transporter activity, GO:0140413 zinc:bicarbonate:selenite symporter activity, GO:0015633 ABC-type zinc transporter activity]
CHEBI:64627 microcin	[GO:0015638 microcin transmembrane transporter activity]
CHEBI:18186 tyrosine	[GO:0070908 tyrosine:tyramine antiporter activity]
CHEBI:26977 thiosulfate	[GO:0015117 thiosulfate transmembrane transporter activity]
CHEBI:16961 monoacylglycerol phosphate	[GO:0051978 lysophospholipid:sodium symporter activity]
CHEBI:17698 chloramphenicol	[GO:0042896 chloramphenicol transmembrane transporter activity]
CHEBI:32682 L-argininium(1+)	[GO:0097626 low-affinity L-arginine transmembrane transporter activity, GO:0005289 high-affinity L-arginine transmembrane transporter activity]
CHEBI:57812 CMP-N-acetyl-beta-neuraminate(2-)	[GO:0005456 CMP-N-acetylneuraminate transmembrane transporter activity]
CHEBI:30411 cobalamin	[GO:0015420 ABC-type vitamin B12 transporter activity]
CHEBI:15792 malonate(2-)	[GO:0044668 sodium:malonate symporter activity]
CHEBI:36264 molybdate	[GO:0015412 ABC-type molybdate transporter activity]
CHEBI:26986 threonine	[GO:0015565 threonine efflux transmembrane transporter activity]
CHEBI:28053 melibiose	[GO:0043887 melibiose:sodium symporter activity]
CHEBI:30823 oleate	[GO:1901480 oleate transmembrane transporter activity]
CHEBI:33118 boric acid	[GO:0046715 active borate transmembrane transporter activity, GO:0080139 borate efflux transmembrane transporter activity]
CHEBI:28793 beta-D-glucan	[GO:0015441 ABC-type beta-glucan transporter activity]
CHEBI:26667 sialic acid	[GO:0015538 sialic acid:proton symporter activity]
CHEBI:33384 L-serine zwitterion	[GO:0015194 L-serine transmembrane transporter activity]
CHEBI:17634 D-glucose	[GO:0055055 D-glucose:proton symporter activity]
CHEBI:57287 coenzyme A(4-)	[GO:0015325 acetyl-CoA:CoA antiporter activity]
CHEBI:37239 molybdenum cation	[GO:0042888 molybdenum ion transmembrane transporter activity]
CHEBI:25140 maltodextrin	[GO:0042958 maltodextrin transmembrane transporter activity]
CHEBI:33871 glycerate	[GO:1901974 glycerate transmembrane transporter activity]
CHEBI:25155 manganese cation	[GO:0015410 ABC-type manganese transporter activity, GO:0140613 P-type manganese transporter activity, GO:0010486 manganese:proton antiporter activity]
CHEBI:28199 ferrienterobactin(3-)	[GO:0015624 ABC-type ferric-enterobactin transporter activity, GO:0015345 ferric enterobactin:proton symporter activity]
CHEBI:37592 organic phosphonate	[GO:0015416 ABC-type phosphonate transporter activity]
CHEBI:18310 alkane	[GO:0015567 alkane transmembrane transporter activity]
CHEBI:59789 S-adenosyl-L-methionine zwitterion	[GO:0000095 S-adenosyl-L-methionine transmembrane transporter activity]
CHEBI:30413 heme	[GO:0015439 ABC-type heme transporter activity]
CHEBI:24592 hexuronic acid	[GO:0015134 hexuronate transmembrane transporter activity]
CHEBI:63137 4-hydroxyproline zwitterion	[GO:0034590 L-hydroxyproline transmembrane transporter activity]
CHEBI:33569 noradrenaline	[GO:0005334 norepinephrine:sodium symporter activity]

Continued on next page

Table 42: C2GO map with ChEBI terms and their corresponding GO terms

ChEBI term	GO terms
CHEBI:17996 chloride	[GO:0019182 histamine-gated chloride channel activity, GO:0015377 cation:chloride symporter activity, GO:0005229 intracellular calcium activated chloride channel activity, GO:0005254 chloride channel activity, GO:0062158 chloride:proton antiporter activity, GO:0005260 intracellularly ATP-gated chloride channel activity, GO:0008511 sodium:potassium:chloride symporter activity, GO:0015379 potassium:chloride symporter activity, GO:0005247 voltage-gated chloride channel activity, GO:0005335 serotonin:sodium symporter activity, GO:0061778 intracellular chloride channel activity, GO:0072320 volume-sensitive chloride channel activity, GO:0005334 norepinephrine:sodium symporter activity, GO:0015108 chloride transmembrane transporter activity, GO:0022852 glycine-gated chloride ion channel activity, GO:0005330 dopamine:sodium symporter activity, GO:0008068 extracellularly glutamate-gated chloride channel activity, GO:0022851 GABA-gated chloride ion channel activity, GO:0061797 pH-gated chloride channel activity]
CHEBI:57861 dethiobiotin(1-)	[GO:1901604 dethiobiotin transmembrane transporter activity]
CHEBI:326268 1,4-butanediammonium	[GO:0015496 putrescine:ornithine antiporter activity, GO:0043861 agmatine:putrescine antiporter activity, GO:0015594 ABC-type putrescine transporter activity]
CHEBI:26288 propanediol	[GO:0015255 propanediol channel activity]
CHEBI:29991 L-aspartate(1-)	[GO:0062057 L-aspartate:fumarate antiporter activity]
CHEBI:57761 pyridoxaminium(1+)	[GO:0031927 pyridoxamine transmembrane transporter activity]
CHEBI:61375 ferric-vibriobactin	[GO:0019535 ferric-vibriobactin transmembrane transporter activity]
CHEBI:23336 cobalt cation	[GO:0032778 P-type cobalt transporter activity]
CHEBI:30616 ATP(4-)	[GO:0005471 ATP:ADP antiporter activity]
CHEBI:16347 (R)-carnitine	[GO:0044667 (R)-carnitine:4-(trimethylammonio)butanoate antiporter activity]
CHEBI:17750 glycine betaine	[GO:0031459 ABC-type glycine betaine transporter activity, GO:0015653 glycine betaine:proton symporter activity]
CHEBI:24175 galacturonate	[GO:0015550 galacturonate transmembrane transporter activity]
CHEBI:58432 histaminium	[GO:0070907 histidine:histamine antiporter activity]
CHEBI:59062 polymyxin	[GO:0042897 polymyxin transmembrane transporter activity]
CHEBI:25197 mercury cation	[GO:0015097 mercury ion transmembrane transporter activity]
CHEBI:27306 vitamin B6	[GO:0031924 vitamin B6 transmembrane transporter activity]
CHEBI:14321 glutamate(1-)	[GO:0015327 cystine:glutamate antiporter activity, GO:0070909 glutamate:gamma-aminobutyric acid antiporter activity]
CHEBI:64702 nucleotide-sulfate	[GO:0005340 nucleotide-sulfate transmembrane transporter activity]
CHEBI:57277 3-(3-hydroxyphenyl)propanoate	[GO:0042926 3-hydroxyphenylpropionic acid transmembrane transporter activity]
CHEBI:63063 cadmium cation	[GO:0015434 ABC-type cadmium transporter activity, GO:0008551 P-type cadmium transporter activity]
CHEBI:58027 glyceraldehyde 3-phosphate(2-)	[GO:0051408 glyceraldehyde 3-phosphate:inorganic phosphate antiporter activity]
CHEBI:35584 purine	[GO:0015390 purine-specific nucleoside:sodium symporter activity]
CHEBI:57586 biotinate	[GO:0015225 biotin transmembrane transporter activity]
CHEBI:16454 pantothenate	[GO:0015498 pantothenate:sodium symporter activity]
CHEBI:60272 aluminium ion	[GO:0015083 aluminum ion transmembrane transporter activity]
CHEBI:16189 sulfate	[GO:0015381 high-affinity sulfate transmembrane transporter activity, GO:0008271 secondary active sulfate transmembrane transporter activity, GO:0008512 sulfate:proton symporter activity, GO:0005340 nucleotide-sulfate transmembrane transporter activity, GO:0015419 ABC-type sulfate transporter activity, GO:0009676 low-affinity sulfate:proton symporter activity, GO:0009675 high-affinity sulfate:proton symporter activity, GO:0015382 sodium:sulfate symporter activity, GO:0015116 sulfate transmembrane transporter activity, GO:0015383 sulfate:bicarbonate antiporter activity]
CHEBI:37684 mannose	[GO:0055053 mannose:proton symporter activity]
CHEBI:57305 glycine zwitterion	[GO:0015375 glycine:sodium symporter activity]
CHEBI:17562 cytidine	[GO:0015212 cytidine transmembrane transporter activity]
CHEBI:15361 pyruvate	[GO:0005477 pyruvate secondary active transmembrane transporter activity]
CHEBI:15710 betaine aldehyde	[GO:0031458 ABC-type betaine transporter activity]
CHEBI:16704 uridine	[GO:0015394 uridine:proton symporter activity]
CHEBI:15889 sterol	[GO:0034041 ABC-type sterol transporter activity]
CHEBI:30623 oxalate(2-)	[GO:0019531 oxalate transmembrane transporter activity]
CHEBI:17310 pyridoxal	[GO:0031925 pyridoxal transmembrane transporter activity]
CHEBI:50562 tartrate salt	[GO:0015516 tartrate:succinate antiporter activity]

Continued on next page

Table 42: C2GO map with ChEBI terms and their corresponding GO terms

ChEBI term	GO terms
CHEBI:35491 L-cystine zwitterion	[GO:0015184 L-cystine transmembrane transporter activity]
CHEBI:50383 nicotinamide mononucleotide	[GO:0015663 nicotinamide mononucleotide transmembrane transporter activity]
CHEBI:62533 oligogalacturonide	[GO:0033154 ABC-type oligogalacturonide transporter activity]
CHEBI:17234 glucose	[GO:0015304 glucose uniporter activity]
CHEBI:16301 nitrite	[GO:0015514 nitrite efflux transmembrane transporter activity, GO:0015113 nitrite transmembrane transporter activity]
CHEBI:48999 4-hydroxyphenylacetate	[GO:1901241 4-hydroxyphenylacetate transmembrane transporter activity]
CHEBI:60584 bicozamycin	[GO:0015545 bicyclomycin transmembrane transporter activity]
CHEBI:32395 arachidonate	[GO:1903962 arachidonate transmembrane transporter activity]
CHEBI:65327 D-xylose	[GO:0015614 ABC-type D-xylose transporter activity]
CHEBI:16708 adenine	[GO:0015207 adenine transmembrane transporter activity]
CHEBI:58252 S-methyl-L-methionine zwitterion	[GO:0000100 S-methylmethionine transmembrane transporter activity]
CHEBI:57632 UDP-alpha-D-xylose(2-)	[GO:0005464 UDP-xylose transmembrane transporter activity]
CHEBI:30653 homoserine	[GO:0042970 homoserine transmembrane transporter activity]
CHEBI:12931 D-galactonate	[GO:0042881 D-galactonate transmembrane transporter activity]
CHEBI:60979 alpha-glucoside	[GO:0005352 alpha-glucoside:proton symporter activity]
CHEBI:66916 alanine zwitterion	[GO:0070906 aspartate:alanine antiporter activity, GO:0015655 alanine:sodium symporter activity]
CHEBI:29990 D-aspartate(1-)	[GO:0140010 D-aspartate transmembrane transporter activity]
CHEBI:17536 allantoate	[GO:0015124 allantoate transmembrane transporter activity]
CHEBI:26333 prostaglandin	[GO:0015132 prostaglandin transmembrane transporter activity]
CHEBI:507393 taurine zwitterion	[GO:0005369 taurine:sodium symporter activity, GO:0015411 ABC-type taurine transporter transporter activity]
CHEBI:60039 L-proline zwitterion	[GO:0015193 L-proline transmembrane transporter activity]
CHEBI:57947 creatine zwitterion	[GO:0005309 creatine:sodium symporter activity]
CHEBI:51349 polyamine macromolecule	[GO:0015312 polyamine:proton antiporter activity]
CHEBI:16480 nitric oxide	[GO:0030184 nitric oxide transmembrane transporter activity]
CHEBI:39128 magnesium ion	[GO:0061768 magnesium:sodium antiporter activity]
CHEBI:46819 urate salt	[GO:0015143 urate transmembrane transporter activity]
CHEBI:15927 N-ribosylnicotinamide	[GO:0034257 nicotinamide riboside transmembrane transporter activity]
CHEBI:25676 oligopeptide	[GO:0015421 ABC-type oligopeptide transporter activity]
CHEBI:28498 acadesine	[GO:1903089 5-amino-1-ribofuranosylimidazole-4-carboxamide transmembrane transporter activity]
CHEBI:26707 glycerol phosphate	[GO:0015527 glycerol-phosphate:inorganic phosphate antiporter activity]
CHEBI:58070 dehydroascorbide(1-)	[GO:0033300 dehydroascorbic acid transmembrane transporter activity]
CHEBI:57427 L-leucine zwitterion	[GO:0015190 L-leucine transmembrane transporter activity]
CHEBI:77805 enterobactin(1-)	[GO:0042931 enterobactin transmembrane transporter activity]
CHEBI:17754 glycerol	[GO:0015254 glycerol channel activity]
CHEBI:35247 D-serine zwitterion	[GO:0042945 D-serine transmembrane transporter activity]
CHEBI:16199 urea	[GO:0015265 urea channel activity, GO:0033221 ATPase-coupled urea transmembrane transporter activity, GO:0015401 urea:sodium symporter activity]
CHEBI:17393 D-allose	[GO:0015615 D-allose-importing ATPase activity]
CHEBI:24337 glutathione derivative	[GO:0015431 ABC-type glutathione S-conjugate transporter activity]
CHEBI:25516 nickel cation	[GO:0015413 ABC-type nickel transporter activity]
CHEBI:58315 L-tyrosine zwitterion	[GO:0005302 L-tyrosine transmembrane transporter activity]
CHEBI:29032 (R)-pantothenate	[GO:0015233 pantothenate transmembrane transporter activity]
CHEBI:16634 raffinose	[GO:0015529 raffinose:proton symporter activity]
CHEBI:62501 folate(2-)	[GO:0140211 folic acid:proton symporter activity, GO:0008518 folate:anion antiporter activity]
CHEBI:35780 phosphate ion	[GO:0061513 glucose 6-phosphate:inorganic phosphate antiporter activity, GO:0015415 ATPase-coupled phosphate ion transmembrane transporter activity, GO:0015364 dicarboxylate:inorganic phosphate antiporter activity, GO:0051407 glycerone phosphate:inorganic phosphate antiporter activity, GO:0015121 phosphoenolpyruvate:phosphate antiporter activity, GO:0051408 glyceraldehyde 3-phosphate:inorganic phosphate antiporter activity, GO:0015527 glycerol-phosphate:inorganic phosphate antiporter activity]
CHEBI:57834 spermidine(3+)	[GO:0015606 spermidine transmembrane transporter activity]
CHEBI:15075 selenate	[GO:0140413 zinc:bicarbonate:selenite symporter activity]
CHEBI:28163 iron(III) hydroxamate	[GO:0015625 ABC-type ferric hydroxamate transporter activity]
CHEBI:17051 fluoride	[GO:0062054 fluoride channel activity, GO:1903425 fluoride transmembrane transporter activity]
CHEBI:24287 glucosylglycerol	[GO:0051474 glucosylglycerol transmembrane transporter activity]
CHEBI:15318 xanthine	[GO:0042907 xanthine transmembrane transporter activity]

Continued on next page

Table 42: C2GO map with ChEBI terms and their corresponding GO terms

ChEBI term	GO terms
CHEBI:29108 calcium(2+)	[GO:0140417 ATP-sensitive calcium-release channel activity, GO:1905056 P-type calcium transporter activity involved in regulation of presynaptic cytosolic calcium ion concentration, GO:0098697 ryanodine-sensitive calcium-release channel activity involved in regulation of postsynaptic cytosolic calcium levels, GO:1905055 calcium:cation antiporter activity involved in regulation of presynaptic cytosolic calcium ion concentration, GO:1905059 P-type calcium transporter activity involved in regulation of postsynaptic cytosolic calcium ion concentration, GO:0008331 high voltage-gated calcium channel activity, GO:0086058 voltage-gated calcium channel activity involved in Purkinje myocyte cell action potential, GO:0008273 calcium, potassium:sodium antiporter activity, GO:0097365 stretch-activated, cation-selective, calcium channel activity involved in regulation of cardiac muscle cell action potential, GO:0099604 ligand-gated calcium channel activity, GO:0008332 low voltage-gated calcium channel activity]
CHEBI:327995 tyraminium	[GO:0070908 tyrosine:tyramine antiporter activity]
CHEBI:46912 ornithinium(1+)	[GO:0043858 arginine:ornithine antiporter activity]
CHEBI:65340 5-formyltetrahydrofolate(2-)	[GO:0015231 5-formyltetrahydrofolate transmembrane transporter activity]
CHEBI:58048 L-asparagine zwitterion	[GO:0015182 L-asparagine transmembrane transporter activity]
CHEBI:46502 tungstate	[GO:1901238 ABC-type tungstate transporter activity]
CHEBI:46911 L-ornithinium(1+)	[GO:0015496 putrescine:ornithine antiporter activity]
CHEBI:57705 UDP-N-acetyl-alpha-D-glucosamine(2-)	[GO:0005462 UDP-N-acetylglucosamine transmembrane transporter activity]
CHEBI:62071 2-cis-abscisate	[GO:0090440 abscisic acid transmembrane transporter activity]
CHEBI:30049 teichoic acid	[GO:0015438 ABC-type teichoic acid transporter activity]
CHEBI:17796 L-idonate	[GO:0015568 L-idonate transmembrane transporter activity]
CHEBI:15595 malate(2-)	[GO:0015366 malate:proton symporter activity, GO:0043882 malate:sodium symporter activity, GO:0015367 oxoglutarate:malate antiporter activity]
CHEBI:32544 nicotinate	[GO:0090416 nicotinate transmembrane transporter activity]
CHEBI:57416 D-alanine zwitterion	[GO:0042944 D-alanine transmembrane transporter activity]
CHEBI:50681 methotrexate(2-)	[GO:0015350 methotrexate transmembrane transporter activity]
CHEBI:32696 argininium(1+)	[GO:0043862 arginine:agmatine antiporter activity, GO:0043858 arginine:ornithine antiporter activity]
CHEBI:57844 L-methionine zwitterion	[GO:0000102 L-methionine secondary active transmembrane transporter activity]
CHEBI:16244 4-(trimethylammonio)butanoate	[GO:0044667 (R)-carnitine:4-(trimethylammonio)butanoate antiporter activity]
CHEBI:16947 citrate(3-)	[GO:0015515 citrate:succinate antiporter activity, GO:0015531 citrate:proton symporter activity]
CHEBI:58937 thiamine(1+) diphosphate(3-)	[GO:0090422 thiamine pyrophosphate transmembrane transporter activity]
CHEBI:2639 amiloride	[GO:0015308 amiloride:proton antiporter activity]
CHEBI:16382 iodide	[GO:0008507 sodium:iodide symporter activity, GO:0015111 iodide transmembrane transporter activity]
CHEBI:36208 shikimate	[GO:0015533 shikimate:proton symporter activity]
CHEBI:49709 chlorate	[GO:0015107 chlorate transmembrane transporter activity]
CHEBI:24250 gibberellin	[GO:1905201 gibberellin transmembrane transporter activity]
CHEBI:10983 (R)-3-hydroxybutyrate	[GO:0097253 beta-hydroxybutyrate transmembrane transporter activity]
CHEBI:26271 proline	[GO:0005298 proline:sodium symporter activity, GO:0005297 proline:proton symporter activity]
CHEBI:64608 GDP-fucose	[GO:0005457 GDP-fucose transmembrane transporter activity]
CHEBI:32564 lysinium(1+)	[GO:0043872 lysine:cadaverine antiporter activity, GO:0015493 lysine:proton symporter activity]
CHEBI:22590 arabiNaN	[GO:0042901 arabiNaN transmembrane transporter activity]
CHEBI:57642 glycerone phosphate(2-)	[GO:0051407 glycerone phosphate:inorganic phosphate antiporter activity]
CHEBI:22605 arabinitol	[GO:0015167 arabitol transmembrane transporter activity]
CHEBI:29195 cyanate	[GO:0015541 secondary active cyanate transmembrane transporter activity]
CHEBI:26441 pyrimidine nucleotide	[GO:0015218 pyrimidine nucleotide transmembrane transporter activity]
CHEBI:16040 cytosine	[GO:0015504 cytosine:proton symporter activity]
CHEBI:18123 N-methylnicotinate	[GO:0090417 N-methylnicotinate transmembrane transporter activity]
CHEBI:60903 N-(4-aminobenzoyl)-L-glutamate	[GO:0015558 secondary active p-aminobenzoyl-L-glutamate transmembrane transporter activity]
CHEBI:30477 tellurite	[GO:0015654 tellurite transmembrane transporter activity]
CHEBI:24265 gluconate	[GO:0015128 gluconate transmembrane transporter activity]
CHEBI:58339 3'-phosphonato-5'-adenylyl sulfate(4-)	[GO:0046964 3'-phosphoadenosine 5'-phosphosulfate transmembrane transporter activity]
CHEBI:15740 formate	[GO:0015660 formate efflux transmembrane transporter activity]
CHEBI:16526 carbon dioxide	[GO:0035379 carbon dioxide transmembrane transporter activity]

Continued on next page

Table 42: C2GO map with ChEBI terms and their corresponding GO terms

ChEBI term	GO terms
CHEBI:456215 adenosine 5'-monophosphate(2-)	[GO:0080122 AMP transmembrane transporter activity]
CHEBI:24636 proton	[GO:0008512 sulfate:proton symporter activity, GO:0005360 insulin-responsive glucose:proton symporter activity, GO:0005278 acetylcholine:proton antiporter activity, GO:0005358 high-affinity glucose:proton symporter activity, GO:0009672 auxin:proton symporter activity, GO:0009678 pyrophosphate hydrolysis-driven proton transmembrane transporter activity, GO:0097079 selenite:proton symporter activity, GO:0015345 ferric enterobactin:proton symporter activity, GO:0015311 monoamine:proton antiporter activity, GO:0015252 proton channel activity, GO:0090448 glucosinolate:proton symporter activity, GO:0009679 hexose:proton symporter activity, GO:0015537 xanthosine:proton symporter activity, GO:0015520 tetracycline:proton antiporter activity, GO:0140211 folic acid:proton symporter activity, GO:0015295 solute:proton symporter activity, GO:0003957 NAD(P)+ transhydrogenase (B-specific) activity, GO:0030171 voltage-gated proton channel activity, GO:0015366 malate:proton symporter activity]
CHEBI:30612 D-glucarate(2-)	[GO:0042878 D-glucarate transmembrane transporter activity]
CHEBI:71501 c-GMP-AMP(2-)	[GO:0140360 cyclic-GMP-AMP transmembrane transporter activity]
CHEBI:24279 glucosinolate	[GO:0090448 glucosinolate:proton symporter activity]
CHEBI:58045 L-isoleucine zwitterion	[GO:0015188 L-isoleucine transmembrane transporter activity]
CHEBI:35692 dicarboxylic acid	[GO:0015364 dicarboxylate:inorganic phosphate antiporter activity]
CHEBI:35404 chromate(2-)	[GO:0015109 chromate transmembrane transporter activity]
CHEBI:77932 tetracycline zwitterion	[GO:0015520 tetracycline:proton antiporter activity]
CHEBI:57986 riboflavin(1-)	[GO:0032217 riboflavin transmembrane transporter activity]
CHEBI:26020 phosphate	[GO:0015317 phosphate:proton symporter activity]
CHEBI:57966 beta-alanine zwitterion	[GO:0001761 beta-alanine transmembrane transporter activity]
CHEBI:15355 acetylcholine	[GO:0005278 acetylcholine:proton antiporter activity]
CHEBI:61346 achromobactin	[GO:0042934 achromobactin transmembrane transporter activity]
CHEBI:67062 lipo-chitin oligosaccharide	[GO:1901514 ATPase-coupled lipo-chitin oligosaccharide transmembrane transporter activity]
CHEBI:57925 glutathionate(1-)	[GO:0034634 glutathione transmembrane transporter activity]
CHEBI:131429 carcinine	[GO:1905131 carcinine transmembrane transporter activity]
CHEBI:28675 dextrin	[GO:0042957 dextrin transmembrane transporter activity]
CHEBI:15373 biopterin	[GO:0015224 biopterin transmembrane transporter activity]
CHEBI:16670 peptide	[GO:0015333 peptide:proton symporter activity]
CHEBI:22563 anion	[GO:0008518 folate:anion antiporter activity]
CHEBI:3015 benomyl	[GO:0015310 benomyl:proton antiporter activity]
CHEBI:60253 silver cation	[GO:0015445 P-type silver transporter activity]
CHEBI:48081 bacteriocin	[GO:0043214 ABC-type bacteriocin transporter activity]
CHEBI:24591 hexuronate	[GO:0015539 hexuronate:cation symporter activity]
CHEBI:27641 cycloheximide	[GO:0015309 cycloheximide:proton antiporter activity]
CHEBI:17359 sulfite	[GO:0000319 sulfite transmembrane transporter activity]
CHEBI:59338 methylammonium	[GO:0015264 methylammonium channel activity]
CHEBI:33709 amino acid	[GO:0017032 amino acid:potassium symporter activity]
CHEBI:17306 maltose	[GO:0015423 ABC-type maltose transporter activity, GO:0005364 maltose:proton symporter activity]
CHEBI:22629 arsenate ion	[GO:0015446 ATPase-coupled arsenite transmembrane transporter activity]
CHEBI:21168 GDP-mannose	[GO:0005458 GDP-mannose transmembrane transporter activity]
CHEBI:16235 guanine	[GO:0008558 ABC-type guanine transporter activity]
CHEBI:29103 potassium(1+)	[GO:0060072 large conductance calcium-activated potassium channel activity, GO:0005391 P-type sodium:potassium-exchanging transporter activity, GO:0022894 Intermediate conductance calcium-activated potassium channel activity, GO:0005249 voltage-gated potassium channel activity, GO:0008273 calcium, potassium:sodium antiporter activity, GO:0070089 chloride-activated potassium channel activity, GO:0015272 ATP-activated inward rectifier potassium channel activity, GO:0015387 potassium:proton symporter activity, GO:0086008 voltage-gated potassium channel activity involved in cardiac muscle cell action potential repolarization, GO:0086090 voltage-gated potassium channel activity involved in SA node cell action potential repolarization, GO:0015467 G-protein activated inward rectifier potassium channel activity, GO:0005267 potassium channel activity, GO:0086088 voltage-gated potassium channel activity involved in Purkinje myocyte action potential repolarization, GO:0009674 potassium:sodium

Continued on next page

Table 42: C2GO map with ChEBI terms and their corresponding GO terms

ChEBI term	GO terms
CHEBI:29033 iron(2+)	[GO:0015093 ferrous iron transmembrane transporter activity, GO:0061840 high-affinity ferrous iron transmembrane transporter activity]
CHEBI:17814 salicin	[GO:0042950 salicin transmembrane transporter activity]
CHEBI:58243 5'-adenylyl sulfate(2-)	[GO:1902557 5'-adenylyl sulfate transmembrane transporter activity]
CHEBI:25706 organomercury compound	[GO:0015548 organomercurial transmembrane transporter activity]
CHEBI:60249 lead ion	[GO:0015094 lead ion transmembrane transporter activity]
CHEBI:60836 phytochelatin	[GO:0044604 ABC-type phytochelatin transporter activity]
CHEBI:27082 trehalose	[GO:0044693 trehalose:proton symporter activity]
CHEBI:443725 fosmidomycin	[GO:0042898 fosmidomycin transmembrane transporter activity]
CHEBI:57597 sn-glycerol 3-phosphate(2-)	[GO:0015430 ABC-type glycerol-3-phosphate transporter activity]
CHEBI:29805 glycolate	[GO:0043879 glycolate transmembrane transporter activity]
CHEBI:37554 fatty acyl-CoA	[GO:0015607 ABC-type fatty-acyl-CoA transporter activity]
CHEBI:57746 3',5'-cyclic GMP(1-)	[GO:1905948 ABC-type 3',5'-cyclic GMP transmembrane transporter activity]
CHEBI:36080 protein	[GO:0008564 protein-exporting ATPase activity]
CHEBI:17992 sucrose	[GO:0008506 sucrose:proton symporter activity]
CHEBI:45725 spermine(4+)	[GO:0000297 spermine transmembrane transporter activity]
CHEBI:18305 hydroquinone O-beta-D-glucopyranoside	[GO:0042951 arbutin transmembrane transporter activity]
CHEBI:17716 lactose	[GO:0015528 lactose:proton symporter activity]
CHEBI:16150 benzoate	[GO:0042925 benzoate transmembrane transporter activity]
CHEBI:18212 selenite(2-)	[GO:0097079 selenite:proton symporter activity]
CHEBI:36420 acridine	[GO:0042962 acridine:proton antiporter activity, GO:0042911 acridine transmembrane transporter activity]
CHEBI:15904 long-chain fatty acid	[GO:0015483 long-chain fatty acid transporting porin activity]
CHEBI:9337 sulfathiazole	[GO:0015546 sulfathiazole transmembrane transporter activity]
CHEBI:17268 myo-inositol	[GO:0005367 myo-inositol:sodium symporter activity, GO:0005366 myo-inositol:proton symporter activity]
CHEBI:57288 acetyl-CoA(4-)	[GO:0015325 acetyl-CoA:CoA antiporter activity]
CHEBI:30408 iron-sulfur cluster	[GO:0140481 ABC-type iron-sulfur cluster transporter activity]
CHEBI:15377 water	[GO:0015250 water channel activity]
CHEBI:16709 pyridoxine	[GO:0031928 pyridoxine transmembrane transporter activity]
CHEBI:85413 5-hydroxyectoine zwitterion	[GO:0033288 ATPase-coupled hydroxyectoine transmembrane transporter activity]
CHEBI:18107 xanthosine	[GO:0015537 xanthosine:proton symporter activity]
CHEBI:14314 D-glucose 6-phosphate	[GO:0061513 glucose 6-phosphate:inorganic phosphate antiporter activity]
CHEBI:58515 ectoine zwitterion	[GO:0033286 ATPase-coupled ectoine transmembrane transporter activity]
CHEBI:58384 cadaverine(2+)	[GO:0043872 lysine:cadaverine antiporter activity]
CHEBI:16991 deoxyribonucleic acid	[GO:0051035 DNA transmembrane transporter activity]
CHEBI:28260 galactose	[GO:0015371 galactose:sodium symporter activity, GO:0015517 galactose:proton symporter activity]
CHEBI:17632 nitrate	[GO:0010542 nitrate efflux transmembrane transporter activity, GO:0015112 nitrate transmembrane transporter activity, GO:0015414 ABC-type nitrate transporter activity, GO:0009671 nitrate:proton symporter activity, GO:0080054 low-affinity nitrate transmembrane transporter activity, GO:0015513 high-affinity secondary active nitrite transmembrane transporter activity]
CHEBI:48030 tetrapeptide	[GO:1901584 tetrapeptide transmembrane transporter activity]
CHEBI:61310 chromium ion	[GO:0070835 chromium ion transmembrane transporter activity]
CHEBI:49870 antimonous acid	[GO:0042961 ATPase-coupled antimonite transmembrane transporter activity]
CHEBI:58145 agmatinium(2+)	[GO:0043862 arginine:agmatine antiporter activity, GO:0043861 agmatine:putrescine antiporter activity]
CHEBI:16412 lipopolysaccharide	[GO:0015221 lipopolysaccharide transmembrane transporter activity]
CHEBI:3259 CCCP	[GO:0015549 carbonyl cyanide m-chlorophenylhydrazone transmembrane transporter activity]
CHEBI:383703 3,6-diamino-10-methylacridinium chloride	[GO:0015566 acriflavine transmembrane transporter activity]
CHEBI:30849 L-arabinose	[GO:0015612 ABC-type L-arabinose transporter activity]
CHEBI:17126 carnitine	[GO:0005476 carnitine:acyl carnitine antiporter activity]
CHEBI:28044 phenylalanine	[GO:0015492 phenylalanine:proton symporter activity]
CHEBI:24040 flavin adenine dinucleotide	[GO:0015230 FAD transmembrane transporter activity]

Continued on next page

Table 42: C2GO map with ChEBI terms and their corresponding GO terms

ChEBI term	GO terms
CHEBI:29101 sodium(1+)	[GO:0015370 solute:sodium symporter activity, GO:0005415 nucleoside:sodium symporter activity, GO:0015361 low-affinity sodium:dicarboxylate symporter activity, GO:0086061 voltage-gated sodium channel activity involved in bundle of His cell action potential, GO:0022818 sodium ion uniporter activity, GO:0015081 sodium ion transmembrane transporter activity, GO:0005248 voltage-gated sodium channel activity, GO:0008520 L-ascorbate:sodium symporter activity, GO:0005391 P-type sodium:potassium-exchanging transporter activity,GO:0005283 amino acid:sodium symporter activity, GO:0015371 galactose:sodium symporter activity, GO:0005436 sodium:phosphate symporter activity, GO:0005328 neurotransmitter:sodium symporter activity, GO:0005307 choline:sodium symporter activity, GO:0015362 high-affinity sodium:dicarboxylate symporter activity, GO:0008273 calcium, potassium:sodium antiporter activity, GO:0008507 sodium:iodide symporter activity, GO:0086040 sodium:proton antiporter activity involved i
CHEBI:33984 fucose	[GO:0015535 fucose:proton symporter activity]
CHEBI:61304 phosphoglycerate	[GO:0015120 phosphoglycerate transmembrane transporter activity]
CHEBI:40036 amitrole	[GO:0015314 aminotriazole:proton antiporter activity]
CHEBI:57932 D-methionine zwitterion	[GO:0033232 ABC-type D-methionine transporter activity]
CHEBI:49552 copper(1+)	[GO:0140581 P-type monovalent copper transporter activity]
CHEBI:59905 dopaminium(1+)	[GO:0005330 dopamine:sodium symporter activity]
CHEBI:29806 fumarate(2-)	[GO:0005469 succinate:fumarate antiporter activity, GO:0062057 L-aspartate:fumarate antiporter activity]
CHEBI:26675 silicic acid	[GO:0015115 silicate transmembrane transporter activity]
CHEBI:29242 arsenite(1-)	[GO:0015105 arsenite transmembrane transporter activity, GO:0008490 arsenite secondary active transmembrane transporter activity]
CHEBI:57990 2-dehydro-3-deoxy-D-gluconate	[GO:0015649 2-keto-3-deoxygluconate:proton symporter activity]
CHEBI:58702 phosphoenolpyruvate	[GO:0015121 phosphoenolpyruvate:phosphate antiporter activity]
CHEBI:57418 (2-aminoethyl)phosphonic acid zwitterion	[GO:0033225 ATPase-coupled 2-aminoethylphosphonate transporter activity]
CHEBI:22599 arabinose	[GO:0015518 arabinose:proton symporter activity]
CHEBI:61496 alpha-D-	[GO:0015436 ABC-type capsular-polysaccharide transporter activity]
CHEBI:16810 2-oxoglutarate(2-)	[GO:0015367 oxoglutarate:malate antiporter activity, GO:0015532 alpha-ketoglutarate:proton symporter activity]
CHEBI:30031 succinate(2-)	[GO:0097434 succinate:proton symporter activity, GO:0015515 citrate:succinate antiporter activity, GO:0015516 tartrate:succinate antiporter activity, GO:0005469 succinate:fumarate antiporter activity]
CHEBI:64816 doxorubicin(1+)	[GO:1901242 ABC-type doxorubicin transporter activity]
CHEBI:143624 D-galactofuranose	[GO:0103116 ABC-type D-galactofuranose transporter]
CHEBI:17387 O-acylcarnitine	[GO:0005476 carnitine:acyl carnitine antiporter activity]
CHEBI:3098 bile acid	[GO:0008508 bile acid:sodium symporter activity, GO:0015432 ABC-type bile acid transporter activity]
CHEBI:50211 retinol	[GO:0034632 retinol transmembrane transporter activity]
CHEBI:597326 pyridoxal 5'-phosphate(2-)	[GO:0031926 pyridoxal phosphate transmembrane transporter activity]
CHEBI:65228 3-hydroxyphenyl propanoate	[GO:0015540 3-hydroxyphenyl propionate:proton symporter activity]
CHEBI:30089 acetate	[GO:0015360 acetate:proton symporter activity]
CHEBI:39127 magnesium cation	[GO:0015444 P-type magnesium transporter activity]
CHEBI:356416 5-aminolevulinic acid	[GO:0140485 5-aminolevulinic acid transmembrane transporter activity]
CHEBI:33542 trioxidosulfanidosulfate(1-)	[GO:0102025 ABC-type thiosulfate transporter activity]
CHEBI:48871 galactaric acid anion	[GO:1902301 galactarate transmembrane transporter activity]
CHEBI:29105 zinc(2+)	[GO:0015341 zinc efflux active transmembrane transporter activity, GO:0005385 zinc ion transmembrane transporter activity, GO:0000006 high-affinity zinc transmembrane transporter activity, GO:0000007 low-affinity zinc ion transmembrane transporter activity, GO:0022883 zinc efflux transmembrane transporter activity]
CHEBI:28938 ammonium	[GO:0015400 low-affinity secondary active ammonium transmembrane transporter activity, GO:0008519 ammonium transmembrane transporter activity, GO:0015398 high-affinity secondary active ammonium transmembrane transporter activity]
CHEBI:15676 allantoin	[GO:0005274 allantoin:proton symporter activity]
CHEBI:62070 nalidixic acid anion	[GO:0015547 nalidixic acid transmembrane transporter activity]

Continued on next page

Table 42: C2GO map with ChEBI terms and their corresponding GO terms

ChEBI term	GO terms
CHEBI:36915 inorganic cation	[GO:0015539]hexuronate:cation symporter activity, GO:0097604 temperature-gated cation channel activity, GO:0022850 serotonin-gated cation-selective channel activity, GO:0015306 sialate:cation symporter activity, GO:0022848 acetylcholine-gated cation-selective channel activity, GO:0015505 uracil:cation symporter activity]
CHEBI:77847 glycerol 1-phosphodiester(1-)	[GO:0001406]glycerophosphodiester transmembrane transporter activity]
CHEBI:62436 (+)-abscisic acid D-glucopyranosyl ester	[GO:1902417 (+)-abscisic acid D-glucopyranosyl ester transmembrane transporter activity]
CHEBI:58189 GDP(3-)	[GO:0010292 GTP:GDP antiporter activity]
CHEBI:59888 gamma-aminobutyric acid zwitterion	[GO:0005332 gamma-aminobutyric acid:sodium symporter activity, GO:0070909 glutamate:gamma-aminobutyric acid antiporter activity, GO:0015495 gamma-aminobutyric acid:proton symporter activity]
CHEBI:32551 L-lysinium(1+)	[GO:0015661 L-lysine efflux transmembrane transporter activity]
CHEBI:16452 oxaloacetate(2-)	[GO:0000227 oxaloacetate secondary active transmembrane transporter activity]
CHEBI:15936 aldehydo-D-xylose	[GO:0015519 D-xylose:proton symporter activity]
CHEBI:25350 mevalonate	[GO:0015130 mevalonate transmembrane transporter activity]
CHEBI:61082 lipid-linked peptidoglycan	[GO:0015648 lipid-linked peptidoglycan transporter activity]
CHEBI:58210 FMN(3-)	[GO:0044610 FMN transmembrane transporter activity]
CHEBI:28757 fructose	[GO:0015284 fructose uniporter activity, GO:0055054 fructose:proton symporter activity]
CHEBI:35492 cystine zwitterion	[GO:0015327 cystine:glutamate antiporter activity]
CHEBI:58367 UDP-D-glucose(2-)	[GO:0005460 UDP-glucose transmembrane transporter activity]
CHEBI:17540 methyl beta-D-galactoside	[GO:0015592 methylgalactoside transmembrane transporter activity]
CHEBI:15748 D-glucuronate	[GO:0042880 D-glucuronate transmembrane transporter activity]
CHEBI:51057 3-phenylpropionate	[GO:0042890 3-phenylpropionic acid transmembrane transporter activity]
CHEBI:61345 chrysobactin	[GO:0042933 chrysobactin transmembrane transporter activity]

The table presents the mapping from each specific substrate from ChEBI ontology along with the molecular function GO term in Gene Ontology related to transporting the specific substrate.

Appendix B

Dataset Information

B.1 SPEC dataset series

This chapter presents the four extracted datasets for the SPEC project, with protein sequences as samples and the transported substrate as their label. The protein data has been collected from the UniProt database on 2022.10.03.

Table 43: Details of substrates' sample count in four extracted datasets

Dataset	CHEBI	Substrate	SP-SPEC-60	UP-SPEC-60	SP-SPEC-100	UP-SPEC-100
0	CHEBI:24636	proton	443	563	813	1,226
1	CHEBI:29108	calcium(2+)	153	209	347	865
2	CHEBI:29103	potassium(1+)	146	193	299	726
3	CHEBI:17996	chloride	70	124	165	442
4	CHEBI:29101	sodium(1+)	69	83	148	266
5	CHEBI:29105	zinc(2+)	47	51	67	80
6	CHEBI:16189	sulfate	17	33	20	67
7	CHEBI:28938	ammonium	21	26	32	52
8	CHEBI:15377	water	8	9	38	43
9	CHEBI:57986	riboflavin(1-)	22	23	29	37
10	CHEBI:36080	protein	19	18	28	28
11	CHEBI:24337	glutathione derivative	5	7	13	27
12	CHEBI:50211	retinol	4	8	15	27
13	CHEBI:17632	nitrate	17	17	25	26
14	CHEBI:29033	iron(2+)	13	16	16	24
15	CHEBI:15361	pyruvate	8	9	13	20
16	CHEBI:30616	ATP(4-)	2	5	9	18
17	CHEBI:49552	copper(1+)	6	9	7	17
18	CHEBI:16199	urea	4	4	9	16
19	CHEBI:58070	dehydroascorbide(1-)	4	4	13	14
20	CHEBI:15354	choline	6	6	9	13
21	CHEBI:16113	cholesterol	4	3	9	12
22	CHEBI:46911	L-ornithinium(1+)	4	7	7	12
23	CHEBI:35491	L-cystine zwitterion	8	7	11	11
24	CHEBI:61082	lipid-linked peptidoglycan	10	10	11	11
25	CHEBI:3098	bile acid	1	1	4	10
26	CHEBI:29985	L-glutamate(1-)	3	3	8	10
27	CHEBI:16947	citrate(3-)	4	3	8	10
28	CHEBI:30823	oleate	2	3	7	9
29	CHEBI:59789	S-adenosyl-L-methionine zwitterion	3	7	4	9
30	CHEBI:30411	cobalamin	8	8	9	9

Continued on next page

Table 43: Details of substrates' sample count in four extracted datasets

Dataset	CHEBI	Substrate	SP-SPEC-60	UP-SPEC-60	SP-SPEC-100	UP-SPEC-100
31	CHEBI:30413	heme	3	4	6	9
32	CHEBI:58367	UDP-D-glucose(2-)	6	7	8	9
33	CHEBI:57925	glutathionate(1-)	3	3	8	8
34	CHEBI:58339	3'-phosphonato-5'-adenylyl sulfate(4-)	4	4	5	8
35	CHEBI:326268	1,4-butanediammonium	7	6	8	8
36	CHEBI:57705	UDP-N-acetyl-alpha-D-glucosamine(2-)	6	6	6	8
37	CHEBI:29036	copper(2+)	5	7	6	8
38	CHEBI:57586	biotinate	6	6	6	8
39	CHEBI:46819	urate salt	1	3	4	8
40	CHEBI:17992	sucrose	5	5	7	7
41	CHEBI:57288	acetyl-CoA(4-)	0	2	0	7
42	CHEBI:16810	2-oxoglutarate(2-)	3	4	4	7
43	CHEBI:29034	iron(3+)	5	5	7	7
44	CHEBI:17544	hydrogencarbonate	1	2	4	7
45	CHEBI:61304	phosphoglycerate	3	4	3	7
46	CHEBI:17306	maltose	6	5	7	7
47	CHEBI:32682	L-argininium(1+)	3	3	6	6
48	CHEBI:30408	iron-sulfur cluster	1	0	4	6
49	CHEBI:60039	L-proline zwitterion	3	2	6	6
50	CHEBI:17051	fluoride	3	5	4	6
51	CHEBI:13389	NAD	3	3	6	6
52	CHEBI:15740	formate	4	4	4	6
53	CHEBI:15355	acetylcholine	4	4	5	6
54	CHEBI:15927	N-ribosylnicotinamide	5	5	5	5
55	CHEBI:59888	gamma-aminobutyric acid zwitterion	4	4	5	5
56	CHEBI:25197	mercury cation	4	4	5	5
57	CHEBI:37554	fatty acyl-CoA	2	1	4	5
58	CHEBI:63063	cadmium cation	2	2	5	5
59	CHEBI:26441	pyrimidine nucleotide	2	4	3	5
60	CHEBI:64608	GDP-fucose	3	3	3	5
61	CHEBI:35780	phosphate ion	5	5	5	5
62	CHEBI:38290	L-ascorbate	1	3	1	4
63	CHEBI:57932	D-methionine zwitterion	4	4	4	4
64	CHEBI:15904	long-chain fatty acid	3	4	3	4
65	CHEBI:25140	maltodextrin	2	2	4	4
66	CHEBI:57844	L-methionine zwitterion	4	4	4	4
67	CHEBI:57595	L-histidine zwitterion	2	2	3	4
68	CHEBI:26333	prostaglandin	1	1	3	4
69	CHEBI:30849	L-arabinose	4	4	4	4
70	CHEBI:57305	glycine zwitterion	2	2	2	4
71	CHEBI:33118	boric acid	3	3	4	4
72	CHEBI:48081	bacteriocin	2	4	2	4
73	CHEBI:71501	c-GMP-AMP(2-)	1	1	3	3
74	CHEBI:24265	gluconate	2	2	3	3
75	CHEBI:17750	glycine betaine	2	2	3	3
76	CHEBI:60903	N-(4-aminobenzoyl)-L-glutamate	1	3	1	3
77	CHEBI:16412	lipopolysaccharide	1	2	1	3
78	CHEBI:57632	UDP-alpha-D-xylose(2-)	1	1	3	3
79	CHEBI:30049	teichoic acid	2	3	2	3
80	CHEBI:28793	beta-D-glucan	1	1	3	3
81	CHEBI:24040	flavin adenine dinucleotide	2	2	3	3
82	CHEBI:16235	guanine	2	2	3	3
83	CHEBI:62501	folate(2-)	0	1	2	3
84	CHEBI:22629	arsenate ion	3	3	3	3
85	CHEBI:36264	molybdate	3	3	3	3
86	CHEBI:58702	phosphoenolpyruvate	3	3	3	3
87	CHEBI:71063	N-retinylidene phosphatidylethanolamine	1	1	2	3
88	CHEBI:46502	tungstate	3	3	3	3
89	CHEBI:16704	uridine	2	2	3	3
90	CHEBI:39127	magnesium cation	2	2	3	3
91	CHEBI:57597	sn-glycerol 3-phosphate(2-)	2	1	3	3
92	CHEBI:57912	L-tryptophan zwitterion	2	2	3	3
93	CHEBI:15318	xanthine	3	3	3	3
94	CHEBI:57812	CMP-N-acetyl-beta-neuraminate(2-)	1	1	2	3
95	CHEBI:16347	(R)-carnitine	1	1	3	3
96	CHEBI:45725	spermine(4+)	1	1	2	2

Continued on next page

Table 43: Details of substrates’ sample count in four extracted datasets

Dataset	CHEBI	Substrate	SP-SPEC-60	UP-SPEC-60	SP-SPEC-100	UP-SPEC-100
97	CHEBI:77847	glycerol 1-phosphodiester(1-)	1	1	2	2
98	CHEBI:57947	creatine zwitterion	1	1	2	2
99	CHEBI:33384	L-serine zwitterion	1	1	2	2
100	CHEBI:17540	methyl beta-D-galactoside	1	1	2	2
101	CHEBI:77805	enterobactin(1-)	1	1	1	2
102	CHEBI:25155	manganese cation	2	1	2	2
103	CHEBI:58315	L-tyrosine zwitterion	2	2	2	2
104	CHEBI:46081	fluconazole	2	2	2	2
105	CHEBI:17536	allantoate	2	2	2	2
106	CHEBI:17568	uracil	2	2	2	2
107	CHEBI:30089	acetate	2	2	2	2
108	CHEBI:25516	nickel cation	2	2	2	2
109	CHEBI:29242	arsenite(1-)	2	2	2	2
110	CHEBI:32544	nicotinate	1	1	1	2
111	CHEBI:58937	thiamine(1+) diphosphate(3-)	2	2	2	2
112	CHEBI:58359	L-glutamine zwitterion	1	1	2	2
113	CHEBI:57977	bilirubin(2-)	2	2	2	2
114	CHEBI:24592	hexuronic acid	1	1	1	1
115	CHEBI:21168	GDP-mannose	1	1	1	1
116	CHEBI:57926	L-threonine zwitterion	1	1	1	1
117	CHEBI:62436	(+)-abscisic acid D-glucopyranosyl ester	1	1	1	1
118	CHEBI:25350	mevalonate	1	0	1	1
119	CHEBI:16301	nitrite	0	1	0	1
120	CHEBI:29991	L-aspartate(1-)	0	1	0	1
121	CHEBI:30031	succinate(2-)	0	1	0	1
122	CHEBI:26675	silicic acid	0	0	1	1
123	CHEBI:28498	acadesine	0	0	1	1
124	CHEBI:58252	S-methyl-L-methionine zwitterion	0	0	1	1
125	CHEBI:58145	agmatinium(2+)	0	0	1	1
126	CHEBI:16708	adenine	0	0	1	1
127	CHEBI:57427	L-leucine zwitterion	1	1	1	1
128	CHEBI:17562	cytidine	1	1	1	1
129	CHEBI:57287	coenzyme A(4-)	1	1	1	1
130	CHEBI:16150	benzoate	1	1	1	1
131	CHEBI:16040	cytosine	1	1	1	1
132	CHEBI:64627	microcin	1	1	1	1
133	CHEBI:67063	poly-beta-1,6-N-acetyl-D-glucosamine	1	1	1	1
134	CHEBI:62071	2-cis-abscisate	1	1	1	1
135	CHEBI:58095	L-phenylalanine zwitterion	1	1	1	1
136	CHEBI:60836	phytochelatin	1	1	1	1
137	CHEBI:60253	silver cation	1	1	1	1
138	CHEBI:17754	glycerol	1	1	1	1
139	CHEBI:28199	ferrienterobactin(3-)	1	1	1	1
140	CHEBI:16813	galactitol	1	1	1	1
141	CHEBI:30623	oxalate(2-)	1	1	1	1
142	CHEBI:143624	D-galactofuranose	1	1	1	1
143	CHEBI:37592	organic phosphonate	1	1	1	1
144	CHEBI:16988	D-ribose	1	1	1	1
145	CHEBI:17359	sulfite	1	1	1	1
146	CHEBI:29032	(R)-pantothenate	1	1	1	1
147	CHEBI:17387	O-acylcarnitine	1	1	1	1
148	CHEBI:77932	tetracycline zwitterion	0	0	1	1
149	Total	-	1,366	1,724	2,513	4,526

The detailed information of each ChEBI term, their specific substrate, and their number of samples in the four different datasets of SwissProt-SPEC-60, UniProt-SPEC-60, SwissProt-SPEC-100, and UniProt-SPEC-100 are presented in this table. The substrate terms are sorted based on their frequency on UniProt-SPEC-100 (last column).

Appendix C

TooT-SPEC

C.1 Introduction

This chapter presents the detailed results for each specific substrate class using the TooT-SPEC model explained on Section 3.7.3. First, the results for the hold-out test set have been presented in Table 45. Then Section C.3 presents the results for the 3-fold cross-validation experiments.

C.2 TooT-SPEC results on hold-out testset

Table 44: Detailed test results of TooT-SPEC model on UniProt-SPEC-100

#	Substrate	Train	Test	TP	FP	FN	TN	Acc.	Prec.	Rec.	F1.	MCC
0	proton	817	409	407	46	2	1030	0.968	0.898	0.995	0.944	0.924
1	calcium(2+)	576	289	276	21	13	1175	0.977	0.929	0.955	0.942	0.928
2	potassium(1+)	484	242	218	7	24	1236	0.979	0.969	0.901	0.934	0.922
3	chloride	295	147	135	5	12	1333	0.989	0.964	0.918	0.941	0.935
4	sodium(1+)	178	88	72	5	16	1392	0.99	0.94	0.82	0.87	0.87
5	zinc(2+)	53	27	26	1	1	1457	1.00	0.96	0.96	0.96	0.96
6	sulfate	44	23	19	0	4	1462	1.00	1.00	0.83	0.91	0.91
7	ammonium	35	17	17	2	0	1466	1.00	0.90	1.00	0.94	0.94
8	water	29	14	14	0	0	1471	1.00	1.00	1.00	1.00	1.00
9	riboflavin(1-)	24	13	11	0	2	1472	1.00	1.00	0.85	0.92	0.92
10	protein	18	10	7	0	3	1475	1.00	1.00	0.70	0.82	0.84
11	glutathione derivative	18	9	9	0	0	1476	1.0	1.0	1.0	1.0	1.0
12	retinol	18	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
13	nitrate	18	8	5	1	3	1476	1.0	0.8	0.6	0.7	0.7
14	iron(2+)	16	8	8	1	0	1476	1.0	0.9	1.0	0.9	0.9
15	pyruvate	14	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
16	ATP(4-)	12	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
17	copper(1+)	11	6	6	1	0	1478	1.0	0.9	1.0	0.9	0.9
18	urea	10	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
19	dehydroascorbide(1-)	9	5	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
20	choline	9	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
21	cholesterol	8	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
22	L-ornithinium(1+)	8	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
23	lipid-linked peptidoglycan	8	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
24	L-cystine zwitterion	8	3	2	1	1	1481	1.0	0.7	0.7	0.7	0.7
25	bile acid	6	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0

Continued on next page

Table 44: Detailed test results of TooT-SPEC model on UniProt-SPEC-100

#	Substrate	Train	Test	TP	FP	FN	TN	Acc.	Prec.	Rec.	F1.	MCC
26	L-glutamate(1-)	7	3	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
27	citrate(3-)	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
28	cobalamin	6	3	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
29	oleate	6	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
30	heme	6	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
31	S-adenosyl-L-methionine zwitterion	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
32	UDP-D-glucose(2-)	6	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
33	UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
34	1-4-butanediammonium	6	2	0	2	2	1481	1.0	0.0	0.0	0.0	0.0
35	copper(2+)	5	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
36	3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
37	glutathionate(1-)	6	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
38	urate salt	5	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
39	biotinate	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
40	sucrose	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
41	acetyl-CoA(4-)	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
42	2-oxoglutarate(2-)	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
43	maltose	5	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
44	phosphoglycerate	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
45	iron(3+)	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
46	hydrogencarbonate	4	3	2	3	1	1479	1.0	0.4	0.7	0.5	0.5
47	iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
48	L-proline zwitterion	4	2	0	1	2	1482	1.0	0.0	0.0	0.0	0.0
49	L-argininium(1+)	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
50	NAD	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
51	acetylcholine	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
52	fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
53	formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
54	pyrimidine nucleotide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
55	cadmium cation	3	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
56	GDP-fucose	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
57	fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
58	mercury cation	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
59	phosphate ion	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
60	N-ribosylnicotinamide	4	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
61	gamma-aminobutyric acid zwitterion	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
62	long-chain fatty acid	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
63	L-arabinose	2	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
64	bacteriocin	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
65	glycine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
66	L-histidine zwitterion	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
67	L-ascorbate	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
68	L-methionine zwitterion	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
69	boric acid	2	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
70	D-methionine zwitterion	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
71	prostaglandin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
72	maltodextrin	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
73	folate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
74	molybdate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
75	N-retinylidene phosphatidylethanolamine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
76	lipopolysaccharide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
77	beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
78	gluconate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
79	tungstate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
80	sn-glycerol 3-phosphate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
81	UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
82	magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
83	flavin adenine dinucleotide	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
84	(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
85	CMP-N-acetyl-beta-neuraminate(2-)	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
86	c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
87	L-tryptophan zwitterion	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
88	N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
89	phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
90	arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
91	teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0

Continued on next page

Table 44: Detailed test results of TooT-SPEC model on UniProt-SPEC-100

#	Substrate	Train	Test	TP	FP	FN	TN	Acc.	Prec.	Rec.	F1.	MCC
92	guanine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
93	xanthine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
94	glycine betaine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
95	uridine	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
Mac.	-	-	-	-	-	-	-	0.998	-	0.758	0.744	-
Avg.	-	-	-	-	-	-	-	0.923	0.923	0.923	0.923	0.908

The table presents the detailed results for each substrate class in the UniProt-SPEC-100 dataset in five metrics accuracy, precision, recall, f1-score, and MCC. As the table shows, there are NaN values in some minority classes. This is due to the number of zero True Positives (TP) and False Positives (FP) in the class results.

C.3 TooT-SPEC results for 3-Fold Cross-Validation

Table 45: TooT-SPEC 3-fold cross-validation results for UniProt-SPEC-100

#	Substrate	Train	Test	TP	FP	FN	TN	Acc.	Prec.	Rec.	F1.	MCC
0	proton	817	409	400.33	44.00	1032.33	8.33	0.96	0.90	0.98	0.94	0.92
				±	±	±	±	±	±	±	±	±
				9.07	16.09	16.62	8.50	0.01	0.03	0.02	0.01	0.01
1	calcium(2+)	576	289	268.67	12.67	1184.00	19.67	0.98	0.96	0.93	0.94	0.93
				±	±	±	±	±	±	±	±	±
				6.43	9.71	10.15	5.86	0.00	0.03	0.02	0.01	0.01
2	potassium(1+)	484	242	225.67	13.67	1229.33	16.33	0.98	0.94	0.93	0.94	0.93
				±	±	±	±	±	±	±	±	±
				7.51	7.02	7.02	7.51	0.00	0.02	0.03	0.01	0.02
3	chloride	295	147	135.33	5.00	1332.67	12.00	0.99	0.96	0.92	0.94	0.93
				±	±	±	±	±	±	±	±	±
				2.52	3.00	3.51	3.00	0.00	0.02	0.02	0.02	0.02
4	sodium(1+)	178	88	75.33	8.33	1388.00	13.33	0.99	0.91	0.85	0.87	0.87
				±	±	±	±	±	±	±	±	±
				4.93	6.66	6.93	4.62	0.00	0.05	0.04	0.00	0.00
5	zinc(2+)	53	27	25.00	4.00	1454.33	1.67	1.00	0.88	0.94	0.90	0.90
				±	±	±	±	±	±	±	±	±
				1.73	4.36	3.79	2.08	0.00	0.10	0.06	0.05	0.04
6	sulfate	44	23	19.00	1.67	1461.00	3.33	1.00	0.93	0.85	0.88	0.88
				±	±	±	±	±	±	±	±	±
				1.00	2.08	1.73	1.15	0.00	0.07	0.04	0.02	0.02
7	ammonium	35	17	17.00	1.00	1466.67	0.33	1.00	0.95	0.98	0.96	0.96
				±	±	±	±	±	±	±	±	±
				1.00	1.00	0.58	0.58	0.00	0.04	0.03	0.03	0.03
8	water	29	14	14.33	0.00	1470.67	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00
9	riboflavin(1-)	24	13	8.67	2.67	1470.00	3.67	1.00	0.76	0.70	0.72	0.72
				±	±	±	±	±	±	±	±	±
				3.21	2.52	2.00	2.89	0.00	0.17	0.20	0.17	0.17
10	protein	18	10	7.00	0.33	1475.33	2.33	1.00	0.95	0.75	0.84	0.84
				±	±	±	±	±	±	±	±	±
				1.00	0.58	0.58	1.15	0.00	0.07	0.10	0.08	0.08
11	glutathione derivative	18	9	9.00	1.00	1475.00	0.00	1.00	0.91	1.00	0.95	0.95
				±	±	±	±	±	±	±	±	±
				0.00	1.00	1.00	0.00	0.00	0.07	0.00	0.04	0.04
12	retinol	18	9	8.33	0.00	1476.00	0.67	1.00	1.00	0.93	0.96	0.96
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.00	0.58	0.00	0.00	0.05	0.03	0.03
13	nitrate	18	8	6.00	1.67	1474.67	2.67	1.00	0.79	0.69	0.73	0.74
				±	±	±	±	±	±	±	±	±
				1.00	1.15	1.53	0.58	0.00	0.09	0.06	0.07	0.07
14	iron(2+)	16	8	7.67	0.67	1476.33	0.33	1.00	0.93	0.96	0.94	0.94
				±	±	±	±	±	±	±	±	±
				0.58	0.58	0.58	0.58	0.00	0.05	0.06	0.00	0.00
15	pyruvate	14	6	6.00	1.00	1477.33	0.67	1.00	0.89	0.90	0.89	0.89
				±	±	±	±	±	±	±	±	±
				0.00	1.73	2.08	0.58	0.00	0.16	0.07	0.10	0.10
16	ATP(4-)	12	6	6.00	0.00	1479.00	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	copper(1+)	11	6	3.67	0.67	1478.67	2.00	1.00	0.85	0.67	0.61	0.92
				±	±	±	±	±	±	±	±	±
				3.21	0.58	0.58	3.46	0.00	0.01	0.47	0.43	0.01
18	urea	10	6	4.67	0.00	1479.67	0.67	1.00	1.00	0.87	0.92	0.92
				±	±	±	±	±	±	±	±	±
				1.53	0.00	0.58	1.15	0.00	0.00	0.19	0.12	0.11
19	dehydroascorbide(1-)	9	5	4.33	1.00	1479.33	0.33	1.00	0.82	0.92	0.86	0.86
				±	±	±	±	±	±	±	±	±
				1.15	1.00	1.15	0.58	0.00	0.13	0.12	0.10	0.10
20	choline	9	4	4.33	0.00	1480.67	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00

Continued on next page

Table 45: TooT-SPEC 3-fold cross-validation results for UniProt-SPEC-100

#	Substrate	Train	Test	TP	FP	FN	TN	Acc.	Prec.	Rec.	F1.	MCC
21	cholesterol	8	4	4.00	1.00	1480.00	0.00	1.00	0.80	1.00	0.89	0.89
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22	L-ornithinium(1+)	8	4	3.67	0.67	1480.33	0.33	1.00	0.85	0.92	0.88	0.88
				±	±	±	±	±	±	±	±	±
				0.58	0.58	0.58	0.58	0.00	0.11	0.12	0.10	0.10
23	lipid-linked peptidoglycan	8	3	3.00	1.00	1480.33	0.67	1.00	0.86	0.81	0.79	0.81
				±	±	±	±	±	±	±	±	±
				1.00	1.73	2.08	0.58	0.00	0.20	0.14	0.05	0.05
24	L-cystine zwitterion	8	3	2.00	1.00	1480.33	1.67	1.00	0.76	0.56	0.58	0.61
				±	±	±	±	±	±	±	±	±
				1.00	1.00	1.15	1.15	0.00	0.17	0.22	0.13	0.08
25	bile acid	6	4	3.33	0.00	1481.67	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00
26	L-glutamate(1-)	7	3	2.33	2.00	1479.67	1.00	1.00	0.55	0.69	0.61	0.61
				±	±	±	±	±	±	±	±	±
				0.58	1.00	0.58	0.00	0.00	0.15	0.04	0.11	0.10
27	citrate(3-)	7	3	2.33	0.00	1481.67	1.00	1.00	1.00	0.72	0.82	0.84
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.58	1.00	0.00	0.00	0.21	0.14	0.12
28	cobalamin	6	3	1.67	2.67	1479.33	1.33	1.00	0.44	0.56	0.49	0.49
				±	±	±	±	±	±	±	±	±
				0.58	2.08	2.08	0.58	0.00	0.21	0.16	0.19	0.19
29	oleate	6	3	2.33	0.67	1481.33	0.67	1.00	0.87	0.78	0.75	0.78
				±	±	±	±	±	±	±	±	±
				1.15	1.15	1.15	1.15	0.00	0.19	0.31	0.20	0.17
30	heme	6	3	2.00	0.00	1482.00	1.00	1.00	1.00	0.67	0.77	0.80
				±	±	±	±	±	±	±	±	±
				1.00	0.00	0.00	1.00	0.00	0.00	0.27	0.21	0.17
31	S-adenosyl-L-methionine zwitterion	6	3	2.67	0.00	1482.00	0.33	1.00	1.00	0.89	0.93	0.94
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.00	0.58	0.00	0.00	0.16	0.09	0.09
32	UDP-D-glucose(2-)	6	3	2.00	0.67	1481.33	1.00	1.00	0.83	0.67	0.69	0.72
				±	±	±	±	±	±	±	±	±
				1.00	1.15	1.15	1.00	0.00	0.24	0.27	0.22	0.20
33	UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	2.00	1.00	1481.33	0.67	1.00	0.83	0.72	0.71	0.74
				±	±	±	±	±	±	±	±	±
				1.00	1.73	2.08	0.58	0.00	0.24	0.21	0.06	0.05
34	1-4-butanediammonium	6	2	0.33	3.00	1479.33	2.33	1.00	0.05	0.11	0.07	0.07
				±	±	±	±	±	±	±	±	±
				0.58	2.65	2.89	0.58	0.00	0.07	0.16	0.09	0.10
35	copper(2+)	5	3	1.33	2.00	1480.33	1.33	1.00	0.71	0.50	0.56	0.58
				±	±	±	±	±	±	±	±	±
				0.58	3.46	3.79	0.58	0.00	0.40	0.14	0.26	0.26
36	3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	2.00	0.00	1482.33	0.67	1.00	1.00	0.78	0.83	0.86
				±	±	±	±	±	±	±	±	±
				1.00	0.00	0.58	1.15	0.00	0.00	0.31	0.24	0.20
37	glutathionate(1-)	6	2	1.67	0.33	1482.00	1.00	1.00	0.89	0.67	0.70	0.74
				±	±	±	±	±	±	±	±	±
				0.58	0.58	0.00	1.00	0.00	0.16	0.27	0.14	0.11
38	urate salt	5	3	1.67	0.33	1482.00	1.00	1.00	0.83	0.61	0.70	0.71
				±	±	±	±	±	±	±	±	±
				0.58	0.58	0.00	0.00	0.00	0.24	0.08	0.14	0.15
39	biotinate	5	3	1.67	1.00	1481.33	1.00	1.00	0.72	0.67	0.62	0.66
				±	±	±	±	±	±	±	±	±
				0.58	1.00	1.15	1.00	0.00	0.21	0.27	0.13	0.11
40	sucrose	4	3	2.33	0.00	1482.67	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00
41	acetyl-CoA(4-)	5	2	2.33	0.00	1482.67	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00

Continued on next page

Table 45: TooT-SPEC 3-fold cross-validation results for UniProt-SPEC-100

#	Substrate	Train	Test	TP	FP	FN	TN	Acc.	Prec.	Rec.	F1.	MCC
42	2-oxoglutarate(2-)	5	2	1.33	0.33	1482.33	1.00	1.00	0.83	0.56	0.53	0.82
				±	±	±	±	±	±	±	±	±
				1.15	0.58	0.58	1.00	0.00	0.17	0.42	0.38	0.00
43	maltose	5	2	1.00	1.00	1481.67	1.33	1.00	0.50	0.50	0.39	0.60
				±	±	±	±	±	±	±	±	±
				1.00	1.00	0.58	1.53	0.00	0.00	0.41	0.28	0.10
44	phosphoglycerate	4	3	2.33	0.33	1482.33	0.00	1.00	0.89	1.00	0.93	0.94
				±	±	±	±	±	±	±	±	±
				0.58	0.58	0.58	0.00	0.00	0.16	0.00	0.09	0.09
45	iron(3+)	5	2	1.00	0.67	1482.00	1.33	1.00	0.67	0.50	0.47	0.70
				±	±	±	±	±	±	±	±	±
				1.00	1.15	1.00	1.53	0.00	0.33	0.41	0.41	0.30
46	hydrogencarbonate	4	3	1.00	1.33	1481.33	1.33	1.00	0.45	0.39	0.33	0.51
				±	±	±	±	±	±	±	±	±
				1.00	1.53	2.08	0.58	0.00	0.05	0.28	0.24	0.01
47	iron-sulfur cluster	4	2	2.00	0.00	1483.00	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
48	L-proline zwitterion	4	2	1.00	0.67	1482.33	1.00	1.00	0.56	0.50	0.49	0.51
				±	±	±	±	±	±	±	±	±
				1.00	0.58	0.58	1.00	0.00	0.42	0.41	0.35	0.36
49	L-argininium(1+)	4	2	1.33	0.33	1482.67	0.67	1.00	0.67	0.67	0.67	0.67
				±	±	±	±	±	±	±	±	±
				1.15	0.58	0.58	1.15	0.00	0.47	0.47	0.47	0.47
50	NAD	4	2	2.00	0.33	1482.67	0.00	1.00	0.89	1.00	0.93	0.94
				±	±	±	±	±	±	±	±	±
				0.00	0.58	0.58	0.00	0.00	0.16	0.00	0.09	0.09
51	acetylcholine	4	2	1.33	0.00	1483.00	0.67	1.00	1.00	0.67	0.78	0.80
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.00	0.58	0.00	0.00	0.24	0.16	0.14
52	fluoride	4	2	2.00	0.00	1483.00	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
53	formate	4	2	1.67	0.33	1482.67	0.33	1.00	0.89	0.83	0.82	0.84
				±	±	±	±	±	±	±	±	±
				0.58	0.58	0.58	0.58	0.00	0.16	0.24	0.14	0.12
54	pyrimidine nucleotide	3	2	1.67	0.00	1483.33	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00
55	cadmium cation	3	2	1.33	0.33	1483.00	0.33	1.00	0.89	0.83	0.82	0.84
				±	±	±	±	±	±	±	±	±
				0.58	0.58	1.00	0.58	0.00	0.16	0.24	0.14	0.12
56	GDP-fucose	3	2	1.33	0.67	1482.67	0.33	1.00	0.78	0.83	0.72	0.76
				±	±	±	±	±	±	±	±	±
				0.58	1.15	0.58	0.58	0.00	0.31	0.24	0.21	0.18
57	fatty acyl-CoA	3	2	1.67	0.00	1483.33	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00
58	mercury cation	4	1	1.33	0.00	1483.33	0.33	1.00	1.00	0.83	0.89	0.90
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.58	0.58	0.00	0.00	0.24	0.16	0.14
59	phosphate ion	4	1	1.67	0.00	1483.33	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00
60	N-ribosylnicotinamide	4	1	1.00	0.00	1483.33	0.67	1.00	1.00	0.50	0.56	0.85
				±	±	±	±	±	±	±	±	±
				1.00	0.00	0.58	0.58	0.00	0.00	0.41	0.42	0.15
61	gamma-aminobutyric acid zwitterion	3	2	0.33	0.00	1483.33	1.33	1.00	1.00	0.17	0.22	0.71
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.58	0.58	0.00	0.00	0.24	0.31	0.00
62	long-chain fatty acid	3	1	1.33	0.00	1483.67	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00
63	L-arabinose	2	2	0.33	1.00	1482.67	1.00	1.00	0.25	0.33	0.13	0.50
				±	±	±	±	±	±	±	±	±
				0.58	1.73	1.53	1.00	0.00	0.00	0.47	0.19	0.00

Continued on next page

Table 45: TooT-SPEC 3-fold cross-validation results for UniProt-SPEC-100

#	Substrate	Train	Test	TP	FP	FN	TN	Acc.	Prec.	Rec.	F1.	MCC
64	bacteriocin	3	1	1.00	0.33	1483.33	0.33	1.00	0.75	0.67	0.56	0.85
				±	±	±	±	±	±	±	±	±
				1.00	0.58	0.58	0.58	0.00	0.25	0.47	0.42	0.15
65	glycine zwitterion	3	1	1.33	0.00	1483.67	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00
66	L-histidine zwitterion	3	1	0.00	0.33	1483.33	1.33	1.00	0.00	0.00	0.00	-0.00
				±	±	±	±	±	±	±	±	±
				0.00	0.58	1.15	0.58	0.00	0.00	0.00	0.00	0.00
67	L-ascorbate	3	1	1.00	0.33	1483.33	0.33	1.00	0.83	0.83	0.78	0.80
				±	±	±	±	±	±	±	±	±
				0.00	0.58	0.58	0.58	0.00	0.24	0.24	0.16	0.14
68	L-methionine zwitterion	3	1	0.00	0.00	1483.67	1.33	1.00	NaN	0.00	0.00	NaN
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.58	0.58	0.00	NaN	0.00	0.00	NaN
69	boric acid	2	2	0.67	0.33	1483.33	0.67	1.00	0.75	0.67	0.56	0.85
				±	±	±	±	±	±	±	±	±
				0.58	0.58	0.58	1.15	0.00	0.25	0.47	0.42	0.15
70	D-methionine zwitterion	2	2	1.33	0.67	1483.00	0.00	1.00	0.78	1.00	0.83	0.86
				±	±	±	±	±	±	±	±	±
				0.58	1.15	1.00	0.00	0.00	0.31	0.00	0.24	0.20
71	prostaglandin	3	1	1.33	0.00	1483.67	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00
72	maltodextrin	3	1	1.00	0.00	1483.67	0.33	1.00	1.00	0.67	0.67	1.00
				±	±	±	±	±	±	±	±	±
				1.00	0.00	0.58	0.58	0.00	0.00	0.47	0.47	0.00
73	folate(2-)	2	1	0.00	0.00	1484.00	1.00	1.00	NaN	0.00	0.00	NaN
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	NaN	0.00	0.00	NaN
74	molybdate	2	1	0.00	0.00	1484.00	1.00	1.00	NaN	0.00	0.00	NaN
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	NaN	0.00	0.00	NaN
75	N-retinylidenephosphatidylethanolamine	2	1	0.33	0.00	1484.00	0.67	1.00	1.00	0.33	0.33	1.00
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.00	0.58	0.00	0.00	0.47	0.47	0.00
76	lipopolysaccharide	2	1	1.00	0.00	1484.00	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
77	beta-D-glucan	2	1	1.00	0.00	1484.00	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
78	gluconate	2	1	1.00	0.00	1484.00	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
79	tungstate	2	1	0.67	1.00	1483.00	0.33	1.00	0.67	0.67	0.67	0.67
				±	±	±	±	±	±	±	±	±
				0.58	1.73	1.73	0.58	0.00	0.47	0.47	0.47	0.47
80	sn-glycerol 3-phosphate(2-)	2	1	0.00	0.00	1484.00	1.00	1.00	NaN	0.00	0.00	NaN
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	NaN	0.00	0.00	NaN
81	UDP-alpha-D-xylose(2-)	2	1	1.00	0.00	1484.00	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
82	magnesium cation	2	1	1.00	0.00	1484.00	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
83	flavin adenine dinucleotide	2	1	0.33	0.33	1483.67	0.67	1.00	0.50	0.33	0.22	0.71
				±	±	±	±	±	±	±	±	±
				0.58	0.58	0.58	0.58	0.00	0.00	0.47	0.31	0.00
84	(R)-carnitine	2	1	1.00	0.00	1484.00	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Continued on next page

Table 45: TooT-SPEC 3-fold cross-validation results for UniProt-SPEC-100

#	Substrate	Train	Test	TP	FP	FN	TN	Acc.	Prec.	Rec.	F1.	MCC
85	CMP-N-acetyl-beta -neuraminate(2-)	2	1	0.67	0.33	1483.67	0.33	1.00	0.75	0.67	0.56	0.85
				±	±	±	±	±	±	±	±	±
				0.58	0.58	0.58	0.58	0.00	0.25	0.47	0.42	0.15
86	c-GMP-AMP(2-)	2	1	1.00	0.00	1484.00	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
87	L-tryptophan zwitterion	2	1	0.33	1.00	1483.00	0.67	1.00	0.17	0.33	0.17	0.29
				±	±	±	±	±	±	±	±	±
				0.58	1.00	1.00	0.58	0.00	0.17	0.47	0.24	0.29
88	N-(4-aminobenzoyl) -L-glutamate	2	1	1.00	0.00	1484.00	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
89	phosphonatoenolpyruvate	2	1	0.67	0.00	1484.00	0.33	1.00	1.00	0.67	0.67	1.00
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.00	0.58	0.00	0.00	0.47	0.47	0.00
90	arsenate ion	2	1	1.00	0.00	1484.00	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
91	teichoic acid	2	1	1.00	0.00	1484.00	0.00	1.00	1.00	1.00	1.00	1.00
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
92	guanine	2	1	0.67	0.00	1484.00	0.33	1.00	1.00	0.67	0.67	1.00
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.00	0.58	0.00	0.00	0.47	0.47	0.00
93	xanthine	2	1	0.67	0.00	1484.00	0.33	1.00	1.00	0.67	0.67	1.00
				±	±	±	±	±	±	±	±	±
				0.58	0.00	0.00	0.58	0.00	0.00	0.47	0.47	0.00
94	glycine betaine	2	1	0.00	0.00	1484.00	1.00	1.00	NaN	0.00	0.00	NaN
				±	±	±	±	±	±	±	±	±
				0.00	0.00	0.00	0.00	0.00	NaN	0.00	0.00	NaN
95	uridine	2	1	0.67	0.67	1483.33	0.33	1.00	0.50	0.67	0.56	0.57
				±	±	±	±	±	±	±	±	±
				0.58	0.58	0.58	0.58	0.00	0.41	0.47	0.42	0.42
Mac. Avg	-	-	-	-	-	-	-	0.921 ± 0.02	0.921 ± 0.02	0.921 ± 0.02	0.921 ± 0.02	0.906 ± 0.02

The table presents the 3-fold cross-validation results for each substrate class in five metrics. The NaN values for MCC in some of the minority classes show the number zero of true positive (TP) and false positive (FP).

Appendix D

TooT-Open-ICAT

The TooT-Open-ICAT model is used to predict specific substrates for the transportomes of 5 model organisms *Escherichia coli K-12 (Es)*, *Saccharomyces cerevisiae (Sc)*, *Drosophila melanogaste (Dm)*, *Homo sapiens (Hs)*, and *Arabidopsis thaliana (At)*, with a fixed threshold of $\tau=0.5$ on the softmax layer. The results of this experiment are presented in Table 46.

Table 46: TooT-Open-ICAT prediction for transportome of 5 model organisms

Substrate	Es	Sc	Dm	Hs	At
Proteome	4391	5887	13806	20609	27468
Transportome	894 (20.360%)	702 (11.925%)	891 (6.454%)	1767 (8.574%)	1654 (6.022%)
Non-ICAT	707 (79.083%)	372 (52.991%)	396 (44.444%)	508 (28.749%)	892 (53.930%)
proton	135 (15.101%)	183 (26.068%)	135 (15.152%)	366 (20.713%)	341 (20.617%)
potassium(1+)	11 (1.230%)	32 (4.558%)	73 (8.193%)	209 (11.828%)	109 (6.590%)
other ion transporters (reject)	9 (1.007%)	19 (2.707%)	9 (1.010%)	32 (1.811%)	43 (2.600%)
sulfate	9 (1.007%)	10 (1.425%)	23 (2.581%)	18 (1.019%)	26 (1.572%)
sodium(1+)	7 (0.783%)	24 (3.419%)	131 (14.703%)	194 (10.979%)	62 (3.748%)
zinc(2+)	5 (0.559%)	10 (1.425%)	17 (1.908%)	25 (1.415%)	39 (2.358%)
chloride	3 (0.336%)	7 (0.997%)	34 (3.816%)	81 (4.584%)	19 (1.149%)
phosphate ion	3 (0.336%)	0 (0.000%)	1 (0.112%)	4 (0.226%)	5 (0.302%)
iron(2+)	2 (0.224%)	3 (0.427%)	0 (0.000%)	1 (0.057%)	9 (0.544%)
ammonium	1 (0.112%)	7 (0.997%)	2 (0.224%)	5 (0.283%)	6 (0.363%)
copper(1+)	1 (0.112%)	1 (0.142%)	1 (0.112%)	1 (0.057%)	2 (0.121%)
nitrate	1 (0.112%)	2 (0.285%)	2 (0.224%)	12 (0.679%)	22 (1.330%)
calcium(2+)	0 (0.000%)	32 (4.558%)	67 (7.519%)	311 (17.600%)	79 (4.776%)
Total	894	702	891	1767	1654

The analysis for transportome of five model organisms, *Escherichia coli K-12* (*Es*), *Saccharomyces cerevisiae* (*Sc*), *Drosophila melanogaste* (*Dm*), *Homo sapiens* (*Hs*), and *Arabidopsis thaliana* (*At*), using the TooT-Open-ICAT tool with fixed threshold of $\tau=0.5$ on the softmax layer. The numbers in parentheses indicate the proportion of the transportome in relation to the proteome in the second row, and the proportion of the substrate's transporter within the transportome in the subsequent rows.

Appendix E

TooT-Triplet

E.1 Introduction

This chapter presents further results and analysis of the TooT-Triplet model from Chapter 5. The following sections present TooT-Triplet’s detailed results for the hold-out dataset (Section E.2.1) and 3-fold cross-validation (Section E.2.2) on the UniProt-SPEC-100 dataset. The results of the baseline methods for the same dataset are presented in Section E.3.1. The detailed results of the TooT-SPEC model (described in Section 3.7) followed by a K-NN classifier are presented in Section E.3.2. The external validation of the TooT-Triplet model (Section E.4) includes validating the performance with the TooT-SC dataset in Section E.4.1 and the transportomes of five model organisms in E.4.2. Finally, Section E.5 presents the inter and intra-class distances table for the training set after triplet training.

E.2 TooT-Triplet detailed results

This section presents the detailed results of the TooT-Triplet models trained with different distance metrics using the offline sampling strategy. Section E.2.1 includes the detailed results for each class in the UniProt-SPEC-100 hold-out test set. Section E.2.2 presents the results of the TooT-Triplet model trained using offline sampling strategy with Manhattan distance metric on each fold in a 3-fold cross-validation for the same dataset. Finally the 3-fold cross-validation results for the TooT-Triplet model trained and tested in the absence of the majority class (experiment described in Section 5.8.3) are presented in Section E.2.3.

E.2.1 Hold-out test set results

This section presents the detailed results for each class in the hold-out test set of UniProt-SPEC-100 for the TooT-Triplet model trained and tested using an offline sampling strategy with three different distance metrics: Manhattan, Euclidean, and Cosine similarity. Table 23 in Section 5.7 details the performance of the TooT-Triplet model trained using offline sampling and Manhattan distance metric. Figure 47 shows the confusion matrix for the same results. Table 68 and Table 48 show the results of the TooT-Triplet model trained with Euclidean and Cosine similarity distances respectively (both using offline sampling strategy).

Table 47: TooT-Triplet (offline, Euclidean) results for the hold-out test set of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	818	408	379	16	30	1060	0.969	0.959	0.927	0.943	0.922
calcium(2+)	577	288	279	15	9	1182	0.984	0.949	0.969	0.959	0.949
potassium(1+)	484	242	228	19	14	1224	0.978	0.923	0.942	0.933	0.919
chloride	294	148	142	11	5	1327	0.989	0.928	0.966	0.947	0.941
sodium(1+)	177	89	82	10	7	1386	0.99	0.89	0.92	0.91	0.90
zinc(2+)	53	27	22	4	4	1455	0.99	0.85	0.85	0.85	0.84
sulfate	45	22	15	10	7	1453	0.99	0.60	0.68	0.64	0.63
ammonium	35	17	18	0	0	1467	1.00	1.00	1.00	1.00	1.00
water	29	14	15	0	0	1470	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	25	12	8	5	4	1468	0.99	0.61	0.67	0.64	0.64
protein	19	9	7	0	2	1476	1.0	1.0	0.8	0.9	0.9
retinol	18	9	7	0	2	1476	1.0	1.0	0.8	0.9	0.9
glutathione derivative	18	9	9	2	0	1474	1.0	0.8	1.0	0.9	0.9
nitrate	17	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
iron(2+)	16	8	6	3	2	1474	1.0	0.7	0.8	0.7	0.7
pyruvate	13	7	5	0	2	1478	1.0	1.0	0.7	0.8	0.8
ATP(4-)	12	6	5	0	1	1479	1.0	1.0	0.8	0.9	0.9
copper(1+)	11	6	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
urea	11	5	4	0	1	1480	1.0	1.0	0.8	0.9	0.9
dehydroascorbide(1-)	9	5	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
choline	8	5	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
cholesterol	8	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
L-ornithinium(1+)	8	4	2	0	2	1481	1.0	1.0	0.5	0.7	0.7
citrate(3-)	7	3	2	1	2	1480	1.0	0.7	0.5	0.6	0.6
bile acid	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
L-cystine zwitterion	7	4	0	0	4	1481	1.0	-	0.0	0.0	NaN
lipid-linked peptidoglycan	7	4	2	2	2	1479	1.0	0.5	0.5	0.5	0.5
cobalamin	6	3	0	0	3	1482	1.0	-	0.0	0.0	NaN
oleate	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
heme	6	3	1	1	2	1481	1.0	0.5	0.3	0.4	0.4
S-adenosyl-L-methionine zwitterion	6	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
UDP-D-glucose(2-)	6	3	1	1	2	1481	1.0	0.5	0.3	0.4	0.4
L-glutamate(1-)	6	4	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
biotinate	6	2	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
urate salt	5	3	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
acetyl-CoA(4-)	5	2	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
hydrogencarbonate	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
iron(3+)	5	2	0	1	3	1481	1.0	0.0	0.0	0.0	0.0
phosphoglycerate	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
maltose	5	2	0	1	3	1481	1.0	0.0	0.0	0.0	0.0
2-oxoglutarate(2-)	5	2	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
sucrose	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
glutathionate(1-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
copper(2+)	5	3	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
1-4-butanediammonium	5	3	1	2	2	1480	1.0	0.3	0.3	0.3	0.3
UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
formate	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
acetylcholine	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
fatty acyl-CoA	4	1	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
fluoride	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
L-proline zwitterion	4	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
NAD	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
L-argininium(1+)	4	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-arabinose	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
prostaglandin	3	1	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
D-methionine zwitterion	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
boric acid	3	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
L-methionine zwitterion	3	1	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
L-histidine zwitterion	3	1	0	1	2	1482	1.0	0.0	0.0	0.0	0.0

Continued on next page

Table 47: TooT-Triplet (offline, Euclidean) results for the hold-out test set of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
glycine zwitterion	3	1	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
bacteriocin	3	1	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
long-chain fatty acid	3	1	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid	3	2	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
zwitterion											
N-ribosylnicotinamide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphate ion	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
mercury cation	3	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
GDP-fucose	3	2	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
cadmium cation	3	2	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
pyrimidine nucleotide	3	2	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
neuraminate(2-)											
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
N-(4-aminobenzoyl)-L-	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glutamate											
glycine betaine	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
teichoic acid	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
guanine	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
xanthine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
molybdate	2	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
sn-glycerol 3-phosphate(2-)	2	1	0	3	1	1481	1.0	0.0	0.0	0.0	0.0
)											
tungstate	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
gluconate	2	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
N-retinylidene phos-	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phatidylethanolamine											
folate(2-)	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
maltodextrin	2	2	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-ascorbate	2	2	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
uridine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
Overall	2970	1485	1346	-	-	-	0.998	0.906	0.744	0.715	0.889

The table presents the results of the hold-out test set of UniProt-SPEC-100 using TooT-Triplet (offline) trained and tested using Euclidean distance metric and a K-NN classifier (K=1). The overall number for TP presents the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. The overall entry shows the macro average over all the classes for each metric. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The significant digits presented for these metrics correspond to the Test set size.

Table 48: TooT-Triplet (offline, Cosine) results for the hold-out test set of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	818	408	383	25	26	1051	0.966	0.939	0.936	0.938	0.914
calcium(2+)	577	288	266	12	22	1185	0.977	0.957	0.924	0.940	0.926
potassium(1+)	484	242	232	14	10	1229	0.984	0.943	0.959	0.951	0.941
chloride	294	148	137	19	10	1319	0.980	0.878	0.932	0.904	0.894
sodium(1+)	177	89	79	8	10	1388	0.99	0.91	0.89	0.90	0.89
zinc(2+)	53	27	21	3	5	1456	0.99	0.88	0.81	0.84	0.84
sulfate	45	22	15	3	7	1460	0.99	0.83	0.68	0.75	0.75
ammonium	35	17	18	2	0	1465	1.00	0.90	1.00	0.95	0.95
water	29	14	15	0	0	1470	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	25	12	8	5	4	1468	0.99	0.61	0.67	0.64	0.64

Continued on next page

Table 48: TooT-Triplet (offline, Cosine) results for the hold-out test set of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
protein	19	9	8	2	1	1474	1.0	0.8	0.9	0.8	0.8
retinol	18	9	7	4	2	1472	1.0	0.6	0.8	0.7	0.7
glutathione derivative	18	9	9	0	0	1476	1.0	1.0	1.0	1.0	1.0
nitrate	17	9	7	2	2	1474	1.0	0.8	0.8	0.8	0.8
iron(2+)	16	8	6	3	2	1474	1.0	0.7	0.8	0.7	0.7
pyruvate	13	7	6	0	1	1478	1.0	1.0	0.9	0.9	0.9
ATP(4-)	12	6	5	0	1	1479	1.0	1.0	0.8	0.9	0.9
copper(1+)	11	6	4	2	1	1478	1.0	0.7	0.8	0.7	0.7
urea	11	5	4	0	1	1480	1.0	1.0	0.8	0.9	0.9
dehydroascorbide(1-)	9	5	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
choline	8	5	3	1	1	1480	1.0	0.8	0.8	0.8	0.7
cholesterol	8	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
L-ornithinium(1+)	8	4	3	0	1	1481	1.0	1.0	0.8	0.9	0.9
citrate(3-)	7	3	2	3	2	1478	1.0	0.4	0.5	0.4	0.4
bile acid	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
L-cystine zwitterion	7	4	0	0	4	1481	1.0	-	0.0	0.0	NaN
lipid-linked peptidoglycan	7	4	3	1	1	1480	1.0	0.8	0.8	0.8	0.7
cobalamin	6	3	0	0	3	1482	1.0	-	0.0	0.0	NaN
oleate	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
heme	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
S-adenosyl-L-methionine	6	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
zwitterion											
UDP-D-glucose(2-)	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
L-glutamate(1-)	6	4	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
biotinate	6	2	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
urate salt	5	3	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
acetyl-CoA(4-)	5	2	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
hydrogencarbonate	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
iron(3+)	5	2	0	3	3	1479	1.0	0.0	0.0	0.0	0.0
phosphoglycerate	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
maltose	5	2	0	1	3	1481	1.0	0.0	0.0	0.0	0.0
2-oxoglutarate(2-)	5	2	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
3'-phosphonato-5'-	5	3	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
adenyl sulfate(4-)											
sucrose	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
glutathionate(1-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
copper(2+)	5	3	0	0	2	1483	1.0	-	0.0	0.0	NaN
1-4-butanediammonium	5	3	0	1	3	1481	1.0	0.0	0.0	0.0	0.0
UDP-N-acetyl-alpha-D-	5	3	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
glucosamine(2-)											
formate	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
acetylcholine	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
fatty acyl-CoA	4	1	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
fluoride	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
L-proline zwitterion	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
NAD	4	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
L-argininium(1+)	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-arabinose	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
prostaglandin	3	1	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
D-methionine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
boric acid	3	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
L-methionine zwitterion	3	1	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
L-histidine zwitterion	3	1	0	1	2	1482	1.0	0.0	0.0	0.0	0.0
glycine zwitterion	3	1	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
bacteriocin	3	1	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
long-chain fatty acid	3	1	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
gamma-aminobutyric acid	3	2	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
zwitterion											
N-ribosyl nicotinamide	3	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
phosphate ion	3	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
mercury cation	3	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
GDP-fucose	3	2	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
cadmium cation	3	2	0	3	1	1481	1.0	0.0	0.0	0.0	0.0

Continued on next page

Table 48: TooT-Triplet (offline, Cosine) results for the hold-out test set of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
pyrimidine nucleotide	3	2	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuraminate(2-)	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	5	1	1479	1.0	0.0	0.0	0.0	0.0
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
teichoic acid	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
guanine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
xanthine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
molybdate	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
sn-glycerol 3-phosphate(2-)	2	1	0	4	1	1480	1.0	0.0	0.0	0.0	0.0
tungstate	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
gluconate	2	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
N-retinylidene phos-phatidylethanolamine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
maltodextrin	2	2	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-ascorbate	2	2	0	0	1	1484	1.0	-	0.0	0.0	NaN
uridine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
Overall	2970	1485	1328	-	-	-	0.998	0.894	0.716	0.683	0.875

The table presents the results of the hold-out test set of UniProt-SPEC-100 using TooT-Triplet (offline) trained and tested using the Cosine distance metric and a K-NN classifier (K=1). The overall number for TP presents the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric. The significant digits presented for these metrics correspond to the Test set size.

E.2.2 Folds results

TooT-Triplet (Offline, Manhattan) results

Table 49: TooT-Triplet (offline, manhattan) results for Fold 1 of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	817	409	375	40	34	1036	0.950	0.904	0.917	0.910	0.876
calcium(2+)	576	289	259	39	30	1157	0.954	0.869	0.896	0.882	0.854
potassium(1+)	484	242	210	28	32	1215	0.960	0.882	0.868	0.875	0.851
chloride	295	147	113	28	34	1310	0.958	0.801	0.769	0.785	0.762
sodium(1+)	178	88	75	10	13	1387	0.98	0.88	0.85	0.87	0.86
zinc(2+)	53	27	24	2	3	1456	1.00	0.92	0.89	0.91	0.90
sulfate	44	23	12	10	11	1452	0.99	0.55	0.52	0.53	0.53
ammonium	35	17	15	2	2	1466	1.00	0.88	0.88	0.88	0.88
water	29	14	14	0	0	1471	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	24	13	9	2	4	1470	1.00	0.82	0.69	0.75	0.75
protein	18	10	8	2	2	1473	1.00	0.80	0.80	0.80	0.80
glutathione derivative	18	9	9	1	0	1475	1.0	0.9	1.0	0.9	0.9
retinol	18	9	8	1	1	1475	1.0	0.9	0.9	0.9	0.9
nitrate	18	8	5	0	3	1477	1.0	1.0	0.6	0.8	0.8
iron(2+)	16	8	7	2	1	1475	1.0	0.8	0.9	0.8	0.8
pyruvate	14	6	6	1	0	1478	1.0	0.9	1.0	0.9	0.9

Continued on next page

Table 49: TooT-Triplet (offline, manhattan) results for Fold 1 of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
ATP(4-)	12	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
copper(1+)	11	6	6	2	0	1477	1.0	0.8	1.0	0.9	0.9
urea	10	6	5	0	1	1479	1.0	1.0	0.8	0.9	0.9
dehydroascorbide(1-)	9	5	4	0	1	1480	1.0	1.0	0.8	0.9	0.9
choline	9	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
cholesterol	8	4	2	1	2	1480	1.0	0.7	0.5	0.6	0.6
L-ornithinium(1+)	8	4	3	1	1	1480	1.0	0.8	0.8	0.8	0.7
lipid-linked peptidoglycan	8	3	2	3	1	1479	1.0	0.4	0.7	0.5	0.5
L-cystine zwitterion	8	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
L-glutamate(1-)	7	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
citrate(3-)	7	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
S-adenosyl-L-methionine zwitterion	6	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
glutathionate(1-)	6	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
1-4-butanediammonium	6	2	0	2	2	1481	1.0	0.0	0.0	0.0	0.0
UDP-D-glucose(2-)	6	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
bile acid	6	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
heme	6	3	3	3	0	1479	1.0	0.5	1.0	0.7	0.7
oleate	6	3	3	4	0	1478	1.0	0.4	1.0	0.6	0.7
cobalamin	6	3	2	1	1	1481	1.0	0.7	0.7	0.7	0.7
UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	1	1	2	1481	1.0	0.5	0.3	0.4	0.4
iron(3+)	5	2	0	0	2	1483	1.0	-	0.0	0.0	NaN
maltose	5	2	1	2	1	1481	1.0	0.3	0.5	0.4	0.4
2-oxoglutarate(2-)	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
acetyl-CoA(4-)	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
biotinate	5	3	1	2	2	1480	1.0	0.3	0.3	0.3	0.3
urate salt	5	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
copper(2+)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
N-ribosylnicotinamide	4	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
phosphate ion	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
mercury cation	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-proline zwitterion	4	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
NAD	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-argininium(1+)	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
hydrogencarbonate	4	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
phosphoglycerate	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
sucrose	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
bacteriocin	3	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
maltodextrin	3	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
prostaglandin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-methionine zwitterion	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
L-ascorbate	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-histidine zwitterion	3	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
pyrimidine nucleotide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid zwitterion	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
GDP-fucose	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
cadmium cation	3	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
long-chain fatty acid	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuraminate(2-)	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN

Continued on next page

Table 49: TooT-Triplet (offline, manhattan) results for Fold 1 of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
arsenate ion	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
xanthine	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
sn-glycerol 3-phosphate(2-)	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
tungstate	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
gluconate	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	3	0	1481	1.0	0.2	1.0	0.4	0.5
N-retinylidene phos- phatidylethanolamine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
molybdate	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
D-methionine zwitterion	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
boric acid	2	2	0	1	2	1482	1.0	0.0	0.0	0.0	0.0
L-arabinose	2	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
uridine	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
Overall	2970	1485	1271	-	-	-	0.997	0.856	0.709	0.682	0.829

The table presents the Cross-validation results of UniProt-SPEC-100, Fold 1, using TooT-Triplet (offline) trained and tested using the Manhattan distance metric and a K-NN classifier (K=1). The overall number for TP presents the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric. The significant digits presented for these metrics correspond to the Test set size.

Table 50: TooT-Triplet (offline, manhattan) results for Fold 2 of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	817	409	383	16	26	1060	0.972	0.960	0.936	0.948	0.929
calcium(2+)	577	288	280	14	8	1183	0.985	0.952	0.972	0.962	0.953
potassium(1+)	484	242	232	14	10	1229	0.984	0.943	0.959	0.951	0.941
chloride	295	147	141	11	6	1327	0.989	0.928	0.959	0.943	0.937
sodium(1+)	177	89	78	7	11	1389	0.99	0.92	0.88	0.90	0.89
zinc(2+)	54	26	22	4	4	1455	0.99	0.85	0.85	0.85	0.84
sulfate	45	22	16	7	6	1456	0.99	0.70	0.73	0.71	0.71
ammonium	34	18	18	2	0	1465	1.00	0.90	1.00	0.95	0.95
water	28	15	15	0	0	1470	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	25	12	8	4	4	1469	0.99	0.67	0.67	0.67	0.66
protein	19	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
retinol	18	9	7	3	2	1473	1.0	0.7	0.8	0.7	0.7
glutathione derivative	18	9	9	0	0	1476	1.0	1.0	1.0	1.0	1.0
nitrate	17	9	7	0	2	1476	1.0	1.0	0.8	0.9	0.9
iron(2+)	16	8	5	1	3	1476	1.0	0.8	0.6	0.7	0.7
pyruvate	13	7	5	0	2	1478	1.0	1.0	0.7	0.8	0.8
ATP(4-)	12	6	5	0	1	1479	1.0	1.0	0.8	0.9	0.9
copper(1+)	12	5	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
urea	11	5	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
dehydroascorbide(1-)	10	4	4	2	0	1479	1.0	0.7	1.0	0.8	0.8
choline	9	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
cholesterol	8	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
L-ornithinium(1+)	8	4	3	0	1	1481	1.0	1.0	0.8	0.9	0.9
L-glutamate(1-)	7	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
bile acid	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
lipid-linked peptidoglycan	7	4	2	3	2	1478	1.0	0.4	0.5	0.4	0.4
L-cystine zwitterion	7	4	0	1	4	1480	1.0	0.0	0.0	0.0	0.0
copper(2+)	6	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
citrate(3-)	6	4	2	0	2	1481	1.0	1.0	0.5	0.7	0.7
cobalamin	6	3	0	0	3	1482	1.0	-	0.0	0.0	NaN
oleate	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
heme	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6

Continued on next page

Table 50: TooT-Triplet (offline, manhattan) results for Fold 2 of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
S-adenosyl-L-methionine zwitterion	6	3	3	3	0	1479	1.0	0.5	1.0	0.7	0.7
UDP-D-glucose(2-)	6	3	1	1	2	1481	1.0	0.5	0.3	0.4	0.4
UDP-N-acetyl-alpha-D-glucosamine(2-)	6	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
3'-phosphonato-5'-adenylyl sulfate(4-)	6	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
urate salt	6	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
sucrose	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
hydrogencarbonate	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
glutathionate(1-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
biotinate	5	3	2	3	1	1479	1.0	0.4	0.7	0.5	0.5
1-4-butanediammonium	5	3	1	1	2	1481	1.0	0.5	0.3	0.4	0.4
phosphoglycerate	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
pyrimidine nucleotide	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
formate	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
acetylcholine	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
cadmium cation	4	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
GDP-fucose	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid zwitterion	4	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
fluoride	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
L-proline zwitterion	4	2	2	5	0	1478	1.0	0.3	1.0	0.4	0.5
NAD	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
L-argininium(1+)	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
iron(3+)	4	3	0	1	3	1481	1.0	0.0	0.0	0.0	0.0
maltose	4	3	0	0	3	1482	1.0	-	0.0	0.0	NaN
2-oxoglutarate(2-)	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
acetyl-CoA(4-)	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
D-methionine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
boric acid	3	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
L-ascorbate	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
L-arabinose	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
maltodextrin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
N-ribosylnicotinamide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphate ion	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
mercury cation	3	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuraminatate(2-)	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
xanthine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
sn-glycerol 3-phosphate(2-)	2	1	0	4	1	1480	1.0	0.0	0.0	0.0	0.0
tungstate	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
gluconate	2	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
N-retinylidene phosphatidylethanolamine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
molybdate	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
prostaglandin	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0

Continued on next page

Table 50: TooT-Triplet (offline, manhattan) results for Fold 2 of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
L-methionine zwitterion	2	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
L-histidine zwitterion	2	2	0	1	2	1482	1.0	0.0	0.0	0.0	0.0
glycine zwitterion	2	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
bacteriocin	2	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
long-chain fatty acid	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
uridine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
Overall	2970	1485	1353	-	-	-	0.998	0.911	0.757	0.725	0.895

The table presents the Cross-validation results of UniProt-SPEC-100, Fold 2, using TooT-Triplet (offline) trained and tested using the Manhattan distance metric and a K-NN classifier (K=1). The overall number for TP presents the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric. The significant digits presented for these metrics correspond to the Test set size.

Table 51: TooT-Triplet (offline, manhattan) results for Fold 3 of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	818	408	383	19	25	1058	0.970	0.953	0.939	0.946	0.925
calcium(2+)	577	288	265	25	23	1172	0.968	0.914	0.920	0.917	0.897
potassium(1+)	484	242	215	22	27	1221	0.967	0.907	0.888	0.898	0.878
chloride	294	148	126	19	22	1318	0.972	0.869	0.851	0.860	0.845
sodium(1+)	177	89	83	16	6	1380	0.98	0.84	0.93	0.88	0.88
zinc(2+)	53	27	23	1	4	1457	1.00	0.96	0.85	0.90	0.90
sulfate	45	22	15	2	7	1461	0.99	0.88	0.68	0.77	0.77
ammonium	35	17	16	0	1	1468	1.00	1.00	0.94	0.97	0.97
water	29	14	14	0	0	1471	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	25	12	5	3	7	1470	0.99	0.62	0.42	0.50	0.51
protein	19	9	8	1	1	1475	1.0	0.9	0.9	0.9	0.9
glutathione derivative	18	9	9	4	0	1472	1.0	0.7	1.0	0.8	0.8
retinol	18	9	9	1	0	1475	1.0	0.9	1.0	0.9	0.9
nitrate	17	9	6	2	3	1474	1.0	0.8	0.7	0.7	0.7
iron(2+)	16	8	7	3	1	1474	1.0	0.7	0.9	0.8	0.8
pyruvate	13	7	6	1	1	1477	1.0	0.9	0.9	0.9	0.9
ATP(4-)	12	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
urea	11	5	3	1	2	1479	1.0	0.8	0.6	0.7	0.7
copper(1+)	11	6	3	2	3	1477	1.0	0.6	0.5	0.5	0.5
dehydroascorbide(1-)	9	5	5	3	0	1477	1.0	0.6	1.0	0.8	0.8
choline	8	5	4	0	1	1480	1.0	1.0	0.8	0.9	0.9
L-ornithinium(1+)	8	4	3	0	1	1481	1.0	1.0	0.8	0.9	0.9
cholesterol	8	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
bile acid	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
citrate(3-)	7	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
L-cystine zwitterion	7	4	2	2	2	1479	1.0	0.5	0.5	0.5	0.5
lipid-linked peptidoglycan	7	4	3	1	1	1480	1.0	0.8	0.8	0.8	0.7
L-glutamate(1-)	6	4	3	2	1	1479	1.0	0.6	0.8	0.7	0.7
heme	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
oleate	6	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
S-adenosyl-L-methionine zwitterion	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
biotinate	6	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
UDP-D-glucose(2-)	6	3	2	5	1	1477	1.0	0.3	0.7	0.4	0.4
cobalamin	6	3	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
acetyl-CoA(4-)	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
iron(3+)	5	2	0	0	2	1483	1.0	-	0.0	0.0	NaN
phosphoglycerate	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
hydrogencarbonate	5	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
2-oxoglutarate(2-)	5	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
maltose	5	2	1	2	1	1481	1.0	0.3	0.5	0.4	0.4
sucrose	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
urate salt	5	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
glutathionate(1-)	5	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
1-4-butanediammonium	5	3	0	1	3	1481	1.0	0.0	0.0	0.0	0.0

Continued on next page

Table 51: TooT-Triplet (offline, manhattan) results for Fold 3 of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
copper(2+)	5	3	1	2	2	1480	1.0	0.3	0.3	0.3	0.3
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
NAD	4	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
formate	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
fluoride	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
L-argininium(1+)	4	2	0	1	2	1482	1.0	0.0	0.0	0.0	0.0
L-proline zwitterion	4	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
fatty acyl-CoA	4	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
cadmium cation	3	2	1	4	1	1479	1.0	0.2	0.5	0.3	0.3
L-arabinose	3	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
long-chain fatty acid	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid zwitterion	3	2	0	0	2	1483	1.0	-	0.0	0.0	NaN
N-ribosylnicotinamide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphate ion	3	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
mercury cation	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
GDP-fucose	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
pyrimidine nucleotide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
bacteriocin	3	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
glycine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-histidine zwitterion	3	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
L-methionine zwitterion	3	1	0	3	1	1481	1.0	0.0	0.0	0.0	0.0
D-methionine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
boric acid	3	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
prostaglandin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
teichoic acid	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
xanthine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
uridine	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
L-ascorbate	2	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
maltodextrin	2	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
folate(2-)	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
molybdate	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
N-retinylidene phos- phatidylethanolamine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gluconate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
tungstate	2	1	1	5	0	1479	1.0	0.2	1.0	0.3	0.4
sn-glycerol 3-phosphate(2-)	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuraminate(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	0	1	1484	1.0	-	0.0	0.0	NaN
Overall	2970	1485	1309	-	-	-	0.998	0.881	0.729	0.697	0.860

The table presents the Cross-validation results of UniProt-SPEC-100, Fold 2, using TooT-Triplet (offline) trained and tested using the Manhattan distance metric and a K-NN classifier (K=1). The overall number for TP presents the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric. The significant digits presented for these metrics correspond to the Test set size.

E.2.3 TooT-Triplet Proton Excluded Results

Table 52: SPEC results for triplet offline without proton Fold 1

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	817	409	386	35	23	1041	0.961	0.917	0.944	0.930	0.903
calcium(2+)	576	289	268	36	21	1160	0.962	0.882	0.927	0.904	0.880
potassium(1+)	484	242	210	19	32	1224	0.966	0.917	0.868	0.892	0.872
chloride	295	147	122	18	25	1320	0.971	0.871	0.830	0.850	0.834
sodium(1+)	178	88	70	14	18	1383	0.98	0.83	0.80	0.81	0.80
zinc(2+)	53	27	25	6	2	1452	0.99	0.81	0.93	0.86	0.86
sulfate	44	23	13	9	10	1453	0.99	0.59	0.56	0.58	0.57
ammonium	35	17	17	3	0	1465	1.00	0.85	1.00	0.92	0.92
water	29	14	14	0	0	1471	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	24	13	9	4	4	1468	0.99	0.69	0.69	0.69	0.69
protein	18	10	7	2	3	1473	1.00	0.78	0.70	0.74	0.74
glutathione derivative	18	9	9	0	0	1476	1.0	1.0	1.0	1.0	1.0
retinol	18	9	8	1	1	1475	1.0	0.9	0.9	0.9	0.9
nitrate	18	8	4	1	4	1476	1.0	0.8	0.5	0.6	0.6
iron(2+)	16	8	7	3	1	1474	1.0	0.7	0.9	0.8	0.8
pyruvate	14	6	6	1	0	1478	1.0	0.9	1.0	0.9	0.9
ATP(4-)	12	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
copper(1+)	11	6	6	4	0	1475	1.0	0.6	1.0	0.8	0.8
urea	10	6	4	0	2	1479	1.0	1.0	0.7	0.8	0.8
dehydroascorbide(1-)	9	5	4	0	1	1480	1.0	1.0	0.8	0.9	0.9
choline	9	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
cholesterol	8	4	3	1	1	1480	1.0	0.8	0.8	0.8	0.7
L-ornithinium(1+)	8	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
lipid-linked peptidoglycan	8	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
L-cystine zwitterion	8	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
L-glutamate(1-)	7	3	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
citrate(3-)	7	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
S-adenosyl-L-methionine zwitterion	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
glutathionate(1-)	6	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
1-4-butanediammonium	6	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
UDP-D-glucose(2-)	6	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
bile acid	6	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
heme	6	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
oleate	6	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
cobalamin	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
iron(3+)	5	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
maltose	5	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
2-oxoglutarate(2-)	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
acetyl-CoA(4-)	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
biotinate	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
urate salt	5	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
copper(2+)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
N-ribosylnicotinamide	4	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
phosphate ion	4	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
mercury cation	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
fluoride	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
L-proline zwitterion	4	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
NAD	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-argininium(1+)	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
hydrogencarbonate	4	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
phosphoglycerate	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
sucrose	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
bacteriocin	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
maltodextrin	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
prostaglandin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-methionine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-ascorbate	3	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
glycine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0

Continued on next page

Table 52: SPEC results for triplet offline without proton Fold 1

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
L-histidine zwitterion	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
pyrimidine nucleotide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
zwitterion											
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
GDP-fucose	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
cadmium cation	3	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
long-chain fatty acid	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
neuramate(2-)											
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
N-(4-aminobenzoyl)-L-	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glutamate											
glycine betaine	2	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
arsenate ion	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
xanthine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
flavin adenine dinucleotide	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
sn-glycerol 3-phosphate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
)											
tungstate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
gluconate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	3	0	1481	1.0	0.2	1.0	0.4	0.5
N-retinylidene phos-	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phatidylethanolamine											
molybdate	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
D-methionine zwitterion	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
boric acid	2	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
L-arabinose	2	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
uridine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
Overall	2970	1485	1300	-	-	-	0.997	0.875	0.718	0.700	0.852

The table presents the Cross-validation results of the ProtonEx dataset, Fold 1, using TooT-Triplet (offline) trained and tested using the Euclidean distance metric and a K-NN classifier (K=1). The overall number for TP presents the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric. The significant digits presented for these metrics correspond to the Test set size.

Table 53: SPEC results for triplet offline without proton Fold 2

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	818	409	385	19	24	1057	0.971	0.953	0.941	0.947	0.927
calcium(2+)	577	288	280	12	8	1185	0.987	0.959	0.972	0.966	0.957
potassium(1+)	484	242	230	19	12	1224	0.979	0.924	0.950	0.937	0.924
chloride	294	147	140	15	7	1323	0.985	0.903	0.952	0.927	0.919
sodium(1+)	177	89	78	8	11	1388	0.99	0.91	0.88	0.89	0.89
zinc(2+)	53	26	23	2	3	1457	1.00	0.92	0.89	0.90	0.90
sulfate	45	22	17	3	5	1460	0.99	0.85	0.77	0.81	0.81
ammonium	35	18	18	2	0	1465	1.00	0.90	1.00	0.95	0.95
water	29	15	15	0	0	1470	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	25	12	9	4	3	1469	0.99	0.69	0.75	0.72	0.72
protein	19	9	6	0	3	1476	1.0	1.0	0.7	0.8	0.8
retinol	18	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
glutathione derivative	18	9	9	0	0	1476	1.0	1.0	1.0	1.0	1.0
nitrate	17	9	8	3	1	1473	1.0	0.7	0.9	0.8	0.8
iron(2+)	16	8	7	2	1	1475	1.0	0.8	0.9	0.8	0.8
pyruvate	13	7	6	1	1	1477	1.0	0.9	0.9	0.9	0.9

Continued on next page

Table 53: SPEC results for triplet offline without proton Fold 2

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
ATP(4-)	12	6	5	0	1	1479	1.0	1.0	0.8	0.9	0.9
copper(1+)	11	5	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
urea	11	5	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
dehydroascorbide(1-)	9	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
choline	8	4	3	0	1	1481	1.0	1.0	0.8	0.9	0.9
cholesterol	8	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
L-ornithinium(1+)	8	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
citrate(3-)	7	4	2	0	2	1481	1.0	1.0	0.5	0.7	0.7
bile acid	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
L-cystine zwitterion	7	4	0	1	4	1480	1.0	0.0	0.0	0.0	0.0
lipid-linked peptidoglycan	7	4	3	0	1	1481	1.0	1.0	0.8	0.9	0.9
cobalamin	6	3	0	0	3	1482	1.0	NaN	0.0	0.0	NaN
oleate	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
heme	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
S-adenosyl-L-methionine zwitterion	6	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
UDP-D-glucose(2-)	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
L-glutamate(1-)	6	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
biotinate	6	3	2	3	1	1479	1.0	0.4	0.7	0.5	0.5
urate salt	5	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
acetyl-CoA(4-)	5	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
hydrogencarbonate	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
iron(3+)	5	3	0	2	3	1480	1.0	0.0	0.0	0.0	0.0
phosphoglycerate	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
maltose	5	3	0	2	3	1480	1.0	0.0	0.0	0.0	0.0
2-oxoglutarate(2-)	5	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
3'-phosphonato-5'-adenylyl sulfate(4-)	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
sucrose	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
glutathionate(1-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
copper(2+)	5	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
1-4-butanediammonium	5	3	1	3	2	1479	1.0	0.2	0.3	0.3	0.3
UDP-N-acetyl-alpha-D-glucosamine(2-)	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
formate	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
acetylcholine	4	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
fatty acyl-CoA	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-proline zwitterion	4	2	2	4	0	1479	1.0	0.3	1.0	0.5	0.6
NAD	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
L-argininium(1+)	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-arabinose	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
prostaglandin	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
D-methionine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
boric acid	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-methionine zwitterion	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
L-histidine zwitterion	3	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
glycine zwitterion	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
bacteriocin	3	2	2	4	0	1479	1.0	0.3	1.0	0.5	0.6
long-chain fatty acid	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
N-ribosylnicotinamide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphate ion	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
mercury cation	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
GDP-fucose	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
cadmium cation	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
pyrimidine nucleotide	3	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuraminase(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0

Continued on next page

Table 53: SPEC results for triplet offline without proton Fold 2

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
glycine betaine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
xanthine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
molybdate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
sn-glycerol 3-phosphate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
tungstate	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
gluconate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
N-retinylidene phos- phatidylethanolamine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
maltodextrin	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-ascorbate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
uridine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
Overall	2970	1485	1361	-	-	-	0.998	0.916	0.782	0.765	0.901

The table presents the Cross-validation results of the ProtonEx dataset, Fold 2, using TooT-Triplet (offline) trained and tested using the Euclidean distance metric and a K-NN classifier (K=1). The overall number for TP presents the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric. The significant digits presented for these metrics correspond to the Test set size.

Table 54: SPEC results for triplet offline without proton Fold 3

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	818	408	391	25	17	1052	0.972	0.940	0.958	0.949	0.930
calcium(2+)	577	288	267	23	21	1174	0.970	0.921	0.927	0.924	0.905
potassium(1+)	484	242	221	16	21	1227	0.975	0.932	0.913	0.923	0.908
chloride	294	148	128	9	20	1328	0.980	0.934	0.865	0.898	0.888
sodium(1+)	177	89	84	9	5	1387	0.99	0.90	0.94	0.92	0.92
zinc(2+)	53	27	23	1	4	1457	1.00	0.96	0.85	0.90	0.90
sulfate	45	22	15	4	7	1459	0.99	0.79	0.68	0.73	0.73
ammonium	35	17	16	1	1	1467	1.00	0.94	0.94	0.94	0.94
water	29	14	14	1	0	1470	1.00	0.93	1.00	0.97	0.97
riboflavin(1-)	25	12	5	1	7	1472	0.99	0.83	0.42	0.56	0.59
protein	19	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
retinol	18	9	9	1	0	1475	1.0	0.9	1.0	0.9	0.9
glutathione derivative	18	9	8	4	1	1472	1.0	0.7	0.9	0.8	0.8
nitrate	17	9	5	3	4	1473	1.0	0.6	0.6	0.6	0.6
iron(2+)	16	8	8	4	0	1473	1.0	0.7	1.0	0.8	0.8
pyruvate	13	7	6	0	1	1478	1.0	1.0	0.9	0.9	0.9
ATP(4-)	12	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
copper(1+)	11	6	4	4	2	1475	1.0	0.5	0.7	0.6	0.6
urea	11	5	3	1	2	1479	1.0	0.8	0.6	0.7	0.7
dehydroascorbide(1-)	9	5	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
choline	8	5	4	1	1	1479	1.0	0.8	0.8	0.8	0.8
cholesterol	8	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
L-ornithinium(1+)	8	4	2	0	2	1481	1.0	1.0	0.5	0.7	0.7
citrate(3-)	7	3	2	1	1	1481	1.0	0.7	0.7	0.7	0.7
bile acid	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
L-cystine zwitterion	7	4	3	2	1	1479	1.0	0.6	0.8	0.7	0.7
lipid-linked peptidoglycan	7	4	4	2	0	1479	1.0	0.7	1.0	0.8	0.8
cobalamin	6	3	3	3	0	1479	1.0	0.5	1.0	0.7	0.7
oleate	6	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
heme	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
S-adenosyl-L-methionine zwitterion	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6

Continued on next page

Table 54: SPEC results for triplet offline without proton Fold 3

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
UDP-D-glucose(2-)	6	3	2	4	1	1478	1.0	0.3	0.7	0.4	0.5
L-glutamate(1-)	6	4	3	2	1	1479	1.0	0.6	0.8	0.7	0.7
biotinate	6	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
urate salt	5	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
acetyl-CoA(4-)	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
hydrogencarbonate	5	2	1	2	1	1481	1.0	0.3	0.5	0.4	0.4
iron(3+)	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphoglycerate	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
maltose	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
2-oxoglutarate(2-)	5	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
sucrose	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
glutathionate(1-)	5	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
copper(2+)	5	3	1	3	2	1479	1.0	0.2	0.3	0.3	0.3
1-4-butanediammonium	5	3	1	2	2	1480	1.0	0.3	0.3	0.3	0.3
UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
fatty acyl-CoA	4	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-proline zwitterion	4	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
NAD	4	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
L-argininium(1+)	4	2	0	1	2	1482	1.0	0.0	0.0	0.0	0.0
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-arabinose	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
prostaglandin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
D-methionine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
boric acid	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
L-methionine zwitterion	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
L-histidine zwitterion	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
glycine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
bacteriocin	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
long-chain fatty acid	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid zwitterion	3	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
N-ribosylnicotinamide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphate ion	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
mercury cation	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
GDP-fucose	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
cadmium cation	3	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
pyrimidine nucleotide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuramate(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
xanthine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
molybdate	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
sn-glycerol 3-phosphate(2-)	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
tungstate	2	1	1	7	0	1477	1.0	0.1	1.0	0.2	0.4
gluconate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0

Continued on next page

Table 54: SPEC results for triplet offline without proton Fold 3

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
N-retinylidene phos- phatidylethanolamine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
maltodextrin	2	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
L-ascorbate	2	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
uridine	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
Overall	2970	1485	1332	-	-	-	0.998	0.897	0.754	0.722	0.878

The table presents the Cross-validation results of the ProtonEx dataset, Fold 3, using TooT-Triplet (offline) trained and tested using the Euclidean distance metric and a K-NN classifier (K=1). The overall number for TP presents the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric. The significant digits presented for these metrics correspond to the Test set size.

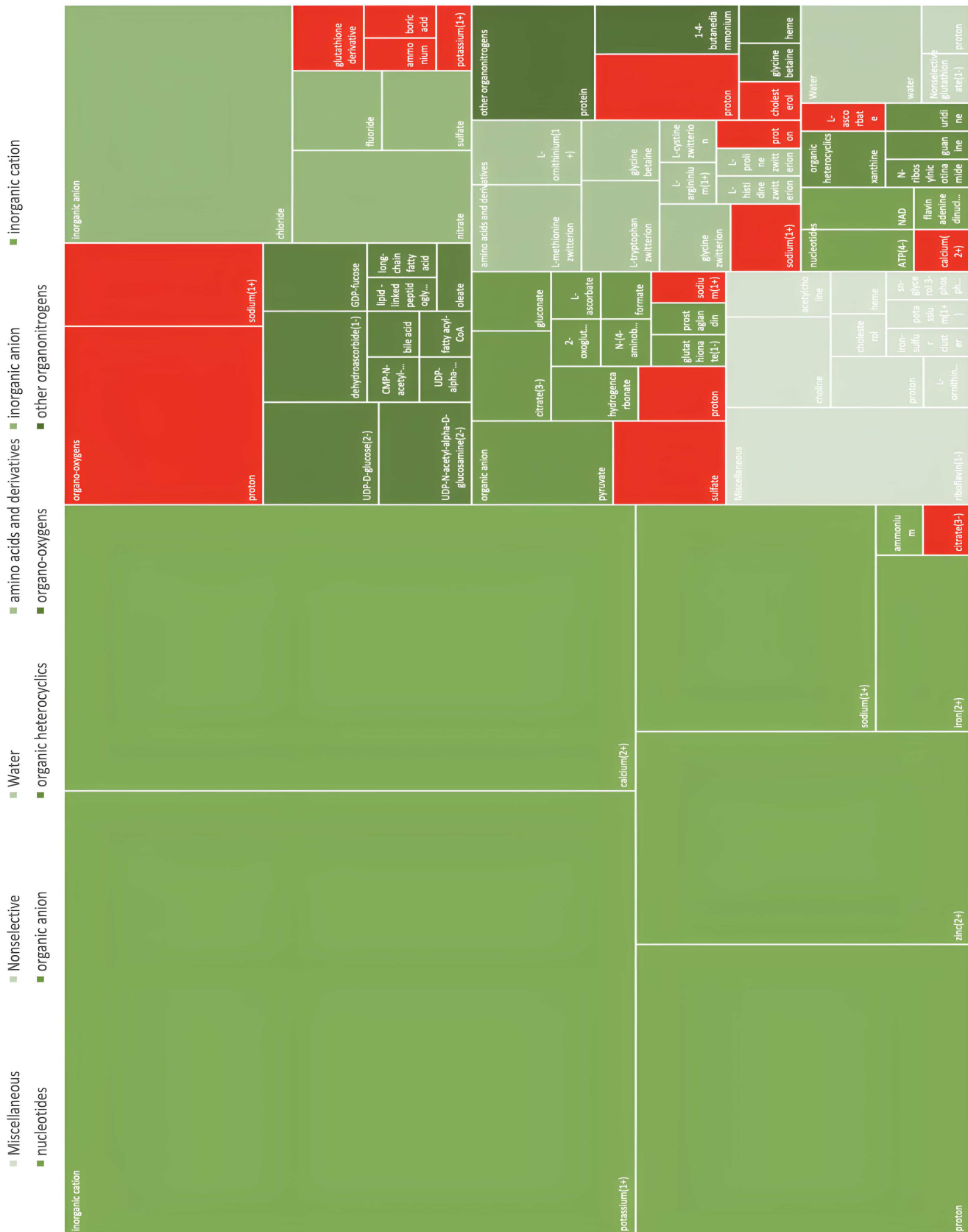


Figure 48: TooT-SC dataset results using TooT-TripleT

Figure presents a treemap comparing the results of the TooT-SC dataset with TooT-TripleT. Each shade of green represents a class in TooT-SC, subdivided into multiple subclasses in TooT-TripleT-SPEC. Red tiles indicate instances where the two models assigned inconsistent labels. The size of each tile is proportional to the sample size in the test set.

E.3 Other method results

E.3.1 Baseline Results

Table 55: Results for UniProt-SPEC-100 hold-out test set using Baseline 1

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	817	409	371	30	38	1046	0.954	0.925	0.907	0.916	0.885
calcium(2+)	576	289	259	41	30	1155	0.952	0.863	0.896	0.879	0.850
potassium(1+)	484	242	208	39	34	1204	0.951	0.842	0.860	0.851	0.821
chloride	295	147	115	27	32	1311	0.960	0.810	0.782	0.796	0.774
sodium(1+)	178	88	74	11	14	1386	0.98	0.87	0.84	0.85	0.85
zinc(2+)	53	27	24	7	3	1451	0.99	0.77	0.89	0.83	0.83
sulfate	44	23	10	9	13	1453	0.98	0.53	0.43	0.48	0.47
ammonium	35	17	16	2	1	1466	1.00	0.89	0.94	0.91	0.91
water	29	14	14	1	0	1470	1.00	0.93	1.00	0.97	0.97
riboflavin(1-)	24	13	9	1	4	1471	1.00	0.90	0.69	0.78	0.79
protein	18	10	8	2	2	1473	1.00	0.80	0.80	0.80	0.80
glutathione derivative	18	9	9	1	0	1475	1.0	0.9	1.0	0.9	0.9
retinol	18	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
nitrate	18	8	5	1	3	1476	1.0	0.8	0.6	0.7	0.7
iron(2+)	16	8	7	2	1	1475	1.0	0.8	0.9	0.8	0.8
pyruvate	14	6	6	1	0	1478	1.0	0.9	1.0	0.9	0.9
ATP(4-)	12	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
copper(1+)	11	6	6	1	0	1478	1.0	0.9	1.0	0.9	0.9
urea	10	6	5	0	1	1479	1.0	1.0	0.8	0.9	0.9
dehydroascorbide(1-)	9	5	4	0	1	1480	1.0	1.0	0.8	0.9	0.9
choline	9	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
cholesterol	8	4	2	1	2	1480	1.0	0.7	0.5	0.6	0.6
L-ornithinium(1+)	8	4	3	1	1	1480	1.0	0.8	0.8	0.8	0.7
lipid-linked peptidoglycan	8	3	2	4	1	1478	1.0	0.3	0.7	0.4	0.5
L-cystine zwitterion	8	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
L-glutamate(1-)	7	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
citrate(3-)	7	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
S-adenosyl-L-methionine zwitterion	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
glutathionate(1-)	6	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
1-4-butanediammonium	6	2	0	1	2	1482	1.0	0.0	0.0	0.0	0.0
UDP-D-glucose(2-)	6	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
bile acid	6	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
heme	6	3	3	3	0	1479	1.0	0.5	1.0	0.7	0.7
oleate	6	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
cobalamin	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
iron(3+)	5	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
maltose	5	2	1	2	1	1481	1.0	0.3	0.5	0.4	0.4
2-oxoglutarate(2-)	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
acetyl-CoA(4-)	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
biotinate	5	3	1	1	2	1481	1.0	0.5	0.3	0.4	0.4
urate salt	5	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
copper(2+)	5	3	1	1	2	1481	1.0	0.5	0.3	0.4	0.4
N-ribosylnicotinamide	4	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
phosphate ion	4	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
mercury cation	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-proline zwitterion	4	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
NAD	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-argininium(1+)	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
hydrogencarbonate	4	3	2	1	1	1481	1.0	0.7	0.7	0.7	0.7
phosphoglycerate	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
sucrose	4	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9

Continued on next page

Table 55: Results for UniProt-SPEC-100 hold-out test set using Baseline 1

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
bacteriocin	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
maltodextrin	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
prostaglandin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-methionine zwitterion	3	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
L-ascorbate	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-histidine zwitterion	3	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
pyrimidine nucleotide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid zwitterion	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
GDP-fucose	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
cadmium cation	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
long-chain fatty acid	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuraminate(2-)	2	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
c-GMP-AMP(2-)	2	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
L-tryptophan zwitterion	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
arsenate ion	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
xanthine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
sn-glycerol 3-phosphate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
tungstate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
gluconate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	3	0	1481	1.0	0.2	1.0	0.4	0.5
N-retinylidene phos-phatidylethanolamine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
molybdate	2	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
D-methionine zwitterion	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
boric acid	2	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
L-arabinose	2	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
uridine	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
Overall	2970	1485	1264	-	-	-	0.997	0.851	0.705	0.672	0.824

Baseline 1 includes the frozen Prot-BERT-BFD representations followed by a K-NN classifier trained and tested using the Manhattan distance metric. The columns Train and Test show the number of samples for each substrate class in the training and test sets. TP, FP, FN, and TN columns show the number of True Positives, False Positives, False Negatives, and True Negatives correspondingly for each class in the results of TooT- Triplet predictions. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric.

Table 56: Results for UniProt-SPEC-100 hold-out test set using Baseline 2

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	817	409	374	27	35	1049	0.958	0.933	0.914	0.923	0.895
calcium(2+)	576	289	260	43	29	1153	0.952	0.858	0.900	0.878	0.848
potassium(1+)	484	242	207	34	35	1209	0.954	0.859	0.855	0.857	0.829
chloride	295	147	115	33	32	1305	0.956	0.777	0.782	0.780	0.755
sodium(1+)	178	88	75	15	13	1382	0.98	0.83	0.85	0.84	0.83
zinc(2+)	53	27	24	5	3	1453	0.99	0.83	0.89	0.86	0.85
sulfate	44	23	9	9	14	1453	0.98	0.50	0.39	0.44	0.43
ammonium	35	17	16	2	1	1466	1.00	0.89	0.94	0.91	0.91
water	29	14	14	0	0	1471	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	24	13	9	1	4	1471	1.00	0.90	0.69	0.78	0.79

Continued on next page

Table 56: Results for UniProt-SPEC-100 hold-out test set using Baseline 2

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
protein	18	10	8	2	2	1473	1.00	0.80	0.80	0.80	0.80
glutathione derivative	18	9	9	1	0	1475	1.0	0.9	1.0	0.9	0.9
retinol	18	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
nitrate	18	8	5	1	3	1476	1.0	0.8	0.6	0.7	0.7
iron(2+)	16	8	7	1	1	1476	1.0	0.9	0.9	0.9	0.9
pyruvate	14	6	6	1	0	1478	1.0	0.9	1.0	0.9	0.9
ATP(4-)	12	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
copper(1+)	11	6	6	1	0	1478	1.0	0.9	1.0	0.9	0.9
urea	10	6	5	0	1	1479	1.0	1.0	0.8	0.9	0.9
dehydroascorbide(1-)	9	5	4	0	1	1480	1.0	1.0	0.8	0.9	0.9
choline	9	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
cholesterol	8	4	2	1	2	1480	1.0	0.7	0.5	0.6	0.6
L-ornithinium(1+)	8	4	3	1	1	1480	1.0	0.8	0.8	0.8	0.7
lipid-linked peptidoglycan	8	3	2	4	1	1478	1.0	0.3	0.7	0.4	0.5
L-cystine zwitterion	8	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
L-glutamate(1-)	7	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
citrate(3-)	7	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
S-adenosyl-L-methionine zwitterion	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
glutathionate(1-)	6	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
1-4-butanediammonium	6	2	0	1	2	1482	1.0	0.0	0.0	0.0	0.0
UDP-D-glucose(2-)	6	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
bile acid	6	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
heme	6	3	3	3	0	1479	1.0	0.5	1.0	0.7	0.7
oleate	6	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
cobalamin	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
iron(3+)	5	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
maltose	5	2	1	2	1	1481	1.0	0.3	0.5	0.4	0.4
2-oxoglutarate(2-)	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
acetyl-CoA(4-)	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
biotinate	5	3	1	1	2	1481	1.0	0.5	0.3	0.4	0.4
urate salt	5	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
copper(2+)	5	3	1	1	2	1481	1.0	0.5	0.3	0.4	0.4
N-ribosylnicotinamide	4	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
phosphate ion	4	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
mercury cation	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-proline zwitterion	4	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
NAD	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-argininium(1+)	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
hydrogencarbonate	4	3	2	1	1	1481	1.0	0.7	0.7	0.7	0.7
phosphoglycerate	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
sucrose	4	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
bacteriocin	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
maltodextrin	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
prostaglandin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-methionine zwitterion	3	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
L-ascorbate	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine zwitterion	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
L-histidine zwitterion	3	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
pyrimidine nucleotide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid zwitterion	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
GDP-fucose	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
cadmium cation	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
long-chain fatty acid	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuraminase(2-)	2	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6

Continued on next page

Table 56: Results for UniProt-SPEC-100 hold-out test set using Baseline 2

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
arsenate ion	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
xanthine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
sn-glycerol 3-phosphate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
tungstate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
gluconate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
N-retinylidene phos-phatidylethanolamine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
molybdate	2	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
D-methionine zwitterion	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
boric acid	2	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
L-arabinose	2	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
uridine	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
Overall	2970	1485	1267	-	-	-	0.997	0.853	0.704	0.672	0.824

Baseline 2 includes the frozen Prot-BERT-BFD representations followed by a K-NN classifier trained and tested using the Euclidean distance metric. The columns Train and Test show the number of samples for each substrate class in the training and test sets. TP, FP, FN, and TN columns show the number of True Positives, False Positives, False Negatives, and True Negatives correspondingly for each class in the results of TooT- Triplet predictions. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric.

Table 57: Results for UniProt-SPEC-100 hold-out test set using Baseline 3

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	817	409	376	34	33	1042	0.955	0.917	0.919	0.918	0.887
calcium(2+)	576	289	263	48	26	1148	0.950	0.846	0.910	0.877	0.846
potassium(1+)	484	242	211	30	31	1213	0.959	0.876	0.872	0.874	0.849
chloride	295	147	114	30	33	1308	0.958	0.792	0.776	0.784	0.760
sodium(1+)	178	88	74	11	14	1386	0.98	0.87	0.84	0.85	0.85
zinc(2+)	53	27	24	1	3	1457	1.00	0.96	0.89	0.92	0.92
sulfate	44	23	9	7	14	1455	0.99	0.56	0.39	0.46	0.46
ammonium	35	17	15	2	2	1466	1.00	0.88	0.88	0.88	0.88
water	29	14	14	1	0	1470	1.00	0.93	1.00	0.97	0.97
riboflavin(1-)	24	13	9	2	4	1470	1.00	0.82	0.69	0.75	0.75
protein	18	10	8	3	2	1472	1.00	0.73	0.80	0.76	0.76
glutathione derivative	18	9	9	1	0	1475	1.0	0.9	1.0	0.9	0.9
retinol	18	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
nitrate	18	8	5	0	3	1477	1.0	1.0	0.6	0.8	0.8
iron(2+)	16	8	7	0	1	1477	1.0	1.0	0.9	0.9	0.9
pyruvate	14	6	6	1	0	1478	1.0	0.9	1.0	0.9	0.9
ATP(4-)	12	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
copper(1+)	11	6	6	1	0	1478	1.0	0.9	1.0	0.9	0.9
urea	10	6	5	0	1	1479	1.0	1.0	0.8	0.9	0.9
dehydroascorbide(1-)	9	5	4	0	1	1480	1.0	1.0	0.8	0.9	0.9
choline	9	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
cholesterol	8	4	2	1	2	1480	1.0	0.7	0.5	0.6	0.6
L-ornithinium(1+)	8	4	3	1	1	1480	1.0	0.8	0.8	0.8	0.7
lipid-linked peptidoglycan	8	3	2	1	1	1481	1.0	0.7	0.7	0.7	0.7
L-cystine zwitterion	8	3	2	1	1	1481	1.0	0.7	0.7	0.7	0.7
L-glutamate(1-)	7	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
citrate(3-)	7	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9

Continued on next page

Table 57: Results for UniProt-SPEC-100 hold-out test set using Baseline 3

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
S-adenosyl-L-methionine zwitterion	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
glutathionate(1-)	6	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
1-4-butanediammonium	6	2	0	2	2	1481	1.0	0.0	0.0	0.0	0.0
UDP-D-glucose(2-)	6	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
bile acid	6	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
heme	6	3	3	3	0	1479	1.0	0.5	1.0	0.7	0.7
oleate	6	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
cobalamin	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
iron(3+)	5	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
maltose	5	2	1	2	1	1481	1.0	0.3	0.5	0.4	0.4
2-oxoglutarate(2-)	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
acetyl-CoA(4-)	5	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
biotinate	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
urate salt	5	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
copper(2+)	5	3	1	1	2	1481	1.0	0.5	0.3	0.4	0.4
N-ribosylnicotinamide	4	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
phosphate ion	4	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
mercury cation	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-proline zwitterion	4	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
NAD	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-argininium(1+)	4	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
hydrogencarbonate	4	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
phosphoglycerate	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
sucrose	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
bacteriocin	3	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
maltodextrin	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
prostaglandin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-methionine zwitterion	3	1	1	3	0	1481	1.0	0.2	1.0	0.4	0.5
L-ascorbate	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-histidine zwitterion	3	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
pyrimidine nucleotide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid zwitterion	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
GDP-fucose	3	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
cadmium cation	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
long-chain fatty acid	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphonateoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuraminate(2-)	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
arsenate ion	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
xanthine	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
sn-glycerol 3-phosphate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN

Continued on next page

Table 57: Results for UniProt-SPEC-100 hold-out test set using Baseline 3

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
tungstate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
gluconate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
N-retinylidene phos- phatidylethanolamine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
molybdate	2	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
D-methionine zwitterion	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
boric acid	2	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
L-arabinose	2	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
uridine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
Overall	2970	1485	1273	-	-	-	0.997	0.857	0.704	0.682	0.831

Baseline 3 includes the frozen Prot-BERT-BFD representations followed by a K-NN classifier trained and tested using the cosine distance metric. The columns Train and Test show the number of samples for each substrate class in the training and test sets. TP, FP, FN, and TN columns show the number of True Positives, False Positives, False Negatives, and True Negatives correspondingly for each class in the results of TooT- Triplet predictions. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric.

E.3.2 Fine-tuned TooT-SPEC + KNN results

Table 58: Results for UniProt-SPEC-100 hold-out test set using TooT-SPEC and KNN (Manhattan)

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	817	409	406	30	3	1046	0.978	0.931	0.993	0.961	0.946
calcium(2+)	576	289	275	10	14	1186	0.984	0.965	0.952	0.958	0.948
potassium(1+)	484	242	223	9	19	1234	0.981	0.961	0.921	0.941	0.930
chloride	295	147	134	9	13	1329	0.985	0.937	0.912	0.924	0.916
sodium(1+)	178	88	74	7	14	1390	0.99	0.91	0.84	0.88	0.87
zinc(2+)	53	27	26	0	1	1458	1.00	1.00	0.96	0.98	0.98
sulfate	44	23	19	0	4	1462	1.00	1.00	0.83	0.91	0.91
ammonium	35	17	17	2	0	1466	1.00	0.90	1.00	0.94	0.94
water	29	14	14	0	0	1471	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	24	13	11	0	2	1472	1.00	1.00	0.85	0.92	0.92
retinol	18	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
glutathione derivative	18	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
protein	18	10	8	1	2	1474	1.00	0.89	0.80	0.84	0.84
nitrate	18	8	5	0	3	1477	1.0	1.0	0.6	0.8	0.8
iron(2+)	16	8	8	1	0	1476	1.0	0.9	1.0	0.9	0.9
pyruvate	14	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
ATP(4-)	12	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
copper(1+)	11	6	6	3	0	1476	1.0	0.7	1.0	0.8	0.8
urea	10	6	6	1	0	1478	1.0	0.9	1.0	0.9	0.9
choline	9	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
dehydroascorbide(1-)	9	5	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
L-ornithinium(1+)	8	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
lipid-linked peptidoglycan	8	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
L-cystine zwitterion	8	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
cholesterol	8	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
L-glutamate(1-)	7	3	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
citrate(3-)	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
heme	6	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
S-adenosyl-L-methionine zwitterion	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
bile acid	6	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
cobalamin	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
UDP-D-glucose(2-)	6	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
oleate	6	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
glutathionate(1-)	6	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
1-4-butanediammonium	6	2	0	1	2	1482	1.0	0.0	0.0	0.0	0.0
maltose	5	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
acetyl-CoA(4-)	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0

Continued on next page

Table 58: Results for UniProt-SPEC-100 hold-out test set using TooT-SPEC and KNN (Manhattan)

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
2-oxoglutarate(2-)	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
copper(2+)	5	3	0	0	3	1482	1.0	NaN	0.0	0.0	NaN
3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
biotinate	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
urate salt	5	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
iron(3+)	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
N-ribosylnicotinamide phosphate ion	4	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
mercury cation	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphoglycerate	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
sucrose	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
NAD	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-argininium(1+)	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-proline zwitterion	4	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
hydrogencarbonate	4	3	3	10	0	1472	1.0	0.2	1.0	0.4	0.5
glycine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid zwitterion	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
long-chain fatty acid	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
GDP-fucose	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
cadmium cation	3	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
pyrimidine nucleotide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-methionine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
prostaglandin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-histidine zwitterion	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
L-ascorbate	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
maltodextrin	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
bacteriocin	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
xanthine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
guanine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-arabinose	2	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
boric acid	2	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
uridine	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
D-methionine zwitterion	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
molybdate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
N-retinylidene phos-phatidylethanolamine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gluconate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
tungstate	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
sn-glycerol 3-phosphate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuramate(2-)	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7

Continued on next page

Table 58: Results for UniProt-SPEC-100 hold-out test set using TooT-SPEC and KNN (Manhattan)

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
Overall	2970	1485	1377	-	-	-	0.998	0.927	0.786	0.758	0.914

This table includes the results for a K-NN classifier trained and tested using the Manhattan distance metric using protein representations from the fine-tuned TooT-SPEC. Train and Test columns show the number of samples for each substrate class in the training and test sets. TP, FP, FN, and TN columns show the number of True Positives, False Positives, False Negatives, and True Negatives correspondingly for each class in the results of TooT- Triplet predictions. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric.

Table 59: Results for UniProt-SPEC-100 hold-out test set using TooT-SPEC and KNN (Euclidean)

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	817	409	407	30	2	1046	0.978	0.931	0.995	0.962	0.948
calcium(2+)	576	289	275	10	14	1186	0.984	0.965	0.952	0.958	0.948
potassium(1+)	484	242	223	8	19	1235	0.982	0.965	0.921	0.943	0.932
chloride	295	147	134	10	13	1328	0.985	0.931	0.912	0.921	0.912
sodium(1+)	178	88	74	7	14	1390	0.99	0.91	0.84	0.88	0.87
zinc(2+)	53	27	26	0	1	1458	1.00	1.00	0.96	0.98	0.98
sulfate	44	23	19	0	4	1462	1.00	1.00	0.83	0.91	0.91
ammonium	35	17	17	2	0	1466	1.00	0.90	1.00	0.94	0.94
water	29	14	14	0	0	1471	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	24	13	11	0	2	1472	1.00	1.00	0.85	0.92	0.92
retinol	18	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
glutathione derivative	18	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
protein	18	10	8	1	2	1474	1.00	0.89	0.80	0.84	0.84
nitrate	18	8	5	0	3	1477	1.0	1.0	0.6	0.8	0.8
iron(2+)	16	8	8	1	0	1476	1.0	0.9	1.0	0.9	0.9
pyruvate	14	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
ATP(4-)	12	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
copper(1+)	11	6	6	3	0	1476	1.0	0.7	1.0	0.8	0.8
urea	10	6	6	1	0	1478	1.0	0.9	1.0	0.9	0.9
choline	9	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
dehydroascorbide(1-)	9	5	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
L-ornithinium(1+)	8	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
lipid-linked peptidoglycan	8	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
L-cystine zwitterion	8	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
cholesterol	8	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
L-glutamate(1-)	7	3	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
citrate(3-)	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
heme	6	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
S-adenosyl-L-methionine zwitterion	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
bile acid	6	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
cobalamin	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
UDP-D-glucose(2-)	6	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
oleate	6	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
glutathionate(1-)	6	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
1-4-butanediammonium	6	2	0	1	2	1482	1.0	0.0	0.0	0.0	0.0
maltose	5	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
acetyl-CoA(4-)	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
2-oxoglutarate(2-)	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
copper(2+)	5	3	0	0	3	1482	1.0	NaN	0.0	0.0	NaN
3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
biotinate	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
urate salt	5	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
iron(3+)	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
N-ribosylnicotinamide	4	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN

Continued on next page

Table 59: Results for UniProt-SPEC-100 hold-out test set using TooT-SPEC and KNN (Euclidean)

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
phosphate ion	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
mercury cation	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphoglycerate	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
sucrose	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
NAD	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-argininium(1+)	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-proline zwitterion	4	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
hydrogencarbonate	4	3	3	9	0	1473	1.0	0.2	1.0	0.4	0.5
glycine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid zwitterion	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
long-chain fatty acid	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
GDP-fucose	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
cadmium cation	3	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
pyrimidine nucleotide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-methionine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
prostaglandin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-histidine zwitterion	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
L-ascorbate	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
maltodextrin	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
bacteriocin	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
xanthine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
guanine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-arabinose	2	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
boric acid	2	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
uridine	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
D-methionine zwitterion	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
molybdate	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
N-retinylidene phos- phatidylethanolamine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gluconate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
tungstate	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
sn-glycerol 3-phosphate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta- neuraminate(2-)	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
Overall	2970	1485	1378	-	-	-	0.998	0.928	0.786	0.759	0.915

This table includes the results for a K-NN classifier trained and tested using the Euclidean distance metric using protein representations from the fine-tuned TooT-SPEC. Train and Test columns show the number of samples for each substrate class in the training and test sets. TP, FP, FN, and TN columns show the number of True Positives, False Positives, False Negatives, and True Negatives correspondingly for each class in the results of TooT- Triplet predictions. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric.

Table 60: Results for UniProt-SPEC-100 hold-out test set using TooT-SPEC and KNN(cosine)

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	817	409	406	28	3	1048	0.979	0.935	0.993	0.963	0.949
calcium(2+)	576	289	275	8	14	1188	0.985	0.972	0.952	0.962	0.952
potassium(1+)	484	242	223	9	19	1234	0.981	0.961	0.921	0.941	0.930
chloride	295	147	135	8	12	1330	0.987	0.944	0.918	0.931	0.924
sodium(1+)	178	88	74	7	14	1390	0.99	0.91	0.84	0.88	0.87
zinc(2+)	53	27	26	0	1	1458	1.00	1.00	0.96	0.98	0.98
sulfate	44	23	17	0	6	1462	1.00	1.00	0.74	0.85	0.86
ammonium	35	17	17	2	0	1466	1.00	0.90	1.00	0.94	0.94
water	29	14	14	0	0	1471	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	24	13	11	0	2	1472	1.00	1.00	0.85	0.92	0.92
retinol	18	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
glutathione derivative	18	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
protein	18	10	8	1	2	1474	1.00	0.89	0.80	0.84	0.84
nitrate	18	8	5	0	3	1477	1.0	1.0	0.6	0.8	0.8
iron(2+)	16	8	8	2	0	1475	1.0	0.8	1.0	0.9	0.9
pyruvate	14	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
ATP(4-)	12	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
copper(1+)	11	6	6	3	0	1476	1.0	0.7	1.0	0.8	0.8
urea	10	6	6	1	0	1478	1.0	0.9	1.0	0.9	0.9
choline	9	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
dehydroascorbide(1-)	9	5	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
L-ornithinium(1+)	8	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
lipid-linked peptidoglycan	8	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
L-cystine zwitterion	8	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
cholesterol	8	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
L-glutamate(1-)	7	3	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
citrate(3-)	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
heme	6	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
S-adenosyl-L-methionine	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
zwitterion											
bile acid	6	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
cobalamin	6	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
UDP-D-glucose(2-)	6	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
oleate	6	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
glutathionate(1-)	6	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
1-4-butanediammonium	6	2	0	1	2	1482	1.0	0.0	0.0	0.0	0.0
maltose	5	2	2	4	0	1479	1.0	0.3	1.0	0.5	0.6
acetyl-CoA(4-)	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
2-oxoglutarate(2-)	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
UDP-N-acetyl-alpha-D-glucosamine(2-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
copper(2+)	5	3	0	0	3	1482	1.0	NaN	0.0	0.0	NaN
3'-phosphonato-5'-adenylyl sulfate(4-)	5	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
biotinate	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
urate salt	5	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
iron(3+)	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
N-ribosylnicotinamide	4	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
phosphate ion	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
mercury cation	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphoglycerate	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
sucrose	4	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
NAD	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-argininium(1+)	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-proline zwitterion	4	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
hydrogencarbonate	4	3	3	11	0	1471	1.0	0.2	1.0	0.4	0.5
glycine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid	3	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
zwitterion											
long-chain fatty acid	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0

Continued on next page

Table 60: Results for UniProt-SPEC-100 hold-out test set using TooT-SPEC and KNN(cosine)

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
GDP-fucose	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
cadmium cation	3	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
pyrimidine nucleotide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-methionine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
prostaglandin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-histidine zwitterion	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
L-ascorbate	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
maltodextrin	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
bacteriocin	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
xanthine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
guanine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-arabinose	2	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
boric acid	2	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
uridine	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
D-methionine zwitterion	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
molybdate	2	1	0	3	1	1481	1.0	0.0	0.0	0.0	0.0
N-retinylidene phos- phatidylethanolamine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gluconate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
tungstate	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
sn-glycerol 3-phosphate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuraminate(2-)	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
Overall	2970	1485	1376	-	-	-	0.998	0.927	0.785	0.756	0.913

This table includes the results for a K-NN classifier trained and tested using the cosine distance metric using protein representations from the fine-tuned TooT-SPEC. Train and Test columns show the number of samples for each substrate class in the training and test sets. TP, FP, FN, and TN columns show the number of True Positives, False Positives, False Negatives, and True Negatives correspondingly for each class in the results of TooT- Triplet predictions. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric.

Table 61: Triplet offline results for SC dataset analyzed by GPT4 and CHEBI

Sequence ID	SC label	Triplet prediction	SPEC label	GPT4	CHEBI
P38087	Miscellaneous	L-ornithinium(1+)	L-ornithinium(1+)	True	True
P34711	Miscellaneous	acetylcholine	acetylcholine	True	True
Q16572	Miscellaneous	acetylcholine	acetylcholine	True	True
Q99PE8	Miscellaneous	cholesterol	cholesterol	True	True
Q8IWA5	Miscellaneous	choline	choline	True	True
Q8NCS7	Miscellaneous	choline	choline	True	True
Q94AN2	Miscellaneous	choline	choline	True	True
Q6X893	Miscellaneous	choline	choline	True	True
Q9NP58	Miscellaneous	heme	heme	True	True
O75027	Miscellaneous	iron-sulfur cluster	iron-sulfur cluster	True	False
Q93646	Miscellaneous	potassium(1+)	acetylcholine	True	False
P10870	Miscellaneous	proton	proton	True	False
Q8VZR6	Miscellaneous	proton	proton	True	False
Q9X1G6	Miscellaneous	riboflavin(1-)	riboflavin(1-)	True	True
Q5M614	Miscellaneous	riboflavin(1-)	riboflavin(1-)	True	True
P50726	Miscellaneous	riboflavin(1-)	riboflavin(1-)	True	True
Q08777	Miscellaneous	riboflavin(1-)	riboflavin(1-)	True	True
Q9X2I1	Miscellaneous	riboflavin(1-)	riboflavin(1-)	True	True
Q5M245	Miscellaneous	riboflavin(1-)	riboflavin(1-)	True	True
P0CI36	Miscellaneous	riboflavin(1-)	riboflavin(1-)	True	True
Q9NQ40	Miscellaneous	riboflavin(1-)	riboflavin(1-)	True	True
Q3LFN0	Miscellaneous	riboflavin(1-)	riboflavin(1-)	True	True
G4SDH4	Miscellaneous	riboflavin(1-)	riboflavin(1-)	True	True
P27514	Miscellaneous	sn-glycerol 3-phosphate(2-)	sn-glycerol 3-phosphate(2-)	True	True
P23242	Nonselective	glutathionate(1-)	glutathionate(1-)	True	True
Q60362	Nonselective	proton	proton	True	True
P55064	Water	water	water	True	True
Q41963	Water	water	water	True	True
P06624	Water	water	water	True	True
P60844	Water	water	water	True	True
Q54WT8	Water	water	water	True	True
Q8N8R3	amino acids and derivatives	L-argininium(1+)	L-argininium(1+)	True	False
P0AFT2	amino acids and derivatives	L-cystine zwitterion	L-cystine zwitterion	True	False
P06775	amino acids and derivatives	L-histidine zwitterion	L-histidine zwitterion	True	False
P38734	amino acids and derivatives	L-methionine zwitterion	L-methionine zwitterion	True	False
P39277	amino acids and derivatives	L-methionine zwitterion	L-methionine zwitterion	True	False
P50276	amino acids and derivatives	L-methionine zwitterion	L-methionine zwitterion	True	False
P31547	amino acids and derivatives	L-methionine zwitterion	L-methionine zwitterion	True	False
Q12375	amino acids and derivatives	L-ornithinium(1+)	L-ornithinium(1+)	False	False
Q9Y619	amino acids and derivatives	L-ornithinium(1+)	L-ornithinium(1+)	False	False
Q93XM7	amino acids and derivatives	L-ornithinium(1+)	L-ornithinium(1+)	False	False
P28573	amino acids and derivatives	L-proline zwitterion	L-proline zwitterion	True	False
P48813	amino acids and derivatives	L-tryptophan zwitterion	L-proline zwitterion	True	False
P38967	amino acids and derivatives	L-tryptophan zwitterion	L-tryptophan zwitterion	True	False
P38085	amino acids and derivatives	L-tryptophan zwitterion	L-tryptophan zwitterion	True	False
Q9X4A5	amino acids and derivatives	glycine betaine	glycine betaine	True	True
Q9RR45	amino acids and derivatives	glycine betaine	glycine betaine	True	True

Continued on next page

Table 61: Triplet offline results for SC dataset analyzed by GPT4 and CHEBI

Sequence ID	SC label	Triplet prediction	SPEC label	GPT4	CHEBI
Q07534	amino acids and derivatives	glycine zwitterion	glycine zwitterion	True	False
Q96DW6	amino acids and derivatives	glycine zwitterion	glycine zwitterion	True	False
P64550	amino acids and derivatives	proton	proton	False	False
P0AGE4	amino acids and derivatives	sodium(1+)	sodium(1+)	False	False
P07117	amino acids and derivatives	sodium(1+)	sodium(1+)	False	False
Q03263	inorganic anion	ammonium	boric acid	False	False
P53838	inorganic anion	boric acid	boric acid	False	False
Q96282	inorganic anion	chloride	chloride	True	True
Q9W701	inorganic anion	chloride	chloride	True	True
P35523	inorganic anion	chloride	chloride	True	True
Q95833	inorganic anion	chloride	chloride	True	True
O15247	inorganic anion	chloride	chloride	True	True
Q9R0A1	inorganic anion	chloride	chloride	True	True
Q9NZA1	inorganic anion	chloride	chloride	True	True
Q9BMK9	inorganic anion	chloride	chloride	True	True
O35454	inorganic anion	chloride	chloride	True	True
Q8WQA4	inorganic anion	chloride	chloride	True	True
Q9Y696	inorganic anion	chloride	chloride	True	True
Q9D3A9	inorganic anion	chloride	chloride	True	True
Q9W5A5	inorganic anion	chloride	chloride	True	True
O14764	inorganic anion	chloride	chloride	True	True
Q06393	inorganic anion	chloride	chloride	True	True
Q9C0H2	inorganic anion	chloride	chloride	True	True
Q9NQ90	inorganic anion	chloride	chloride	True	True
G5EBR3	inorganic anion	chloride	chloride	True	True
Q75V66	inorganic anion	chloride	chloride	True	True
Q9HCE9	inorganic anion	chloride	chloride	True	True
A1A5B4	inorganic anion	chloride	chloride	True	True
Q9BYT9	inorganic anion	chloride	chloride	True	True
Q14AT5	inorganic anion	chloride	chloride	True	True
Q08913	inorganic anion	fluoride	fluoride	True	True
P37002	inorganic anion	fluoride	fluoride	True	True
P9WP63	inorganic anion	fluoride	fluoride	True	True
O95255	inorganic anion	glutathione derivative	glutathione derivative	False	False
Q5T3U5	inorganic anion	glutathione derivative	glutathione derivative	False	False
Q05085	inorganic anion	nitrate	nitrate	True	True
Q94JG1	inorganic anion	nitrate	nitrate	True	True
Q9LXH0	inorganic anion	nitrate	nitrate	True	True
Q9M1E2	inorganic anion	nitrate	nitrate	True	True
Q9M817	inorganic anion	nitrate	nitrate	True	True
Q9LFX9	inorganic anion	nitrate	nitrate	True	True
P38750	inorganic anion	nitrate	nitrate	True	True
Q9FGS5	inorganic anion	potassium(1+)	nitrate	False	False
P92946	inorganic anion	sulfate	sulfate	True	True
Q9FY46	inorganic anion	sulfate	sulfate	True	True
P53273	inorganic anion	sulfate	sulfate	True	True
Q9M6N7	inorganic cation	ammonium	ammonium	True	True
Q21121	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q7Z443	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q2EG98	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q9Z1L5	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q8VHW8	inorganic cation	calcium(2+)	calcium(2+)	True	True
A2ARP9	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q24270	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q8NEC5	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q8IQ70	inorganic cation	calcium(2+)	calcium(2+)	True	True
P48995	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q9UBN4	inorganic cation	calcium(2+)	calcium(2+)	True	True

Continued on next page

Table 61: Triplet offline results for SC dataset analyzed by GPT4 and CHEBI

Sequence ID	SC label	Triplet prediction	SPEC label	GPT4	CHEBI
Q13507	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q02641	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q00975	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q94KI8	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q91ZR5	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q80W99	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q8BVN3	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q9BRQ5	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q94AX5	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q09917	inorganic cation	calcium(2+)	calcium(2+)	True	True
E7FKV8	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q8R526	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q96RD7	inorganic cation	calcium(2+)	calcium(2+)	True	True
P84903	inorganic cation	calcium(2+)	calcium(2+)	True	True
P98161	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q8R455	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q18593	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q8BLA8	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q9W0Y6	inorganic cation	calcium(2+)	calcium(2+)	True	True
P19334	inorganic cation	calcium(2+)	calcium(2+)	True	True
D3Z291	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q8TDX9	inorganic cation	calcium(2+)	calcium(2+)	True	True
P41821	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q7Z3S7	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q08289	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q14644	inorganic cation	calcium(2+)	calcium(2+)	True	True
P83094	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q9GZU1	inorganic cation	calcium(2+)	calcium(2+)	True	True
E9Q401	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q95T12	inorganic cation	calcium(2+)	calcium(2+)	True	True
O00555	inorganic cation	calcium(2+)	calcium(2+)	True	True
P50077	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q9LEQ3	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q9U6B8	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q9VJJ7	inorganic cation	calcium(2+)	calcium(2+)	True	True
P48994	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q9Y5S1	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q9EPK8	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q9R186	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q15878	inorganic cation	calcium(2+)	calcium(2+)	True	True
O60840	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q96D31	inorganic cation	calcium(2+)	calcium(2+)	True	True
P38301	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q9U1S7	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q8NE86	inorganic cation	calcium(2+)	calcium(2+)	True	True
P07293	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q13936	inorganic cation	calcium(2+)	calcium(2+)	True	True
P22002	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q99246	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q9P246	inorganic cation	calcium(2+)	calcium(2+)	True	True
G5EF60	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q9BXT2	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q8K424	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q8R4P5	inorganic cation	calcium(2+)	calcium(2+)	True	True
O70578	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q9NY47	inorganic cation	calcium(2+)	calcium(2+)	True	True
P34374	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q09624	inorganic cation	calcium(2+)	calcium(2+)	True	True
Q8BWY7	inorganic cation	citrate(3-)	zinc(2+)	False	False
P47818	inorganic cation	iron(2+)	iron(2+)	True	True
Q09919	inorganic cation	iron(2+)	iron(2+)	True	True
P38310	inorganic cation	iron(2+)	iron(2+)	True	True
P33650	inorganic cation	iron(2+)	iron(2+)	True	True

Continued on next page

Table 61: Triplet offline results for SC dataset analyzed by GPT4 and CHEBI

Sequence ID	SC label	Triplet prediction	SPEC label	GPT4	CHEBI
Q9NP59	inorganic cation	iron(2+)	iron(2+)	True	True
P10566	inorganic cation	iron(2+)	iron(2+)	True	True
P40088	inorganic cation	iron(2+)	iron(2+)	True	True
Q9P6J2	inorganic cation	potassium(1+)	iron(2+)	True	True
Q9DAV9	inorganic cation	potassium(1+)	potassium(1+)	True	True
O35174	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9BQ31	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q8NCM2	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q8TDN2	inorganic cation	potassium(1+)	potassium(1+)	True	True
O95069	inorganic cation	potassium(1+)	potassium(1+)	True	True
O60928	inorganic cation	potassium(1+)	potassium(1+)	True	True
G5EFC3	inorganic cation	potassium(1+)	potassium(1+)	True	True
O88454	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9UGI6	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q19907	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9NA73	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9UK17	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9Y257	inorganic cation	potassium(1+)	potassium(1+)	True	True
P48048	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9NPC2	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q99712	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q8LBL1	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9NA75	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q6VVA6	inorganic cation	potassium(1+)	potassium(1+)	True	True
O17185	inorganic cation	potassium(1+)	potassium(1+)	True	True
AOA086F3E3	inorganic cation	potassium(1+)	potassium(1+)	True	True
S5VBU1	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q57603	inorganic cation	potassium(1+)	potassium(1+)	True	True
P48549	inorganic cation	potassium(1+)	potassium(1+)	True	True
P48051	inorganic cation	potassium(1+)	potassium(1+)	True	True
O54982	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q14B80	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q14003	inorganic cation	potassium(1+)	potassium(1+)	True	True
O95279	inorganic cation	potassium(1+)	potassium(1+)	True	True
O22397	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q16322	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9H3M0	inorganic cation	potassium(1+)	potassium(1+)	True	True
P92960	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9FY75	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q04705	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q8L7Z0	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q16558	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9XFR0	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9S6Z8	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9YDF8	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q98GN8	inorganic cation	potassium(1+)	potassium(1+)	True	True
D9N164	inorganic cation	potassium(1+)	potassium(1+)	True	True
B7U540	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q38898	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q86W47	inorganic cation	potassium(1+)	potassium(1+)	True	True
O15554	inorganic cation	potassium(1+)	potassium(1+)	True	True
O70596	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q94526	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q38998	inorganic cation	potassium(1+)	potassium(1+)	True	True
P12685	inorganic cation	potassium(1+)	potassium(1+)	True	True
P0AGI8	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q18120	inorganic cation	potassium(1+)	potassium(1+)	True	True
P47114	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q39128	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q12809	inorganic cation	potassium(1+)	potassium(1+)	True	True
O54853	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q03611	inorganic cation	potassium(1+)	potassium(1+)	True	True
O95259	inorganic cation	potassium(1+)	potassium(1+)	True	True

Continued on next page

Table 61: Triplet offline results for SC dataset analyzed by GPT4 and CHEBI

Sequence ID	SC label	Triplet prediction	SPEC label	GPT4	CHEBI
Q8TAE7	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q96KK3	inorganic cation	potassium(1+)	potassium(1+)	True	True
P97557	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q8TDN1	inorganic cation	potassium(1+)	potassium(1+)	True	True
P15382	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q2I0M4	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q6ZSA7	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9Y691	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9NPA1	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9Y6J6	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9UIX4	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q03720	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q96T55	inorganic cation	potassium(1+)	potassium(1+)	True	True
P51787	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q02280	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q22271	inorganic cation	potassium(1+)	potassium(1+)	True	True
P15385	inorganic cation	potassium(1+)	potassium(1+)	True	True
P22460	inorganic cation	potassium(1+)	potassium(1+)	True	True
P17970	inorganic cation	potassium(1+)	potassium(1+)	True	True
P08510	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9Z258	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q850M0	inorganic cation	potassium(1+)	potassium(1+)	True	True
P15387	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q15842	inorganic cation	potassium(1+)	potassium(1+)	True	True
O43525	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9NR82	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q12791	inorganic cation	potassium(1+)	potassium(1+)	True	True
K9UJK2	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9BSA9	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9M8S6	inorganic cation	potassium(1+)	potassium(1+)	True	True
P28584	inorganic cation	potassium(1+)	potassium(1+)	True	True
P40310	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q95V25	inorganic cation	potassium(1+)	potassium(1+)	True	True
O14649	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q6VV64	inorganic cation	potassium(1+)	potassium(1+)	True	True
P70604	inorganic cation	potassium(1+)	potassium(1+)	True	True
O43526	inorganic cation	potassium(1+)	potassium(1+)	True	True
P17658	inorganic cation	potassium(1+)	potassium(1+)	True	True
O54852	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9UJ90	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q5M8M9	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q5VT99	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q8WWG9	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9NS61	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q63113	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9H427	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9JJV7	inorganic cation	potassium(1+)	potassium(1+)	True	True
Q9SZY7	inorganic cation	proton	proton	False	True
Q9HBG4	inorganic cation	proton	proton	False	True
P32842	inorganic cation	proton	proton	True	True
P68699	inorganic cation	proton	proton	True	True
P37296	inorganic cation	proton	proton	True	True
P0A769	inorganic cation	proton	proton	True	True
Q9M0Z3	inorganic cation	proton	proton	True	True
P25515	inorganic cation	proton	proton	True	True
Q9LJI5	inorganic cation	proton	proton	True	True
Q13488	inorganic cation	proton	proton	True	True
Q91V37	inorganic cation	proton	proton	False	True
Q9Y487	inorganic cation	proton	proton	True	True
Q54E04	inorganic cation	proton	proton	True	True
P31412	inorganic cation	proton	proton	True	True
Q9LX65	inorganic cation	proton	proton	True	True
P31407	inorganic cation	proton	proton	True	True

Continued on next page

Table 61: Triplet offline results for SC dataset analyzed by GPT4 and CHEBI

Sequence ID	SC label	Triplet prediction	SPEC label	GPT4	CHEBI
P0ABA4	inorganic cation	proton	proton	True	True
Q99766	inorganic cation	proton	proton	True	True
Q3E7B6	inorganic cation	proton	proton	True	True
O15342	inorganic cation	proton	proton	True	True
P22203	inorganic cation	proton	proton	False	True
P43457	inorganic cation	proton	proton	False	True
P06821	inorganic cation	proton	proton	True	True
P61829	inorganic cation	proton	proton	True	True
P00846	inorganic cation	proton	proton	True	True
P41807	inorganic cation	proton	proton	True	True
Q96D96	inorganic cation	proton	proton	True	True
P32366	inorganic cation	proton	proton	True	True
P23968	inorganic cation	proton	proton	True	True
P32563	inorganic cation	proton	proton	True	True
P0ABA6	inorganic cation	proton	proton	True	True
P38925	inorganic cation	proton	proton	False	True
P0ABA0	inorganic cation	proton	proton	True	True
P03928	inorganic cation	proton	proton	True	True
P56385	inorganic cation	proton	proton	True	True
Q9FT52	inorganic cation	proton	proton	True	True
O75964	inorganic cation	proton	proton	True	True
P0ABJ6	inorganic cation	proton	proton	True	True
O75947	inorganic cation	proton	proton	False	True
P81449	inorganic cation	proton	proton	True	True
Q8W4S4	inorganic cation	proton	proton	True	True
Q16515	inorganic cation	sodium(1+)	sodium(1+)	True	True
P51172	inorganic cation	sodium(1+)	sodium(1+)	True	True
Q01118	inorganic cation	sodium(1+)	sodium(1+)	True	True
Q96FT7	inorganic cation	sodium(1+)	sodium(1+)	True	True
Q07699	inorganic cation	sodium(1+)	sodium(1+)	True	True
O01811	inorganic cation	sodium(1+)	sodium(1+)	True	True
P35499	inorganic cation	sodium(1+)	sodium(1+)	True	True
Q09274	inorganic cation	sodium(1+)	sodium(1+)	True	True
Q708S7	inorganic cation	sodium(1+)	sodium(1+)	True	True
Q708S6	inorganic cation	sodium(1+)	sodium(1+)	True	True
Q9Y5Y9	inorganic cation	sodium(1+)	sodium(1+)	True	True
Q9WU39	inorganic cation	sodium(1+)	sodium(1+)	True	True
P08104	inorganic cation	sodium(1+)	sodium(1+)	True	True
Q61180	inorganic cation	sodium(1+)	sodium(1+)	True	True
Q14524	inorganic cation	sodium(1+)	sodium(1+)	True	True
P34886	inorganic cation	sodium(1+)	sodium(1+)	True	True
P24612	inorganic cation	sodium(1+)	sodium(1+)	True	True
P54900	inorganic cation	sodium(1+)	sodium(1+)	True	True
Q8IWT1	inorganic cation	sodium(1+)	sodium(1+)	True	True
Q8BXR5	inorganic cation	sodium(1+)	sodium(1+)	True	True
Q86LG1	inorganic cation	sodium(1+)	sodium(1+)	True	True
Q9VR07	inorganic cation	sodium(1+)	sodium(1+)	True	True
Q13796	inorganic cation	sodium(1+)	sodium(1+)	True	True
P75757	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q9NY26	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q5U1X7	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q9NP94	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q94DG6	inorganic cation	zinc(2+)	zinc(2+)	True	True
P32804	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q12436	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q9EYX5	inorganic cation	zinc(2+)	zinc(2+)	True	True
O94639	inorganic cation	zinc(2+)	zinc(2+)	True	True
P34240	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q9XUC4	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q6XR72	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q12067	inorganic cation	zinc(2+)	zinc(2+)	True	True
P32798	inorganic cation	zinc(2+)	zinc(2+)	True	True
P20107	inorganic cation	zinc(2+)	zinc(2+)	True	True

Continued on next page

Table 61: Triplet offline results for SC dataset analyzed by GPT4 and CHEBI

Sequence ID	SC label	Triplet prediction	SPEC label	GPT4	CHEBI
Q6ZMH5	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q9ULF5	inorganic cation	zinc(2+)	zinc(2+)	True	True
O35149	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q9Y6M5	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q95QW4	inorganic cation	zinc(2+)	zinc(2+)	True	True
P53735	inorganic cation	zinc(2+)	zinc(2+)	True	True
P97441	inorganic cation	zinc(2+)	zinc(2+)	True	True
P40544	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q13433	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q78IQ7	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q8TAD4	inorganic cation	zinc(2+)	zinc(2+)	True	True
P39832	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q8IWU4	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q9ZT63	inorganic cation	zinc(2+)	zinc(2+)	True	True
A8DYE2	inorganic cation	zinc(2+)	zinc(2+)	True	True
Q9M024	nucleotides	ATP(4-)	ATP(4-)	True	False
Q6KCM7	nucleotides	ATP(4-)	ATP(4-)	True	False
Q8RWA5	nucleotides	NAD	NAD	True	True
P39953	nucleotides	NAD	NAD	True	True
P39719	nucleotides	calcium(2+)	flavin adenine dinucleotide	False	False
Q08967	nucleotides	flavin adenine dinucleotide	flavin adenine dinucleotide	True	True
Q41364	organic anion	2-oxoglutarate(2-)	2-oxoglutarate(2-)	True	True
Q8GX78	organic anion	L-ascorbate	L-ascorbate	True	True
P46133	organic anion	N-(4-aminobenzoyl)-L-glutamate	N-(4-aminobenzoyl)-L-glutamate	True	True
P55069	organic anion	citrate(3-)	citrate(3-)	True	True
P32089	organic anion	citrate(3-)	citrate(3-)	True	True
Q9SFB0	organic anion	citrate(3-)	citrate(3-)	True	True
P37327	organic anion	formate	formate	True	True
P0AC96	organic anion	gluconate	gluconate	True	True
Q46892	organic anion	gluconate	gluconate	True	True
O15440	organic anion	glutathionate(1-)	glutathionate(1-)	True	True
P48751	organic anion	hydrogencarbonate	hydrogencarbonate	False	True
P39660	organic anion	hydrogencarbonate	hydrogencarbonate	True	True
Q9EPT5	organic anion	prostaglandin	prostaglandin	True	False
Q8L7Z9	organic anion	proton	L-glutamate(1-)	False	False
Q46839	organic anion	proton	proton	False	False
O95563	organic anion	pyruvate	pyruvate	True	True
Q9Y5U8	organic anion	pyruvate	pyruvate	True	True
Q1EBV7	organic anion	pyruvate	pyruvate	True	True
P53311	organic anion	pyruvate	pyruvate	True	True
P53157	organic anion	pyruvate	pyruvate	True	True
P0AER8	organic anion	sodium(1+)	sodium(1+)	False	False
Q12325	organic anion	sulfate	sulfate	False	False
P38359	organic anion	sulfate	sulfate	False	False
Q86WA9	organic anion	sulfate	sulfate	False	False
P53394	organic anion	sulfate	sulfate	False	False
P77328	organic heterocyclics	L-ascorbate	xanthine	False	False
D2ZZC1	organic heterocyclics	N-ribosylnicotinamide	N-ribosylnicotinamide	True	True
Q46817	organic heterocyclics	guanine	guanine	True	True
Q61672	organic heterocyclics	uridine	uridine	True	True
P0AGM9	organic heterocyclics	xanthine	xanthine	True	True
P67444	organic heterocyclics	xanthine	xanthine	True	True
P78382	organo-oxygens	CMP-N-acetyl-beta-neuraminate(2-)	CMP-N-acetyl-beta-neuraminate(2-)	False	False
Q96A29	organo-oxygens	GDP-fucose	GDP-fucose	False	True
Q84L08	organo-oxygens	GDP-fucose	GDP-fucose	False	True
Q9VHT4	organo-oxygens	GDP-fucose	GDP-fucose	False	True
Q12520	organo-oxygens	UDP-D-glucose(2-)	UDP-D-glucose(2-)	False	False
O64503	organo-oxygens	UDP-D-glucose(2-)	UDP-D-glucose(2-)	False	False
Q94B65	organo-oxygens	UDP-D-glucose(2-)	UDP-D-glucose(2-)	False	False
Q0D9V6	organo-oxygens	UDP-D-glucose(2-)	UDP-D-glucose(2-)	False	False
Q5W7C1	organo-oxygens	UDP-D-glucose(2-)	UDP-D-glucose(2-)	False	False

Continued on next page

Table 61: Triplet offline results for SC dataset analyzed by GPT4 and CHEBI

Sequence ID	SC label	Triplet prediction	SPEC label	GPT4	CHEBI
P40004	organo-oxygens	UDP-N-acetyl-alpha-D-glucosamine(2-)	UDP-N-acetyl-alpha-D-glucosamine(2-)	False	False
Q76EJ3	organo-oxygens	UDP-N-acetyl-alpha-D-glucosamine(2-)	UDP-N-acetyl-alpha-D-glucosamine(2-)	False	False
Q93890	organo-oxygens	UDP-N-acetyl-alpha-D-glucosamine(2-)	UDP-N-acetyl-alpha-D-glucosamine(2-)	False	False
Q18779	organo-oxygens	UDP-N-acetyl-alpha-D-glucosamine(2-)	UDP-N-acetyl-alpha-D-glucosamine(2-)	False	False
Q8RXL8	organo-oxygens	UDP-alpha-D-xylose(2-)	UDP-alpha-D-xylose(2-)	False	False
O95342	organo-oxygens	bile acid	bile acid	False	True
Q9JJZ1	organo-oxygens	dehydroascorbide(1-)	dehydroascorbide(1-)	False	False
Q9UGQ3	organo-oxygens	dehydroascorbide(1-)	dehydroascorbide(1-)	False	False
P11169	organo-oxygens	dehydroascorbide(1-)	dehydroascorbide(1-)	False	False
P11168	organo-oxygens	dehydroascorbide(1-)	dehydroascorbide(1-)	False	False
P33897	organo-oxygens	fatty acyl-CoA	fatty acyl-CoA	False	True
P0AF16	organo-oxygens	lipid-linked peptidoglycan	lipid-linked peptidoglycan	False	True
P10384	organo-oxygens	long-chain fatty acid	long-chain fatty acid	False	True
Q9Y2P4	organo-oxygens	oleate	oleate	False	False
Q9P3U6	organo-oxygens	proton	proton	False	False
Q10PW9	organo-oxygens	proton	proton	False	False
P42833	organo-oxygens	proton	proton	False	False
P54854	organo-oxygens	proton	proton	False	False
Q56ZZ7	organo-oxygens	proton	proton	False	False
P39932	organo-oxygens	proton	proton	False	False
P41036	organo-oxygens	proton	proton	False	False
P0AEP1	organo-oxygens	proton	proton	False	False
P23585	organo-oxygens	proton	proton	False	False
P02920	organo-oxygens	proton	proton	False	False
Q9NRA2	organo-oxygens	proton	proton	False	False
P11551	organo-oxygens	proton	proton	False	False
O23492	organo-oxygens	proton	proton	False	False
P15325	organo-oxygens	proton	proton	False	False
Q12407	organo-oxygens	proton	proton	False	False
A0PJK1	organo-oxygens	sodium(1+)	sodium(1+)	False	False
Q2M3M2	organo-oxygens	sodium(1+)	sodium(1+)	False	False
P31639	organo-oxygens	sodium(1+)	sodium(1+)	False	False
P13866	organo-oxygens	sodium(1+)	sodium(1+)	False	False
Q12908	organo-oxygens	sodium(1+)	sodium(1+)	False	False
Q96EP9	organo-oxygens	sodium(1+)	sodium(1+)	False	False
O08705	organo-oxygens	sodium(1+)	sodium(1+)	False	False
P0AAH4	other organonitrogens	1-4-butanediammonium	1-4-butanediammonium	False	False
P0AAH8	other organonitrogens	1-4-butanediammonium	1-4-butanediammonium	False	False
P0AGH5	other organonitrogens	1-4-butanediammonium	1-4-butanediammonium	False	False
P12428	other organonitrogens	cholesterol	guanine	False	False
P0AGH3	other organonitrogens	glycine betaine	1-4-butanediammonium	False	False
Q9P5N0	other organonitrogens	heme	heme	False	False
P9WGP5	other organonitrogens	protein	protein	False	True
P9WGP3	other organonitrogens	protein	protein	False	True
Q9X1R4	other organonitrogens	protein	protein	False	True
O06446	other organonitrogens	protein	protein	False	True
P0DJP3	other organonitrogens	protein	protein	False	True
P10408	other organonitrogens	protein	protein	False	True
P39276	other organonitrogens	proton	proton	False	False
Q16348	other organonitrogens	proton	proton	False	False
O01840	other organonitrogens	proton	proton	False	False
Q9JIP7	other organonitrogens	proton	proton	False	False

The alignment between GPT-4 analysis and CHEBI ontology terms is represented by colors. Green indicates that both GPT-4 and CHEBI define the TooT-Triplet specific prediction as consistent with the general label in the SC dataset. Red indicates that both GPT-4 and CHEBI ontology find the TooT-Triplet prediction does not match the SC label. Yellow signifies that GPT-4 finds the TooT-Triplet prediction matching the SC label, but CHEBI ontology does not. Orange indicates that GPT-4 analysis finds the TooT-Triplet predictions do not match the SC label, while CHEBI ontology does. It is important to note that both GPT-4 and CHEBI analysis are subject to errors, as discussed in Section 5.9.

E.4.2 Organism analysis

The TooT-Triplet model trained with offline sampling strategy using Manhattan distance metric and a K-NN classifier (K=1) is used to predict the specific substrates for protein sequences in transportomes of 5 model organisms *Escherichia coli K-12 (Es)*, *Saccharomyces cerevisiae (Sc)*, *Drosophila melanogaste (Dm)*, *Homo sapiens (Hs)*, and *Arabidopsis thaliana (At)*. The results of this experiment are presented in Table 62.

Table 62: TooT-Triplet predictions for transportome of five model organisms

Substrate	Es	Sc	Dm	Hs	At	# UniProt- SPEC-100
Proteome	4,391	5,887	13,806	20,609	27,468	
Transportome	894 (20.360%)	702 (11.925%)	891 (6.454%)	1767 (8.574%)	1654 (6.022%)	NaN
proton	321 (7.310%)	219 (3.720%)	209 (1.514%)	291 (1.412%)	370 (1.347%)	1226
potassium(1+)	51 (1.161%)	68 (1.155%)	85 (0.616%)	222 (1.077%)	159 (0.579%)	726
sodium(1+)	50 (1.139%)	19 (0.323%)	92 (0.666%)	144 (0.699%)	28 (0.102%)	266
sulfate	33 (0.752%)	10 (0.170%)	10 (0.072%)	14 (0.068%)	16 (0.058%)	67
maltodextrin	32 (0.729%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	4
chloride	28 (0.638%)	44 (0.747%)	78 (0.565%)	234 (1.135%)	99 (0.360%)	442
maltose	28 (0.638%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	7
L-cystine zwitterion	26 (0.592%)	6 (0.102%)	11 (0.080%)	9 (0.044%)	1 (0.004%)	11
1-4-butanediammonium	26 (0.592%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	1 (0.004%)	8
iron(2+)	23 (0.524%)	10 (0.170%)	4 (0.029%)	13 (0.063%)	19 (0.069%)	24
iron(3+)	20 (0.455%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	7
zinc(2+)	15 (0.342%)	22 (0.374%)	25 (0.181%)	37 (0.180%)	42 (0.153%)	80
cobalamin	15 (0.342%)	0 (0.000%)	0 (0.000%)	1 (0.005%)	2 (0.007%)	9
lipid-linked peptidoglycan	15 (0.342%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	11
protein	14 (0.319%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	3 (0.011%)	28
calcium(2+)	13 (0.296%)	138 (2.344%)	130 (0.942%)	428 (2.077%)	186 (0.677%)	865
riboflavin(1-)	13 (0.296%)	6 (0.102%)	13 (0.094%)	14 (0.068%)	16 (0.058%)	37
L-tryptophan zwitterion	10 (0.228%)	7 (0.119%)	1 (0.007%)	0 (0.000%)	2 (0.007%)	3
L-arabinose	9 (0.205%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	4
bacteriocin	9 (0.205%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	4
water	8 (0.182%)	4 (0.068%)	7 (0.051%)	22 (0.107%)	43 (0.157%)	43
L-histidine zwitterion	8 (0.182%)	2 (0.034%)	1 (0.007%)	2 (0.010%)	0 (0.000%)	4
L-glutamate(1-)	8 (0.182%)	1 (0.017%)	5 (0.036%)	15 (0.073%)	16 (0.058%)	10
long-chain fatty acid	7 (0.159%)	0 (0.000%)	1 (0.007%)	0 (0.000%)	0 (0.000%)	4
gluconate	7 (0.159%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	3
citrate(3-)	6 (0.137%)	3 (0.051%)	6 (0.043%)	8 (0.039%)	12 (0.044%)	10
L-methionine zwitterion	6 (0.137%)	2 (0.034%)	1 (0.007%)	1 (0.005%)	0 (0.000%)	4
guanine	6 (0.137%)	1 (0.017%)	6 (0.043%)	0 (0.000%)	7 (0.025%)	3
glycine betaine	6 (0.137%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	3
pyruvate	5 (0.114%)	16 (0.272%)	34 (0.246%)	34 (0.165%)	43 (0.157%)	20
N-ribosylnicotinamide	5 (0.114%)	7 (0.119%)	4 (0.029%)	1 (0.005%)	11 (0.040%)	5
gamma-aminobutyric acid	5 (0.114%)	2 (0.034%)	2 (0.014%)	3 (0.015%)	35 (0.127%)	5
zwitterion						
glutathionate(1-)	5 (0.114%)	2 (0.034%)	0 (0.000%)	14 (0.068%)	10 (0.036%)	8
formate	5 (0.114%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	6
molybdate	5 (0.114%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	3
biotinate	4 (0.091%)	6 (0.102%)	0 (0.000%)	4 (0.019%)	0 (0.000%)	8
ammonium	4 (0.091%)	5 (0.085%)	4 (0.029%)	6 (0.029%)	14 (0.051%)	52
nitrate	4 (0.091%)	2 (0.034%)	4 (0.029%)	8 (0.039%)	78 (0.284%)	26
phosphate ion	4 (0.091%)	1 (0.017%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	5
fluoride	3 (0.068%)	2 (0.034%)	0 (0.000%)	0 (0.000%)	1 (0.004%)	6
heme	3 (0.068%)	0 (0.000%)	1 (0.007%)	1 (0.005%)	2 (0.007%)	9
lipopolysaccharide	3 (0.068%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	3
xanthine	3 (0.068%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	3
oleate	2 (0.046%)	1 (0.017%)	1 (0.007%)	4 (0.019%)	1 (0.004%)	9
UDP-D-glucose(2-)	2 (0.046%)	1 (0.017%)	1 (0.007%)	1 (0.005%)	8 (0.029%)	9
cadmium cation	2 (0.046%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	2 (0.007%)	5
copper(2+)	2 (0.046%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	2 (0.007%)	8

Continued on next page

Table 62: TooT-Triplet predictions for transportome of five model organisms

Substrate	Es	Sc	Dm	Hs	At	# UniProt- SPEC-100
mercury cation	2 (0.046%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	5
teichoic acid	2 (0.046%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	3
L-ornithinium(1+)	1 (0.023%)	8 (0.136%)	10 (0.072%)	10 (0.049%)	4 (0.015%)	12
sn-glycerol 3-phosphate(2-)	1 (0.023%)	3 (0.051%)	1 (0.007%)	2 (0.010%)	5 (0.018%)	3
urea	1 (0.023%)	2 (0.034%)	3 (0.022%)	10 (0.049%)	15 (0.055%)	16
fatty acyl-CoA	1 (0.023%)	2 (0.034%)	2 (0.014%)	5 (0.024%)	4 (0.015%)	5
bile acid	1 (0.023%)	1 (0.017%)	5 (0.036%)	4 (0.019%)	26 (0.095%)	10
copper(1+)	1 (0.023%)	1 (0.017%)	2 (0.014%)	2 (0.010%)	4 (0.015%)	17
phosphoglycerate	1 (0.023%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	6 (0.022%)	7
D-methionine zwitterion	1 (0.023%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	4
N-(4-aminobenzoyl)-L-glutamate	1 (0.023%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	3
magnesium cation	1 (0.023%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	3
tungstate	1 (0.023%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	3
L-argininium(1+)	0 (0.000%)	11 (0.187%)	4 (0.029%)	11 (0.053%)	65 (0.237%)	6
NAD	0 (0.000%)	9 (0.153%)	15 (0.109%)	11 (0.053%)	13 (0.047%)	6
L-proline zwitterion	0 (0.000%)	9 (0.153%)	4 (0.029%)	8 (0.039%)	6 (0.022%)	6
boric acid	0 (0.000%)	9 (0.153%)	0 (0.000%)	0 (0.000%)	18 (0.066%)	4
UDP-N-acetyl-alpha-D-glucosamine(2-)	0 (0.000%)	5 (0.085%)	7 (0.051%)	15 (0.073%)	70 (0.255%)	8
glutathione derivative	0 (0.000%)	5 (0.085%)	6 (0.043%)	18 (0.087%)	28 (0.102%)	27
flavin adenine dinucleotide	0 (0.000%)	5 (0.085%)	0 (0.000%)	4 (0.019%)	1 (0.004%)	3
phosphonateolpyruvate	0 (0.000%)	4 (0.068%)	3 (0.022%)	10 (0.049%)	24 (0.087%)	3
iron-sulfur cluster	0 (0.000%)	3 (0.051%)	4 (0.029%)	4 (0.019%)	3 (0.011%)	6
S-adenosyl-L-methionine zwitterion	0 (0.000%)	3 (0.051%)	3 (0.022%)	5 (0.024%)	11 (0.040%)	9
urate salt	0 (0.000%)	2 (0.034%)	26 (0.188%)	19 (0.092%)	2 (0.007%)	8
ATP(4-)	0 (0.000%)	2 (0.034%)	3 (0.022%)	5 (0.024%)	12 (0.044%)	18
pyrimidine nucleotide	0 (0.000%)	2 (0.034%)	3 (0.022%)	3 (0.015%)	1 (0.004%)	5
choline	0 (0.000%)	2 (0.034%)	2 (0.014%)	5 (0.024%)	3 (0.011%)	13
retinol	0 (0.000%)	2 (0.034%)	0 (0.000%)	8 (0.039%)	3 (0.011%)	27
glycine zwitterion	0 (0.000%)	2 (0.034%)	0 (0.000%)	2 (0.010%)	0 (0.000%)	4
cholesterol	0 (0.000%)	1 (0.017%)	14 (0.101%)	9 (0.044%)	28 (0.102%)	12
uridine	0 (0.000%)	1 (0.017%)	2 (0.014%)	3 (0.015%)	0 (0.000%)	3
acetyl-CoA(4-)	0 (0.000%)	1 (0.017%)	1 (0.007%)	1 (0.005%)	0 (0.000%)	7
2-oxoglutarate(2-)	0 (0.000%)	0 (0.000%)	8 (0.058%)	9 (0.044%)	5 (0.018%)	7
folate(2-)	0 (0.000%)	0 (0.000%)	5 (0.036%)	6 (0.029%)	5 (0.018%)	3
dehydroascorbide(1-)	0 (0.000%)	0 (0.000%)	4 (0.029%)	6 (0.029%)	5 (0.018%)	14
acetylcholine	0 (0.000%)	0 (0.000%)	4 (0.029%)	5 (0.024%)	2 (0.007%)	6
prostaglandin	0 (0.000%)	0 (0.000%)	2 (0.014%)	10 (0.049%)	0 (0.000%)	4
hydrogencarbonate	0 (0.000%)	0 (0.000%)	2 (0.014%)	7 (0.034%)	0 (0.000%)	7
L-ascorbate	0 (0.000%)	0 (0.000%)	2 (0.014%)	3 (0.015%)	19 (0.069%)	4
GDP-fucose	0 (0.000%)	0 (0.000%)	2 (0.014%)	2 (0.010%)	11 (0.040%)	5
3'-phosphonate-5'-adenylyl sulfate(4-)	0 (0.000%)	0 (0.000%)	2 (0.014%)	2 (0.010%)	4 (0.015%)	8
(R)-carnitine	0 (0.000%)	0 (0.000%)	2 (0.014%)	2 (0.010%)	0 (0.000%)	3
sucrose	0 (0.000%)	0 (0.000%)	2 (0.014%)	1 (0.005%)	16 (0.058%)	7
c-GMP-AMP(2-)	0 (0.000%)	0 (0.000%)	0 (0.000%)	6 (0.029%)	0 (0.000%)	3
CMP-N-acetyl-beta-neuraminate(2-)	0 (0.000%)	0 (0.000%)	0 (0.000%)	2 (0.010%)	0 (0.000%)	3
N-retinylidene phos-phatidylethanolamine	0 (0.000%)	0 (0.000%)	0 (0.000%)	1 (0.005%)	0 (0.000%)	3
UDP-alpha-D-xylose(2-)	0 (0.000%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	6 (0.022%)	3

The analysis for transportome of five model organisms, *Escherichia coli K-12* (*Es*), *Saccharomyces cerevisiae* (*Sc*), *Drosophila melanogaster* (*Dm*), *Homo sapiens* (*Hs*), and *Arabidopsis thaliana* (*At*), using the TooT-Triplet (offline) with manhattan distance metric. The numbers in parentheses indicate the proportion of the transportome in relation to the proteome in the second row, and the proportion of the substrate's transporter within the transportome in the subsequent rows.

E.5 TooT-Triplet distances

Table 63: Inter and intra-class distances for training set after triplet training

#	Class	Max. Intra-class Distance	Min. Inter-class Distance	Closest Class	Number of Samples
0	calcium(2+)	7.087268	0.739988	potassium(1+)	576
1	proton	6.970360	0.692443	dehydroascorbide(1-)	817
2	sulfate	6.724621	1.327341	biotinate	44
3	chloride	5.410194	0.782051	hydrogencarbonate	295
4	sodium(1+)	5.239126	0.602044	L-glutamate(1-)	178
5	UDP-D-glucose(2-)	5.110350	1.203223	UDP-N-acetyl-alpha-D-glucosamine(2-)	6
6	riboflavin(1-)	5.018446	1.485888	folate(2-)	24
7	potassium(1+)	4.986507	0.726045	glutathione derivative	484
8	zinc(2+)	4.453715	0.854667	calcium(2+)	53
9	nitrate	4.224268	0.723873	proton	18
10	iron(2+)	4.101011	1.153663	potassium(1+)	16
11	L-cystine zwitterion	3.766217	1.255391	sodium(1+)	8
12	glutathionate(1-)	3.515260	1.979016	maltose	6
13	copper(2+)	3.476302	1.223574	copper(1+)	5
14	pyruvate	3.303638	1.457731	sodium(1+)	14
15	water	3.268059	1.422141	formate	29
16	ammonium	3.208606	1.463116	potassium(1+)	35
17	heme	3.206852	0.783541	glutathione derivative	6
18	cobalamin	3.195211	1.359603	fatty acyl-CoA	6
19	protein	3.079437	0.792145	proton	18
20	1-4-butanediammonium	3.021985	0.954919	calcium(2+)	6
21	iron(3+)	3.006204	1.092148	bacteriocin	5
22	retinol	2.971929	1.189521	proton	18
23	L-ornithinium(1+)	2.924465	1.411958	L-argininium(1+)	8
24	sn-glycerol 3-phosphate(2-)	2.908992	1.347679	maltose	2
25	urea	2.851290	1.764421	calcium(2+)	10
26	lipid-linked peptidoglycan	2.706540	1.619802	sodium(1+)	8
27	L-histidine zwitterion	2.680913	1.182001	L-tryptophan zwitterion	3
28	mercury cation	2.593642	2.116028	potassium(1+)	4
29	biotinate	2.578195	1.327341	sulfate	5
30	2-oxoglutarate(2-)	2.561870	2.313682	L-ornithinium(1+)	5
31	copper(1+)	2.557269	1.019259	cadmium cation	11
32	glutathione derivative	2.535668	0.726045	potassium(1+)	18
33	D-methionine zwitterion	2.515095	2.706762	1-4-butanediammonium	2
34	phosphate ion	2.458009	1.774786	sn-glycerol 3-phosphate(2-)	4
35	maltose	2.444420	1.347679	sn-glycerol 3-phosphate(2-)	5
36	cadmium cation	2.443852	1.019259	copper(1+)	3
37	choline	2.432940	2.340667	proton	9
38	L-glutamate(1-)	2.361895	0.602044	sodium(1+)	7
39	S-adenosyl-L-methionine zwitterion	2.353216	2.590028	L-ornithinium(1+)	6
40	acetyl-CoA(4-)	2.323310	2.513087	sodium(1+)	5
41	oleate	2.314025	1.690717	cholesterol	6
42	bacteriocin	2.310396	1.092148	iron(3+)	3
43	molybdate	2.297308	1.736326	riboflavin(1-)	2
44	gamma-aminobutyric acid zwitterion	2.281194	1.947815	L-methionine zwitterion	3
45	L-arabinose	2.230834	1.688630	iron(3+)	2
46	gluconate	2.227398	3.755613	L-proline zwitterion	2
47	folate(2-)	2.216003	1.212603	NAD	2
48	citrate(3-)	2.204025	1.680504	sodium(1+)	7
49	L-methionine zwitterion	2.195706	1.800568	formate	3
50	ATP(4-)	2.190429	1.219103	NAD	12
51	cholesterol	2.184289	0.838569	N-retinylidene phos- phatidylethanolamine	8
52	fluoride	2.177861	1.980611	biotinate	4
53	hydrogencarbonate	2.160288	0.782051	chloride	4
54	urate salt	2.119289	1.634609	dehydroascorbide(1-)	5
55	acetylcholine	2.119265	1.746673	calcium(2+)	4
56	L-argininium(1+)	2.068744	0.866785	proton	4

Continued on next page

Table 63: Inter and intra-class distances for training set after triplet training

#	Class	Max. Intra-class Distance	Min. Inter-class Distance	Closest Class	Number of Samples
57	L-proline zwitterion	2.058571	0.717097	sodium(1+)	4
58	L-ascorbate	2.024324	1.277490	sodium(1+)	3
59	dehydroascorbide(1-)	1.925059	0.692443	proton	9
60	N-ribosylnicotinamide	1.838826	1.792039	proton	4
61	xanthine	1.837820	3.199968	gamma-aminobutyric acid zwitterion	2
62	glycine betaine	1.724304	1.852068	proton	2
63	formate	1.649253	1.422141	water	4
64	tungstate	1.552319	1.525497	riboflavin(1-)	2
65	NAD	1.549938	1.212603	folate(2-)	4
66	UDP-N-acetyl-alpha-D-glucosamine(2-)	1.481495	1.092177	CMP-N-acetyl-beta-neuraminate(2-)	5
67	fatty acyl-CoA	1.462792	1.359603	cobalamin	3
68	long-chain fatty acid	1.299051	1.655820	maltodextrin	3
69	teichoic acid	1.291712	1.863575	sodium(1+)	2
70	phosphoglycerate	1.257884	1.813916	UDP-N-acetyl-alpha-D-glucosamine(2-)	4
71	sucrose	1.250674	3.273122	UDP-D-glucose(2-)	4
72	3'-phosphonato-5'-adenylyl sulfate(4-)	1.145893	1.410140	UDP-D-glucose(2-)	5
73	boric acid	1.104457	1.783482	UDP-N-acetyl-alpha-D-glucosamine(2-)	2
74	L-tryptophan zwitterion	1.040369	1.182001	L-histidine zwitterion	2
75	GDP-fucose	1.004864	1.480767	water	3
76	phosphonatoenol pyruvate	0.952900	1.726359	UDP-N-acetyl-alpha-D-glucosamine(2-)	2
77	N-(4-aminobenzoyl)-L-glutamate	0.919788	1.902746	L-methionine zwitterion	2
78	glycine zwitterion	0.892344	1.632827	3'-phosphonato-5'-adenylyl sulfate(4-)	3
79	pyrimidine nucleotide	0.850052	1.290201	ATP(4-)	3
80	arsenate ion	0.836491	1.731424	proton	2
81	bile acid	0.798696	1.380384	chloride	6
82	N-retinylidene phos-phatidylethanolamine (R)-carnitine	0.743437	0.838569	cholesterol	2
83	beta-D-glucan	0.737835	2.994943	uridine	2
84	iron-sulfur cluster	0.689745	1.963037	proton	2
85	flavin adenine dinucleotide	0.621080	1.682328	glutathione derivative	4
86	maltodextrin	0.575825	2.481646	potassium(1+)	2
87	c-GMP-AMP(2-)	0.506182	1.655820	long-chain fatty acid	3
88	prostaglandin	0.499086	1.979439	sulfate	2
89	guanine	0.348476	3.511202	flavin adenine dinucleotide	3
90		0.335260	2.479913	gamma-aminobutyric acid zwitterion	2
91	magnesium cation	0.297410	2.549086	calcium(2+)	2
92	lipopolysaccharide	0.257660	2.323100	sodium(1+)	2
93	CMP-N-acetyl-beta-neuraminate(2-)	0.253909	1.092177	UDP-N-acetyl-alpha-D-glucosamine(2-)	2
94	uridine	0.248626	2.994943	(R)-carnitine	2
95	UDP-alpha-D-xylose(2-)	0.190204	2.929723	GDP-fucose	2

The table presents the maximum intra-class and minimum inter-class distances for each class in the dataset after training with TooT-Triplet. The last column indicates the label of the closest sample from a different class. Classes are sorted in descending order based on their maximum intra-class distances.

Table 64: Average distance of the nearest neighbor in TooT-SPEC training set after TooT-Triplet training on Proton Excluded dataset

Class	Avg. distance of nearest neighbor with same label	Std. distance of nearest neighbor with same label	Avg. distance of nearest neighbor with different labels	Std. distance of nearest neighbor with different labels	size
calcium(2+)	0.648238	0.602394	1.690853	0.341107	577
potassium(1+)	0.721905	0.617354	1.693184	0.302800	484

Continued on next page

Table 64: Average distance of the nearest neighbor in TooT-SPEC training set after TooT-Triplet training on Proton Excluded dataset

Class	Avg. distance of nearest neighbor with same label	Std. distance of nearest neighbor with same label	Avg. distance of nearest neighbor with different labels	Std. distance of nearest neighbor with different labels	size
chloride	0.806464	0.638811	1.733471	0.282879	295
sodium(1+)	0.717143	0.616870	1.565560	0.246339	177
zinc(2+)	1.054478	0.612919	1.803798	0.293641	54
sulfate	1.216781	0.698119	1.812318	0.323806	45
ammonium	0.794238	0.615068	1.488303	0.223543	34
water	0.612739	0.461758	1.397590	0.197612	28
riboflavin(1-)	1.126741	0.357245	1.692518	0.219335	25
protein	0.890787	0.391379	1.701419	0.100223	19
retinol	0.792095	0.713903	1.835203	0.295096	18
glutathione derivative	0.556792	0.544772	1.075629	0.264949	18
nitrate	1.277808	0.577336	1.639610	0.201950	17
iron(2+)	1.126354	0.906319	1.691785	0.316017	16
pyruvate	0.789139	0.825418	1.760801	0.161847	13
copper(1+)	0.994716	0.481453	1.319893	0.167251	12
ATP(4-)	0.416413	0.296423	1.372387	0.170698	12
urea	0.632758	0.550360	1.984410	0.369591	11
dehydroascorbide(1-)	0.432674	0.273086	1.163248	0.243898	10
choline	0.439092	0.235714	2.303445	0.134730	9
cholesterol	0.899115	1.004765	1.229116	0.558015	8
L-ornithinium(1+)	0.998137	0.573227	1.283176	0.470497	8
bile acid	0.190215	0.190766	1.125153	0.025777	7
L-glutamate(1-)	1.128083	0.646663	0.969707	0.377632	7
L-cystine zwitterion	1.276799	1.264684	1.372461	0.342621	7
lipid-linked peptidoglycan	1.598139	0.675505	1.793218	0.086560	7
cobalamin	1.446684	0.872403	1.568579	0.441226	6
UDP-D-glucose(2-)	1.300658	1.803829	1.193245	0.243256	6
citrate(3-)	0.639489	0.661117	1.843370	0.114747	6
S-adenosyl-L-methionine zwitterion	1.114852	0.491404	1.120259	0.242433	6
3'-phosphonato-5'-adenylyl sulfate(4-)	0.434631	0.177974	1.170170	0.096369	6
UDP-N-acetyl-alpha-D-glucosamine(2-)	1.011253	0.359613	1.189723	0.101291	6
heme	0.732731	0.425125	1.285589	0.222599	6
copper(2+)	1.530870	0.234042	1.316049	0.179566	6
urate salt	0.743123	0.500514	1.372410	0.348641	6
oleate	0.612325	0.694700	1.802474	0.209206	6
phosphoglycerate	0.366495	0.227976	1.040813	0.094349	5
glutathionate(1-)	0.666745	1.059025	2.003184	0.474866	5
biotinate	0.788030	0.378246	1.477964	0.124585	5
hydrogencarbonate	1.114859	0.800102	1.246645	0.566908	5
1-4-butanediammonium	1.475092	1.014412	1.417263	0.376249	5
sucrose	0.493597	0.135227	1.640059	0.100030	5
acetylcholine	1.177119	0.727625	1.716498	0.112488	4
maltose	1.415317	0.237511	2.015225	0.099432	4
cadmium cation	1.038253	0.460884	1.150836	0.149538	4
L-argininium(1+)	1.260530	0.359778	1.419924	0.126518	4
L-proline zwitterion	1.533755	1.003032	1.090665	0.386439	4
NAD	0.549261	0.225510	1.165083	0.126846	4
formate	0.851060	0.302776	1.537579	0.095907	4
iron-sulfur cluster	0.509213	0.122377	1.434188	0.076683	4
iron(3+)	1.918874	0.403130	1.445292	0.239219	4
GDP-fucose	0.410795	0.338596	1.044937	0.047119	4
acetyl-CoA(4-)	1.491801	0.499587	2.168197	0.092671	4
fluoride	0.922354	0.328617	1.745132	0.052780	4
gamma-aminobutyric acid zwitterion	1.387952	0.948072	1.795089	0.269659	4
2-oxoglutarate(2-)	1.797965	0.398155	1.914179	0.309661	4
pyrimidine nucleotide	0.458052	0.304623	0.970559	0.007741	4
L-arabinose	2.122877	0.099646	1.727016	0.249596	3
N-ribosylnicotinamide	2.008186	0.217408	1.873600	0.269996	3
D-methionine zwitterion	1.150518	0.393287	1.884484	0.189942	3

Continued on next page

Table 64: Average distance of the nearest neighbor in TooT-SPEC training set after TooT-Triplet training on Proton Excluded dataset

Class	Avg. distance of nearest neighbor with same label	Std. distance of nearest neighbor with same label	Avg. distance of nearest neighbor with different labels	Std. distance of nearest neighbor with different labels	size
maltodextrin	1.069568	0.949415	1.802209	0.026760	3
mercury cation	1.410187	0.884778	2.401351	0.160334	3
boric acid	1.359252	1.137427	1.202889	0.208605	3
phosphate ion	1.479602	0.094748	1.869563	0.055470	3
L-ascorbate	1.914632	0.281661	1.746250	0.273565	3
fatty acyl-CoA	0.600569	0.526683	1.288111	0.039497	3
phosphonatoenolpyruvate	0.774486	0.000000	1.001280	0.071941	2
glycine betaine	3.041117	0.000000	1.726025	0.144184	2
arsenate ion	1.918183	0.000000	1.613810	0.081373	2
guanine	4.058576	0.000000	1.701224	0.837102	2
xanthine	3.655659	0.000000	2.520874	0.732891	2
tungstate	1.775443	0.000000	1.447581	0.041673	2
bacteriocin	2.348023	0.000000	1.255130	0.311999	2
glycine zwitterion	0.502808	0.000000	1.397517	0.015499	2
L-tryptophan zwitterion	2.588317	0.000000	1.218659	0.588051	2
beta-D-glucan	0.699514	0.000000	2.082832	0.044913	2
c-GMP-AMP(2-)	0.153720	0.000000	1.812088	0.013794	2
long-chain fatty acid	1.151358	0.000000	1.738373	0.021284	2
sn-glycerol 3-phosphate(2-)	5.526310	0.000000	1.527460	0.147363	2
uridine	4.361460	0.000000	2.353010	0.710958	2
prostaglandin	0.129707	0.000000	3.626264	0.004145	2
folate(2-)	2.742722	0.000000	1.678946	0.281700	2
L-histidine zwitterion	3.035507	0.000000	0.831849	0.265631	2
CMP-N-acetyl-beta-neuraminate(2-)	0.163766	0.000000	0.940675	0.011336	2
(R)-carnitine	0.137736	0.000000	3.064160	0.000192	2
L-methionine zwitterion	2.243223	0.000000	1.724510	0.115616	2
teichoic acid	0.839965	0.000000	1.706618	0.004168	2
flavin adenine dinucleotide	0.759828	0.000000	1.448004	0.012645	2
N-(4-aminobenzoyl)-L-glutamate	1.433130	0.000000	2.172776	0.043980	2
molybdate	2.325576	0.000000	1.690453	0.284546	2
N-retinylidene phos-phatidylethanolamine	0.318327	0.000000	0.806708	0.039205	2
UDP-alpha-D-xylose(2-)	0.692081	0.000000	2.277358	0.072912	2
gluconate	3.383928	0.000000	2.493544	0.796185	2
lipopolysaccharide	1.012989	0.000000	2.062570	0.015847	2
magnesium cation	0.726717	0.000000	1.559698	0.023450	2

The table shows the maximum intra-class and minimum inter-class distances for each class in the dataset, excluding the majority class **proton**, after training with TooT-Triplet. The last column displays the label of the closest sample from a different class. Classes are sorted in descending order based on their maximum intra-class distances.

Appendix F

TooTranslator

F.1 Results

Table 65: TooTranslator results with weighted MSE loss function for the hold-out test set of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	817	409	394	11	15	1065	0.982	0.973	0.963	0.968	0.956
calcium(2+)	577	288	282	6	6	1191	0.992	0.979	0.979	0.979	0.974
potassium(1+)	484	242	233	4	9	1239	0.991	0.983	0.963	0.973	0.968
chloride	295	147	138	2	9	1336	0.993	0.986	0.939	0.962	0.958
sodium(1+)	177	89	70	5	19	1391	0.98	0.93	0.79	0.85	0.85
zinc(2+)	54	26	25	0	1	1459	1.00	1.00	0.96	0.98	0.98
sulfate	45	22	20	2	2	1461	1.00	0.91	0.91	0.91	0.91
ammonium	34	18	18	0	0	1467	1.00	1.00	1.00	1.00	1.00
water	28	15	15	0	0	1470	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	25	12	10	0	2	1473	1.00	1.00	0.83	0.91	0.91
protein	19	9	9	5	0	1471	1.0	0.6	1.0	0.8	0.8
glutathione derivative	18	9	9	0	0	1476	1.0	1.0	1.0	1.0	1.0
retinol	18	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
nitrate	17	9	9	0	0	1476	1.0	1.0	1.0	1.0	1.0
iron(2+)	16	8	8	1	0	1476	1.0	0.9	1.0	0.9	0.9
pyruvate	13	7	6	0	1	1478	1.0	1.0	0.9	0.9	0.9
ATP(4-)	12	6	5	0	1	1479	1.0	1.0	0.8	0.9	0.9
copper(1+)	12	5	5	1	0	1479	1.0	0.8	1.0	0.9	0.9
urea	11	5	5	1	0	1479	1.0	0.8	1.0	0.9	0.9
dehydroascorbide(1-)	10	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
choline	9	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
L-ornithinium(1+)	8	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
cholesterol	8	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
bile acid	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
L-glutamate(1-)	7	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
L-cystine zwitterion	7	4	1	1	3	1480	1.0	0.5	0.2	0.3	0.4
lipid-linked peptidoglycan	7	4	4	2	0	1479	1.0	0.7	1.0	0.8	0.8
3'-phosphonato-5'-adenylyl sulfate(4-)	6	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
copper(2+)	6	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
UDP-N-acetyl-alpha-D-glucosamine(2-)	6	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
S-adenosyl-L-methionine zwitterion	6	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
citrate(3-)	6	4	2	1	2	1480	1.0	0.7	0.5	0.6	0.6
heme	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
oleate	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6

Continued on next page

Table 65: TooTranslator results with weighted MSE loss function for the hold-out test set of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
cobalamin	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
UDP-D-glucose(2-)	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
urate salt	6	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
hydrogencarbonate	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
sucrose	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphoglycerate	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
biotinate	5	3	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
1-4-butanediammonium	5	3	0	2	3	1480	1.0	0.0	0.0	0.0	0.0
glutathionate(1-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
2-oxoglutarate(2-)	4	3	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
maltose	4	3	0	1	3	1481	1.0	0.0	0.0	0.0	0.0
iron(3+)	4	3	0	1	3	1481	1.0	0.0	0.0	0.0	0.0
acetyl-CoA(4-)	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
pyrimidine nucleotide	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
cadmium cation	4	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
NAD	4	2	2	4	0	1479	1.0	0.3	1.0	0.5	0.6
L-argininium(1+)	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-proline zwitterion	4	2	2	9	0	1474	1.0	0.2	1.0	0.3	0.4
gamma-aminobutyric acid zwitterion	4	1	1	3	0	1481	1.0	0.2	1.0	0.4	0.5
GDP-fucose	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-ascorbate	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-arabinose	3	1	1	3	0	1481	1.0	0.2	1.0	0.4	0.5
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
mercury cation	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
N-ribosylnicotinamide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphate ion	3	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
D-methionine zwitterion	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
boric acid	3	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
maltodextrin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
N-(4-aminobenzoyl)-L-glutamate	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
xanthine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
uridine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
long-chain fatty acid	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-histidine zwitterion	2	2	0	1	2	1482	1.0	0.0	0.0	0.0	0.0
glycine zwitterion	2	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
L-methionine zwitterion	2	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
bacteriocin	2	2	2	5	0	1478	1.0	0.3	1.0	0.4	0.5
prostaglandin	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
molybdate	2	1	0	3	1	1481	1.0	0.0	0.0	0.0	0.0
N-retinylidene\textasciitilde phosphatidylethan...	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
lipopolysaccharide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gluconate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
tungstate	2	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
sn-glycerol 3-phosphate(2-)	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0

Continued on next page

Table 65: TooTranslator results with weighted MSE loss function for the hold-out test set of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
CMP-N-acetyl-beta-neuraminate(2-)	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
Overall	2970	1485	1382	-	-	-	0.999	0.931	0.814	0.748	0.918

The table presents the results of the hold-out test set of UniProt-SPEC-100 using TooTranslator trained using the weighted MSE loss function, predicted by a K-NN classifier (K=1). The overall number for TP presents the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric.

Table 66: TooTranslator results with MAE loss function for the hold-out test set of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	817	409	396	16	13	1060	0.980	0.961	0.968	0.965	0.951
calcium(2+)	577	288	284	4	4	1193	0.995	0.986	0.986	0.986	0.983
potassium(1+)	484	242	235	4	7	1239	0.993	0.983	0.971	0.977	0.973
chloride	295	147	141	10	6	1328	0.989	0.934	0.959	0.946	0.940
sodium(1+)	177	89	79	6	10	1390	0.99	0.93	0.89	0.91	0.90
zinc(2+)	54	26	25	1	1	1458	1.00	0.96	0.96	0.96	0.96
sulfate	45	22	20	3	2	1460	1.00	0.87	0.91	0.89	0.89
ammonium	34	18	18	0	0	1467	1.00	1.00	1.00	1.00	1.00
water	28	15	15	0	0	1470	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	25	12	9	0	3	1473	1.00	1.00	0.75	0.86	0.86
protein	19	9	8	2	1	1474	1.0	0.8	0.9	0.8	0.8
glutathione derivative	18	9	9	0	0	1476	1.0	1.0	1.0	1.0	1.0
retinol	18	9	8	1	1	1475	1.0	0.9	0.9	0.9	0.9
nitrate	17	9	9	0	0	1476	1.0	1.0	1.0	1.0	1.0
iron(2+)	16	8	8	3	0	1474	1.0	0.7	1.0	0.8	0.9
pyruvate	13	7	6	0	1	1478	1.0	1.0	0.9	0.9	0.9
ATP(4-)	12	6	5	0	1	1479	1.0	1.0	0.8	0.9	0.9
copper(1+)	12	5	5	2	0	1478	1.0	0.7	1.0	0.8	0.8
urea	11	5	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
dehydroascorbide(1-)	10	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
choline	9	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
L-ornithinium(1+)	8	4	3	0	1	1481	1.0	1.0	0.8	0.9	0.9
cholesterol	8	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
bile acid	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
L-glutamate(1-)	7	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
L-cystine zwitterion	7	4	1	1	3	1480	1.0	0.5	0.2	0.3	0.4
lipid-linked peptidoglycan	7	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
3'-phosphonato-5'-adenyl sulfate(4-)	6	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
copper(2+)	6	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
UDP-N-acetyl-alpha-D-glucosamine(2-)	6	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
S-adenosyl-L-methionine zwitterion	6	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
citrate(3-)	6	4	2	1	2	1480	1.0	0.7	0.5	0.6	0.6
heme	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
oleate	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
cobalamin	6	3	1	2	2	1480	1.0	0.3	0.3	0.3	0.3
UDP-D-glucose(2-)	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
urate salt	6	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
hydrogencarbonate	5	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
sucrose	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphoglycerate	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
biotinate	5	3	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
1-4-butanediammonium	5	3	0	1	3	1481	1.0	0.0	0.0	0.0	0.0
glutathionate(1-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6

Continued on next page

Table 66: TooTranslator results with MAE loss function for the hold-out test set of UniProt-SPEC-100

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
2-oxoglutarate(2-)	4	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
maltose	4	3	0	1	3	1481	1.0	0.0	0.0	0.0	0.0
iron(3+)	4	3	2	1	1	1481	1.0	0.7	0.7	0.7	0.7
acetyl-CoA(4-)	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
pyrimidine nucleotide	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
cadmium cation	4	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
NAD	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
L-argininium(1+)	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-proline zwitterion	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
gamma-aminobutyric acid zwitterion	4	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
GDP-fucose	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-ascorbate	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-arabinose	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
mercury cation	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
N-ribosylnicotinamide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphate ion	3	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
D-methionine zwitterion	3	1	1	3	0	1481	1.0	0.2	1.0	0.4	0.5
boric acid	3	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
maltodextrin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
xanthine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
uridine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
long-chain fatty acid	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-histidine zwitterion	2	2	0	2	2	1481	1.0	0.0	0.0	0.0	0.0
glycine zwitterion	2	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
L-methionine zwitterion	2	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
bacteriocin	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
prostaglandin	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
molybdate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
N-retinylidene\textasciitilde phosphatidylethan...	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gluconate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
tungstate	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
sn-glycerol 3-phosphate(2-)	2	1	0	2	1	1482	1.0	0.0	0.0	0.0	0.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuraminate(2-)	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
Overall	2970	1485	1395	-	-	-	0.999	0.939	0.797	0.762	0.928

The table presents the results of the hold-out test set of UniProt-SPEC-100 using TooTranslator trained using the MAE loss function, predicted by a K-NN classifier (K=1). The overall number for TP presents the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric.

Table 67: TooTranslator results with MSE loss function for the hold-out test set of UniProt-SPEC-100 (description and SMILES)

Substrate	Train	Test	TP	FP	FN	TN	Acc.	Prec.	Rec.	F1	MCC
proton	817	409	387	2	22	1074	0.984	0.995	0.946	0.970	0.959
calcium(2+)	577	288	281	3	7	1194	0.993	0.989	0.976	0.983	0.978
potassium(1+)	484	242	235	23	7	1220	0.980	0.911	0.971	0.940	0.929
chloride	295	147	142	5	5	1333	0.993	0.966	0.966	0.966	0.962
sodium(1+)	177	89	65	3	24	1393	0.98	0.96	0.73	0.83	0.83
zinc(2+)	54	26	25	0	1	1459	1.00	1.00	0.96	0.98	0.98
sulfate	45	22	20	1	2	1462	1.00	0.95	0.91	0.93	0.93
ammonium	34	18	18	0	0	1467	1.00	1.00	1.00	1.00	1.00
water	28	15	15	0	0	1470	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	25	12	10	0	2	1473	1.00	1.00	0.83	0.91	0.91
protein	19	9	8	14	1	1462	1.0	0.4	0.9	0.5	0.6
retinol	18	9	9	1	0	1475	1.0	0.9	1.0	0.9	0.9
glutathione derivative	18	9	9	1	0	1475	1.0	0.9	1.0	0.9	0.9
nitrate	17	9	9	3	0	1473	1.0	0.8	1.0	0.9	0.9
iron(2+)	16	8	6	1	2	1476	1.0	0.9	0.8	0.8	0.8
pyruvate	13	7	6	0	1	1478	1.0	1.0	0.9	0.9	0.9
ATP(4-)	12	6	5	0	1	1479	1.0	1.0	0.8	0.9	0.9
copper(1+)	12	5	5	2	0	1478	1.0	0.7	1.0	0.8	0.8
urea	11	5	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
dehydroascorbide(1-)	10	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
choline	9	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
cholesterol	8	4	3	2	1	1479	1.0	0.6	0.8	0.7	0.7
L-ornithinium(1+)	8	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
L-glutamate(1-)	7	3	3	3	0	1479	1.0	0.5	1.0	0.7	0.7
bile acid	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
lipid-linked peptidoglycan	7	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
L-cystine zwitterion	7	4	1	1	3	1480	1.0	0.5	0.2	0.3	0.4
copper(2+)	6	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
citrate(3-)	6	4	2	1	2	1480	1.0	0.7	0.5	0.6	0.6
cobalamin	6	3	1	2	2	1480	1.0	0.3	0.3	0.3	0.3
oleate	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
heme	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
S-adenosyl-L-methionine zwitterion	6	3	3	1	0	1481	1.0	0.8	1.0	0.9	0.9
UDP-D-glucose(2-)	6	3	1	3	2	1479	1.0	0.2	0.3	0.3	0.3
UDP-N-acetyl-alpha-D-glucosamine(2-)	6	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
3'-phosphonato-5'-adenylyl sulfate(4-)	6	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
urate salt	6	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
sucrose	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
hydrogencarbonate	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
glutathionate(1-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
biotinate	5	3	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
1-4-butanediammonium	5	3	0	2	3	1480	1.0	0.0	0.0	0.0	0.0
phosphoglycerate	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
pyrimidine nucleotide	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
formate	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
acetylcholine	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
cadmium cation	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
GDP-fucose	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid zwitterion	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-proline zwitterion	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
NAD	4	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
L-argininium(1+)	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0

Continued on next page

Table 67: TooTranslator results with MSE loss function for the hold-out test set of UniProt-SPEC-100 (description and SMILES)

Substrate	Train	Test	TP	FP	FN	TN	Acc.	Prec.	Rec.	F1	MCC
iron(3+)	4	3	0	1	3	1481	1.0	0.0	0.0	0.0	0.0
maltose	4	3	0	3	3	1479	1.0	0.0	0.0	0.0	0.0
2-oxoglutarate(2-)	4	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
acetyl-CoA(4-)	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
D-methionine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
boric acid	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-ascorbate	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-arabinose	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
maltodextrin	3	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
N-ribosylnicotinamide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphate ion	3	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
mercury cation	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuraminate(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
N-(4-aminobenzoyl)-L-glutamate	2	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
xanthine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
sn-glycerol 3-phosphate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
tungstate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
gluconate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	3	0	1481	1.0	0.2	1.0	0.4	0.5
N-retinylidene\textasciitilde phosphatidylethan...	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
molybdate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
prostaglandin	2	2	2	13	0	1470	1.0	0.1	1.0	0.2	0.4
L-methionine zwitterion	2	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
L-histidine zwitterion	2	2	0	1	2	1482	1.0	0.0	0.0	0.0	0.0
glycine zwitterion	2	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
bacteriocin	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
long-chain fatty acid	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
uridine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
2970	1485	1369	-	-	-	0.998	0.922	0.788	0.740	0.908	Overall

The table presents the results of the hold-out test set of UniProt-SPEC-100 using TooTranslator trained using the MSE loss function, predicted by a K-NN classifier (K=1), trained and tested on the contamination of description and SMILES vectors. The overall number for TP presents the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric.

Table 68: TooTranslator results with MSE loss function for the hold-out test set of UniProt-SPEC-100 (SMILES)

Substrate	Train	Test	TP	FP	FN	TN	Acc.	Prec.	Rec.	F1	MCC
proton	817	409	399	18	10	1058	0.981	0.957	0.976	0.966	0.953
calcium(2+)	577	288	283	8	5	1189	0.991	0.973	0.983	0.978	0.972
potassium(1+)	484	242	234	0	8	1243	0.995	1.000	0.967	0.983	0.980

Continued on next page

Table 68: TooTranslator results with MSE loss function for the hold-out test set of UniProt-SPEC-100 (SMILES)

Substrate	Train	Test	TP	FP	FN	TN	Acc.	Prec.	Rec.	F1	MCC
chloride	295	147	138	5	9	1333	0.991	0.965	0.939	0.952	0.947
sodium(1+)	177	89	1	0	88	1396	0.94	1.00	0.01	0.02	0.10
zinc(2+)	54	26	26	0	0	1459	1.00	1.00	1.00	1.00	1.00
sulfate	45	22	19	2	3	1461	1.00	0.91	0.86	0.88	0.88
ammonium	34	18	18	0	0	1467	1.00	1.00	1.00	1.00	1.00
water	28	15	15	0	0	1470	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	25	12	10	0	2	1473	1.00	1.00	0.83	0.91	0.91
protein	19	9	8	4	1	1472	1.0	0.7	0.9	0.8	0.8
retinol	18	9	8	0	1	1476	1.0	1.0	0.9	0.9	0.9
glutathione derivative	18	9	9	1	0	1475	1.0	0.9	1.0	0.9	0.9
nitrate	17	9	9	1	0	1475	1.0	0.9	1.0	0.9	0.9
iron(2+)	16	8	8	4	0	1473	1.0	0.7	1.0	0.8	0.8
pyruvate	13	7	6	0	1	1478	1.0	1.0	0.9	0.9	0.9
ATP(4-)	12	6	5	0	1	1479	1.0	1.0	0.8	0.9	0.9
copper(1+)	12	5	5	83	0	1397	0.9	0.1	1.0	0.1	0.2
urea	11	5	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
dehydroascorbide(1-)	10	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
choline	9	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
cholesterol	8	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
L-ornithinium(1+)	8	4	4	4	0	1477	1.0	0.5	1.0	0.7	0.7
L-glutamate(1-)	7	3	3	4	0	1478	1.0	0.4	1.0	0.6	0.7
bile acid	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
lipid-linked peptidoglycan	7	4	0	0	4	1481	1.0	NaN	0.0	0.0	NaN
L-cystine zwitterion	7	4	1	0	3	1481	1.0	1.0	0.2	0.4	0.5
copper(2+)	6	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
citrate(3-)	6	4	2	0	2	1481	1.0	1.0	0.5	0.7	0.7
cobalamin	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
oleate	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
heme	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
S-adenosyl-L-methionine zwitterion	6	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
UDP-D-glucose(2-)	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
UDP-N-acetyl-alpha-D-glucosamine(2-)	6	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
3'-phosphonato-5'-adenylyl sulfate(4-)	6	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
urate salt	6	2	2	16	0	1467	1.0	0.1	1.0	0.2	0.3
sucrose	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
hydrogencarbonate	5	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
glutathionate(1-)	5	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
biotinate	5	3	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
1-4-butanediammonium	5	3	0	1	3	1481	1.0	0.0	0.0	0.0	0.0
phosphoglycerate	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
pyrimidine nucleotide	4	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
formate	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
acetylcholine	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
cadmium cation	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
GDP-fucose	4	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
gamma-aminobutyric acid zwitterion	4	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
fluoride	4	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
L-proline zwitterion	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
NAD	4	2	2	2	0	1481	1.0	0.5	1.0	0.7	0.7
L-argininium(1+)	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
iron(3+)	4	3	1	1	2	1481	1.0	0.5	0.3	0.4	0.4
maltose	4	3	0	0	3	1482	1.0	NaN	0.0	0.0	NaN
2-oxoglutarate(2-)	4	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
acetyl-CoA(4-)	4	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
D-methionine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
boric acid	3	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
L-ascorbate	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-arabinose	3	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
maltodextrin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0

Continued on next page

Table 68: TooTranslator results with MSE loss function for the hold-out test set of UniProt-SPEC-100 (SMILES)

Substrate	Train	Test	TP	FP	FN	TN	Acc.	Prec.	Rec.	F1	MCC
N-ribosylnicotinamide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphate ion	3	2	2	7	0	1476	1.0	0.2	1.0	0.4	0.5
mercury cation	3	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
CMP-N-acetyl-beta-neuraminate(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
teichoic acid	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
xanthine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
glycine betaine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
magnesium cation	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
folate(2-)	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
sn-glycerol 3-phosphate(2-)	2	1	0	3	1	1481	1.0	0.0	0.0	0.0	0.0
tungstate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gluconate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
beta-D-glucan	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
lipopolysaccharide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
N-retinylidene\textasciitilde phosphatidylethan...	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
molybdate	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
prostaglandin	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-methionine zwitterion	2	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
L-histidine zwitterion	2	2	0	1	2	1482	1.0	0.0	0.0	0.0	0.0
glycine zwitterion	2	2	1	4	1	1479	1.0	0.2	0.5	0.3	0.3
bacteriocin	2	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
long-chain fatty acid	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
uridine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
Overall	2970	1485	1305	-	-	-	0.997	0.879	0.715	0.670	0.861

The table presents the results of the hold-out test set of UniProt-SPEC-100 using TooTranslator trained using the MSE loss function, predicted by a K-NN classifier (K=1), trained and tested on the SMILES vectors. The overall number for TP presents the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric.

Table 69: TooTranslator results with MSE loss function for the hold-out test set of UniProt-SPEC-100 (description)

Substrate	Train	Test	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	817	409	398	24	11	1052	0.976	0.943	0.973	0.958	0.942
calcium(2+)	577	288	283	2	5	1195	0.995	0.993	0.983	0.988	0.985
potassium(1+)	484	242	230	19	12	1224	0.979	0.924	0.950	0.937	0.924
chloride	295	147	139	7	8	1331	0.990	0.952	0.946	0.949	0.943
sodium(1+)	177	89	79	1	10	1395	0.99	0.99	0.89	0.94	0.93
zinc(2+)	54	26	0	0	26	1459	0.98	NaN	0.00	0.00	NaN
sulfate	45	22	20	6	2	1457	0.99	0.77	0.91	0.83	0.83
ammonium	34	18	0	0	18	1467	0.99	NaN	0.00	0.00	NaN
water	28	15	15	0	0	1470	1.00	1.00	1.00	1.00	1.00
riboflavin(1-)	25	12	10	0	2	1473	1.00	1.00	0.83	0.91	0.91
protein	19	9	8	1	1	1475	1.0	0.9	0.9	0.9	0.9
retinol	18	9	9	0	0	1476	1.0	1.0	1.0	1.0	1.0

Continued on next page

Substrate	Train	Test	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
glutathione derivative	18	9	9	0	0	1476	1.0	1.0	1.0	1.0	1.0
nitrate	17	9	8	1	1	1475	1.0	0.9	0.9	0.9	0.9
iron(2+)	16	8	0	0	8	1477	1.0	NaN	0.0	0.0	NaN
pyruvate	13	7	5	0	2	1478	1.0	1.0	0.7	0.8	0.8
ATP(4-)	12	6	6	0	0	1479	1.0	1.0	1.0	1.0	1.0
copper(1+)	12	5	4	1	1	1479	1.0	0.8	0.8	0.8	0.8
urea	11	5	5	0	0	1480	1.0	1.0	1.0	1.0	1.0
dehydroascorbide(1-)	10	4	4	1	0	1480	1.0	0.8	1.0	0.9	0.9
choline	9	4	4	3	0	1478	1.0	0.6	1.0	0.7	0.8
cholesterol	8	4	0	0	4	1481	1.0	NaN	0.0	0.0	NaN
L-ornithinium(1+)	8	4	3	0	1	1481	1.0	1.0	0.8	0.9	0.9
L-glutamate(1-)	7	3	0	0	3	1482	1.0	NaN	0.0	0.0	NaN
bile acid	7	3	3	0	0	1482	1.0	1.0	1.0	1.0	1.0
lipid-linked peptidoglycan	7	4	4	0	0	1481	1.0	1.0	1.0	1.0	1.0
L-cystine zwitterion	7	4	1	0	3	1481	1.0	1.0	0.2	0.4	0.5
copper(2+)	6	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
citrate(3-)	6	4	2	5	2	1476	1.0	0.3	0.5	0.4	0.4
cobalamin	6	3	0	12	3	1470	1.0	0.0	0.0	0.0	0.0
oleate	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
heme	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
S-adenosyl-L-methionine	6	3	3	3	0	1479	1.0	0.5	1.0	0.7	0.7
zwitterion											
UDP-D-glucose(2-)	6	3	1	0	2	1482	1.0	1.0	0.3	0.5	0.6
UDP-N-acetyl-alpha-D-glucosamine(2-)	6	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
3'-phosphonato-5'-adenylyl sulfate(4-)	6	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
urate salt	6	2	2	3	0	1480	1.0	0.4	1.0	0.6	0.6
sucrose	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
hydrogencarbonate	5	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
glutathione(1-)	5	3	1	1	2	1481	1.0	0.5	0.3	0.4	0.4
biotinate	5	3	2	2	1	1480	1.0	0.5	0.7	0.6	0.6
1-4-butanediammonium	5	3	0	0	3	1482	1.0	NaN	0.0	0.0	NaN
phosphoglycerate	5	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
pyrimidine nucleotide	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
formate	4	2	1	0	1	1483	1.0	1.0	0.5	0.7	0.7
acetylcholine	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
cadmium cation	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
GDP-fucose	4	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
gamma-aminobutyric acid	4	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
zwitterion											
fluoride	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-proline zwitterion	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
NAD	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
L-argininium(1+)	4	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
iron-sulfur cluster	4	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
iron(3+)	4	3	0	36	3	1446	1.0	0.0	0.0	0.0	0.0
maltose	4	3	0	3	3	1479	1.0	0.0	0.0	0.0	0.0
2-oxoglutarate(2-)	4	3	2	0	1	1482	1.0	1.0	0.7	0.8	0.8
acetyl-CoA(4-)	4	3	3	2	0	1480	1.0	0.6	1.0	0.8	0.8
D-methionine zwitterion	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
boric acid	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-ascorbate	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-arabinose	3	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
maltodextrin	3	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
N-ribosylnicotinamide	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
phosphate ion	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
mercury cation	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
fatty acyl-CoA	3	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
(R)-carnitine	2	1	1	2	0	1482	1.0	0.3	1.0	0.5	0.6
CMP-N-acetyl-beta-neuraminate(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
c-GMP-AMP(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
L-tryptophan zwitterion	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
N-(4-aminobenzoyl)-L-glutamate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
flavin adenine dinucleotide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
teichoic acid	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0

Continued on next page

Substrate	Train	Test	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
phosphonatoenolpyruvate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
arsenate ion	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
UDP-alpha-D-xylose(2-)	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
guanine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
xanthine	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
glycine betaine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
magnesium cation	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
folate(2-)	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
sn-glycerol 3-phosphate(2-)	2	1	0	5	1	1479	1.0	0.0	0.0	0.0	0.0
tungstate	2	1	1	1	0	1483	1.0	0.5	1.0	0.7	0.7
gluconate	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
beta-D-glucan	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
lipopolysaccharide	2	1	1	0	0	1484	1.0	1.0	1.0	1.0	1.0
N-retinylidene phos- phatidylethanolamine	2	1	0	0	1	1484	1.0	NaN	0.0	0.0	NaN
molybdate	2	1	0	1	1	1483	1.0	0.0	0.0	0.0	0.0
prostaglandin	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
L-methionine zwitterion	2	2	1	1	1	1482	1.0	0.5	0.5	0.5	0.5
L-histidine zwitterion	2	2	0	0	2	1483	1.0	NaN	0.0	0.0	NaN
glycine zwitterion	2	2	2	1	0	1482	1.0	0.7	1.0	0.8	0.8
bacteriocin	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
long-chain fatty acid	2	2	2	0	0	1483	1.0	1.0	1.0	1.0	1.0
uridine	2	1	1	5	0	1479	1.0	0.2	1.0	0.3	0.4
Overall	2970	1485	1330	-	-	-	0.998	0.896	0.744	0.711	0.877

The table presents the results of the hold-out test set of UniProt-SPEC-100 using TooTranslator trained using the MSE loss function, predicted by a K-NN classifier (K=1), trained and tested on the description vectors. The overall number for TP presents the number of True Positive predictions across all the classes. The totals for the other three columns do not provide useful information in a multiclass classification report. Five metrics, accuracy, precision, recall, F1-score, and MCC, are reported in detail to evaluate the model performance for each class. The overall entry shows the macro average over all the classes for each metric.

Table 70: 10 Nearest Neighbors predicted by TooTranslator-Huber for each test sample in TS2

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is predic- tion correct
0	L-serine zwitterion	0.35	L-aspartate(1-)	0.48	False
0	oxalate(2-)	0.41	L-aspartate(1-)	0.48	False
0	nitrite	0.42	L-aspartate(1-)	0.48	False
0	succinate(2-)	0.45	L-aspartate(1-)	0.48	False
0	acetate	0.45	L-aspartate(1-)	0.48	False
0	nicotinate	0.45	L-aspartate(1-)	0.48	False
0	creatine zwitterion	0.48	L-aspartate(1-)	0.48	False
0	benzoate	0.48	L-aspartate(1-)	0.48	False
0	L-phenylalanine zwitterion	0.48	L-aspartate(1-)	0.48	False
0	silver cation	0.48	L-aspartate(1-)	0.48	False
1	silver cation	0.46	poly-beta-1-6-N-acetyl-D-glucosamine	0.66	False
1	nitrite	0.52	poly-beta-1-6-N-acetyl-D-glucosamine	0.66	False
1	manganese cation	0.56	poly-beta-1-6-N-acetyl-D-glucosamine	0.66	False
1	sulfite	0.57	poly-beta-1-6-N-acetyl-D-glucosamine	0.66	False

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
1	phytochelatin	0.59	poly-beta-1-6-N-acetyl-D-glucosamine	0.66	False
1	nickel cation	0.61	poly-beta-1-6-N-acetyl-D-glucosamine	0.66	False
1	benzoate	0.62	poly-beta-1-6-N-acetyl-D-glucosamine	0.66	False
1	mevalonate	0.63	poly-beta-1-6-N-acetyl-D-glucosamine	0.66	False
1	nicotinate	0.64	poly-beta-1-6-N-acetyl-D-glucosamine	0.66	False
1	D-ribose	0.64	poly-beta-1-6-N-acetyl-D-glucosamine	0.66	False
2	L-serine zwitterion	0.36	nicotinate	0.56	False
2	L-aspartate(1-)	0.38	nicotinate	0.56	False
2	O-acylcarnitine	0.40	nicotinate	0.56	False
2	succinate(2-)	0.42	nicotinate	0.56	False
2	L-threonine zwitterion	0.43	nicotinate	0.56	False
2	oxalate(2-)	0.43	nicotinate	0.56	False
2	glycerol 1-phosphodiester(1-)	0.44	nicotinate	0.56	False
2	arsenite(1-)	0.45	nicotinate	0.56	False
2	S-methyl-L-methionine zwitterion	0.46	nicotinate	0.56	False
2	L-leucine zwitterion	0.46	nicotinate	0.56	False
3	sulfite	0.40	nitrite	0.41	False
3	oxalate(2-)	0.41	nitrite	0.41	False
3	nitrite	0.41	nitrite	0.41	True
3	benzoate	0.42	nitrite	0.41	False
3	nicotinate	0.43	nitrite	0.41	False
3	succinate(2-)	0.44	nitrite	0.41	False
3	L-serine zwitterion	0.46	nitrite	0.41	False
3	acetate	0.51	nitrite	0.41	False
3	silver cation	0.52	nitrite	0.41	False
3	L-phenylalanine zwitterion	0.56	nitrite	0.41	False
4	L-serine zwitterion	0.28	succinate(2-)	0.30	False
4	L-aspartate(1-)	0.28	succinate(2-)	0.30	False
4	succinate(2-)	0.30	succinate(2-)	0.30	True
4	oxalate(2-)	0.31	succinate(2-)	0.30	False
4	L-phenylalanine zwitterion	0.36	succinate(2-)	0.30	False
4	L-glutamine zwitterion	0.36	succinate(2-)	0.30	False
4	L-leucine zwitterion	0.40	succinate(2-)	0.30	False
4	L-tyrosine zwitterion	0.41	succinate(2-)	0.30	False
4	creatine zwitterion	0.43	succinate(2-)	0.30	False
4	allantoate	0.44	succinate(2-)	0.30	False
5	O-acylcarnitine	0.34	spermine(4+)	0.68	False
5	L-aspartate(1-)	0.43	spermine(4+)	0.68	False
5	L-serine zwitterion	0.45	spermine(4+)	0.68	False
5	oxalate(2-)	0.47	spermine(4+)	0.68	False
5	S-methyl-L-methionine zwitterion	0.48	spermine(4+)	0.68	False
5	L-leucine zwitterion	0.49	spermine(4+)	0.68	False

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
5	succinate(2-)	0.49	spermine(4+)	0.68	False
5	glycerol 1-phosphodiester(1-)	0.49	spermine(4+)	0.68	False
5	L-threonine zwitterion	0.49	spermine(4+)	0.68	False
5	organic phosphonate	0.49	spermine(4+)	0.68	False
6	O-acylcarnitine	0.33	spermine(4+)	0.70	False
6	organic phosphonate	0.47	spermine(4+)	0.70	False
6	L-aspartate(1-)	0.49	spermine(4+)	0.70	False
6	L-serine zwitterion	0.51	spermine(4+)	0.70	False
6	glycerol 1-phosphodiester(1-)	0.52	spermine(4+)	0.70	False
6	L-threonine zwitterion	0.53	spermine(4+)	0.70	False
6	S-methyl-L-methionine zwitterion	0.54	spermine(4+)	0.70	False
6	creatine zwitterion	0.54	spermine(4+)	0.70	False
6	oxalate(2-)	0.55	spermine(4+)	0.70	False
6	allantoate	0.55	spermine(4+)	0.70	False
7	L-tyrosine zwitterion	0.32	L-glutamine zwitterion	0.34	False
7	L-phenylalanine zwitterion	0.34	L-glutamine zwitterion	0.34	False
7	L-glutamine zwitterion	0.34	L-glutamine zwitterion	0.34	True
7	L-serine zwitterion	0.37	L-glutamine zwitterion	0.34	False
7	S-methyl-L-methionine zwitterion	0.41	L-glutamine zwitterion	0.34	False
7	L-threonine zwitterion	0.41	L-glutamine zwitterion	0.34	False
7	L-aspartate(1-)	0.44	L-glutamine zwitterion	0.34	False
7	L-leucine zwitterion	0.46	L-glutamine zwitterion	0.34	False
7	spermine(4+)	0.55	L-glutamine zwitterion	0.34	False
7	O-acylcarnitine	0.57	L-glutamine zwitterion	0.34	False
8	L-serine zwitterion	0.26	L-threonine zwitterion	0.43	False
8	L-aspartate(1-)	0.30	L-threonine zwitterion	0.43	False
8	creatine zwitterion	0.38	L-threonine zwitterion	0.43	False
8	L-glutamine zwitterion	0.38	L-threonine zwitterion	0.43	False
8	L-phenylalanine zwitterion	0.39	L-threonine zwitterion	0.43	False
8	S-methyl-L-methionine zwitterion	0.40	L-threonine zwitterion	0.43	False
8	allantoate	0.41	L-threonine zwitterion	0.43	False
8	L-leucine zwitterion	0.42	L-threonine zwitterion	0.43	False
8	oxalate(2-)	0.42	L-threonine zwitterion	0.43	False
8	succinate(2-)	0.43	L-threonine zwitterion	0.43	False
9	O-acylcarnitine	0.37	fluconazole	1.09	False
9	L-serine zwitterion	0.39	fluconazole	1.09	False
9	oxalate(2-)	0.40	fluconazole	1.09	False
9	L-aspartate(1-)	0.40	fluconazole	1.09	False
9	succinate(2-)	0.41	fluconazole	1.09	False
9	glycerol 1-phosphodiester(1-)	0.42	fluconazole	1.09	False
9	arsenite(1-)	0.45	fluconazole	1.09	False
9	L-threonine zwitterion	0.45	fluconazole	1.09	False
9	organic phosphonate	0.47	fluconazole	1.09	False
9	glycerol	0.47	fluconazole	1.09	False
10	L-glutamine zwitterion	0.28	agmatinium(2+)	0.70	False
10	L-phenylalanine zwitterion	0.30	agmatinium(2+)	0.70	False
10	L-tyrosine zwitterion	0.30	agmatinium(2+)	0.70	False
10	L-serine zwitterion	0.31	agmatinium(2+)	0.70	False

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
10	L-threonine zwitterion	0.32	agmatinium(2+)	0.70	False
10	S-methyl-L-methionine zwitterion	0.32	agmatinium(2+)	0.70	False
10	L-leucine zwitterion	0.36	agmatinium(2+)	0.70	False
10	L-aspartate(1-)	0.37	agmatinium(2+)	0.70	False
10	creatine zwitterion	0.48	agmatinium(2+)	0.70	False
10	O-acylcarnitine	0.56	agmatinium(2+)	0.70	False
11	oxalate(2-)	0.34	creatine zwitterion	0.49	False
11	succinate(2-)	0.37	creatine zwitterion	0.49	False
11	acetate	0.45	creatine zwitterion	0.49	False
11	allantoate	0.46	creatine zwitterion	0.49	False
11	O-acylcarnitine	0.46	creatine zwitterion	0.49	False
11	L-aspartate(1-)	0.47	creatine zwitterion	0.49	False
11	glycerol 1-phosphodiester(1-)	0.48	creatine zwitterion	0.49	False
11	creatine zwitterion	0.49	creatine zwitterion	0.49	True
11	glycerol	0.49	creatine zwitterion	0.49	False
11	nicotinate	0.49	creatine zwitterion	0.49	False
12	oxalate(2-)	0.33	creatine zwitterion	0.49	False
12	succinate(2-)	0.36	creatine zwitterion	0.49	False
12	acetate	0.45	creatine zwitterion	0.49	False
12	allantoate	0.46	creatine zwitterion	0.49	False
12	L-aspartate(1-)	0.46	creatine zwitterion	0.49	False
12	O-acylcarnitine	0.47	creatine zwitterion	0.49	False
12	glycerol 1-phosphodiester(1-)	0.48	creatine zwitterion	0.49	False
12	nicotinate	0.49	creatine zwitterion	0.49	False
12	creatine zwitterion	0.49	creatine zwitterion	0.49	True
12	glycerol	0.49	creatine zwitterion	0.49	False
13	acadesine	0.42	acetate	1.10	False
13	O-acylcarnitine	0.43	acetate	1.10	False
13	methyl beta-D-galactoside	0.45	acetate	1.10	False
13	cytidine	0.46	acetate	1.10	False
13	hexuronic acid	0.52	acetate	1.10	False
13	galactitol	0.54	acetate	1.10	False
13	(+)-abscisic acid D-glucopyranosyl ester	0.54	acetate	1.10	False
13	organic phosphonate	0.56	acetate	1.10	False
13	tetracycline zwitterion	0.57	acetate	1.10	False
13	D-galactofuranose	0.62	acetate	1.10	False
14	oxalate(2-)	0.29	mevalonate	0.92	False
14	succinate(2-)	0.32	mevalonate	0.92	False
14	acetate	0.41	mevalonate	0.92	False
14	nicotinate	0.44	mevalonate	0.92	False
14	benzoate	0.45	mevalonate	0.92	False
14	L-aspartate(1-)	0.45	mevalonate	0.92	False
14	allantoate	0.47	mevalonate	0.92	False
14	glycerol	0.48	mevalonate	0.92	False
14	creatine zwitterion	0.48	mevalonate	0.92	False
14	L-serine zwitterion	0.48	mevalonate	0.92	False
15	ferrienterobactin(3-)	0.36	ferrienterobactin(3-)	0.36	True

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
15	tetracycline zwitterion	0.41	ferrienterobactin(3-)	0.36	False
15	microcin	0.41	ferrienterobactin(3-)	0.36	False
15	(+)-abscisic acid D-glucopyranosyl ester	0.45	ferrienterobactin(3-)	0.36	False
15	coenzyme A(4-)	0.48	ferrienterobactin(3-)	0.36	False
15	bilirubin(2-)	0.52	ferrienterobactin(3-)	0.36	False
15	enterobactin(1-)	0.60	ferrienterobactin(3-)	0.36	False
15	cytidine	0.62	ferrienterobactin(3-)	0.36	False
15	hexuronic acid	0.64	ferrienterobactin(3-)	0.36	False
15	acadesine	0.67	ferrienterobactin(3-)	0.36	False
16	L-aspartate(1-)	0.19	hexuronic acid	0.95	False
16	L-serine zwitterion	0.29	hexuronic acid	0.95	False
16	L-glutamine zwitterion	0.35	hexuronic acid	0.95	False
16	L-phenylalanine zwitterion	0.38	hexuronic acid	0.95	False
16	L-tyrosine zwitterion	0.39	hexuronic acid	0.95	False
16	L-leucine zwitterion	0.39	hexuronic acid	0.95	False
16	L-threonine zwitterion	0.40	hexuronic acid	0.95	False
16	S-methyl-L-methionine zwitterion	0.41	hexuronic acid	0.95	False
16	O-acylcarnitine	0.45	hexuronic acid	0.95	False
16	creatine zwitterion	0.46	hexuronic acid	0.95	False
17	L-tyrosine zwitterion	0.31	S-methyl-L-methionine zwitterion	0.56	False
17	L-phenylalanine zwitterion	0.40	S-methyl-L-methionine zwitterion	0.56	False
17	L-glutamine zwitterion	0.43	S-methyl-L-methionine zwitterion	0.56	False
17	L-serine zwitterion	0.44	S-methyl-L-methionine zwitterion	0.56	False
17	L-leucine zwitterion	0.50	S-methyl-L-methionine zwitterion	0.56	False
17	L-aspartate(1-)	0.51	S-methyl-L-methionine zwitterion	0.56	False
17	L-threonine zwitterion	0.56	S-methyl-L-methionine zwitterion	0.56	False
17	S-methyl-L-methionine zwitterion	0.56	S-methyl-L-methionine zwitterion	0.56	True
17	O-acylcarnitine	0.74	S-methyl-L-methionine zwitterion	0.56	False
17	creatine zwitterion	0.75	S-methyl-L-methionine zwitterion	0.56	False
18	L-serine zwitterion	0.33	L-leucine zwitterion	0.51	False
18	oxalate(2-)	0.35	L-leucine zwitterion	0.51	False
18	succinate(2-)	0.37	L-leucine zwitterion	0.51	False
18	nicotinate	0.39	L-leucine zwitterion	0.51	False
18	sulfite	0.42	L-leucine zwitterion	0.51	False
18	benzoate	0.42	L-leucine zwitterion	0.51	False
18	L-aspartate(1-)	0.43	L-leucine zwitterion	0.51	False
18	L-phenylalanine zwitterion	0.43	L-leucine zwitterion	0.51	False
18	creatine zwitterion	0.43	L-leucine zwitterion	0.51	False
18	nitrite	0.43	L-leucine zwitterion	0.51	False
19	silicic acid	0.29	arsenite(1-)	0.45	False
19	arsenite(1-)	0.45	arsenite(1-)	0.45	True
19	oxalate(2-)	0.51	arsenite(1-)	0.45	False
19	sulfite	0.51	arsenite(1-)	0.45	False
19	silver cation	0.53	arsenite(1-)	0.45	False
19	succinate(2-)	0.55	arsenite(1-)	0.45	False
19	nitrite	0.57	arsenite(1-)	0.45	False
19	organic phosphonate	0.57	arsenite(1-)	0.45	False

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
19	manganese cation	0.58	arsenite(1-)	0.45	False
19	L-serine zwitterion	0.60	arsenite(1-)	0.45	False
20	coenzyme A(4-)	0.26	GDP-mannose	0.53	False
20	(+)-abscisic acid D-glucopyranosyl ester	0.39	GDP-mannose	0.53	False
20	microcin	0.42	GDP-mannose	0.53	False
20	tetracycline zwitterion	0.42	GDP-mannose	0.53	False
20	ferrienterobactin(3-)	0.46	GDP-mannose	0.53	False
20	enterobactin(1-)	0.49	GDP-mannose	0.53	False
20	GDP-mannose	0.53	GDP-mannose	0.53	True
20	cytidine	0.54	GDP-mannose	0.53	False
20	acadesine	0.58	GDP-mannose	0.53	False
20	bilirubin(2-)	0.59	GDP-mannose	0.53	False
21	O-acylcarnitine	0.42	acetate	1.06	False
21	acadesine	0.42	acetate	1.06	False
21	cytidine	0.46	acetate	1.06	False
21	methyl beta-D-galactoside	0.46	acetate	1.06	False
21	hexuronic acid	0.53	acetate	1.06	False
21	galactitol	0.54	acetate	1.06	False
21	(+)-abscisic acid D-glucopyranosyl ester	0.55	acetate	1.06	False
21	organic phosphonate	0.55	acetate	1.06	False
21	tetracycline zwitterion	0.57	acetate	1.06	False
21	D-galactofuranose	0.63	acetate	1.06	False
22	O-acylcarnitine	0.50	manganese cation	1.07	False
22	sulfite	0.51	manganese cation	1.07	False
22	arsenite(1-)	0.53	manganese cation	1.07	False
22	nitrite	0.55	manganese cation	1.07	False
22	oxalate(2-)	0.56	manganese cation	1.07	False
22	organic phosphonate	0.59	manganese cation	1.07	False
22	succinate(2-)	0.60	manganese cation	1.07	False
22	L-aspartate(1-)	0.61	manganese cation	1.07	False
22	creatine zwitterion	0.63	manganese cation	1.07	False
22	L-serine zwitterion	0.64	manganese cation	1.07	False
23	L-glutamine zwitterion	0.24	O-acylcarnitine	0.59	False
23	L-serine zwitterion	0.27	O-acylcarnitine	0.59	False
23	L-aspartate(1-)	0.28	O-acylcarnitine	0.59	False
23	L-leucine zwitterion	0.31	O-acylcarnitine	0.59	False
23	L-phenylalanine zwitterion	0.37	O-acylcarnitine	0.59	False
23	L-tyrosine zwitterion	0.39	O-acylcarnitine	0.59	False
23	creatine zwitterion	0.43	O-acylcarnitine	0.59	False
23	spermine(4+)	0.44	O-acylcarnitine	0.59	False
23	L-threonine zwitterion	0.45	O-acylcarnitine	0.59	False
23	S-methyl-L-methionine zwitterion	0.46	O-acylcarnitine	0.59	False
24	coenzyme A(4-)	0.29	coenzyme A(4-)	0.29	True
24	microcin	0.41	coenzyme A(4-)	0.29	False
24	(+)-abscisic acid D-glucopyranosyl ester	0.41	coenzyme A(4-)	0.29	False

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
24	tetracycline zwitterion	0.43	coenzyme A(4-)	0.29	False
24	ferrienterobactin(3-)	0.50	coenzyme A(4-)	0.29	False
24	GDP-mannose	0.50	coenzyme A(4-)	0.29	False
24	enterobactin(1-)	0.50	coenzyme A(4-)	0.29	False
24	acadesine	0.51	coenzyme A(4-)	0.29	False
24	cytidine	0.52	coenzyme A(4-)	0.29	False
24	bilirubin(2-)	0.59	coenzyme A(4-)	0.29	False
25	benzoate	0.44	enterobactin(1-)	1.79	False
25	nicotinate	0.45	enterobactin(1-)	1.79	False
25	oxalate(2-)	0.48	enterobactin(1-)	1.79	False
25	sulfite	0.48	enterobactin(1-)	1.79	False
25	nitrite	0.50	enterobactin(1-)	1.79	False
25	succinate(2-)	0.50	enterobactin(1-)	1.79	False
25	L-serine zwitterion	0.53	enterobactin(1-)	1.79	False
25	acetate	0.57	enterobactin(1-)	1.79	False
25	L-phenylalanine zwitterion	0.59	enterobactin(1-)	1.79	False
25	silver cation	0.62	enterobactin(1-)	1.79	False
26	L-aspartate(1-)	0.22	(R)-pantothenate	0.71	False
26	L-serine zwitterion	0.26	(R)-pantothenate	0.71	False
26	L-phenylalanine zwitterion	0.39	(R)-pantothenate	0.71	False
26	L-glutamine zwitterion	0.40	(R)-pantothenate	0.71	False
26	L-tyrosine zwitterion	0.42	(R)-pantothenate	0.71	False
26	L-leucine zwitterion	0.42	(R)-pantothenate	0.71	False
26	L-threonine zwitterion	0.43	(R)-pantothenate	0.71	False
26	S-methyl-L-methionine zwitterion	0.43	(R)-pantothenate	0.71	False
26	O-acylcarnitine	0.44	(R)-pantothenate	0.71	False
26	glycerol 1-phosphodiester(1-)	0.45	(R)-pantothenate	0.71	False
27	sulfite	0.52	glycerol	0.74	False
27	arsenite(1-)	0.54	glycerol	0.74	False
27	succinate(2-)	0.60	glycerol	0.74	False
27	nitrite	0.61	glycerol	0.74	False
27	oxalate(2-)	0.62	glycerol	0.74	False
27	silicic acid	0.63	glycerol	0.74	False
27	glycerol 1-phosphodiester(1-)	0.64	glycerol	0.74	False
27	L-serine zwitterion	0.64	glycerol	0.74	False
27	phytochelatin	0.72	glycerol	0.74	False
27	L-aspartate(1-)	0.72	glycerol	0.74	False
28	cytidine	0.30	uracil	1.23	False
28	acadesine	0.42	uracil	1.23	False
28	methyl beta-D-galactoside	0.45	uracil	1.23	False
28	hexuronic acid	0.52	uracil	1.23	False
28	O-acylcarnitine	0.53	uracil	1.23	False
28	galactitol	0.53	uracil	1.23	False
28	(+)-abscisic acid D-glucopyranosyl ester	0.57	uracil	1.23	False
28	D-galactofuranose	0.57	uracil	1.23	False
28	tetracycline zwitterion	0.62	uracil	1.23	False
28	microcin	0.64	uracil	1.23	False

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
29	benzoate	0.46	glycerol 1-phosphodiester(1-)	0.72	False
29	nicotinate	0.47	glycerol 1-phosphodiester(1-)	0.72	False
29	oxalate(2-)	0.51	glycerol 1-phosphodiester(1-)	0.72	False
29	sulfite	0.52	glycerol 1-phosphodiester(1-)	0.72	False
29	nitrite	0.52	glycerol 1-phosphodiester(1-)	0.72	False
29	succinate(2-)	0.54	glycerol 1-phosphodiester(1-)	0.72	False
29	L-serine zwitterion	0.58	glycerol 1-phosphodiester(1-)	0.72	False
29	acetate	0.60	glycerol 1-phosphodiester(1-)	0.72	False
29	L-phenylalanine zwitterion	0.63	glycerol 1-phosphodiester(1-)	0.72	False
29	silver cation	0.64	glycerol 1-phosphodiester(1-)	0.72	False
30	benzoate	0.46	glycerol 1-phosphodiester(1-)	0.72	False
30	nicotinate	0.47	glycerol 1-phosphodiester(1-)	0.72	False
30	oxalate(2-)	0.51	glycerol 1-phosphodiester(1-)	0.72	False
30	sulfite	0.52	glycerol 1-phosphodiester(1-)	0.72	False
30	nitrite	0.52	glycerol 1-phosphodiester(1-)	0.72	False
30	succinate(2-)	0.54	glycerol 1-phosphodiester(1-)	0.72	False
30	L-serine zwitterion	0.58	glycerol 1-phosphodiester(1-)	0.72	False
30	acetate	0.60	glycerol 1-phosphodiester(1-)	0.72	False
30	L-phenylalanine zwitterion	0.63	glycerol 1-phosphodiester(1-)	0.72	False
30	silver cation	0.64	glycerol 1-phosphodiester(1-)	0.72	False
31	O-acylcarnitine	0.38	fluconazole	1.12	False
31	L-serine zwitterion	0.42	fluconazole	1.12	False
31	L-aspartate(1-)	0.42	fluconazole	1.12	False
31	oxalate(2-)	0.42	fluconazole	1.12	False
31	succinate(2-)	0.44	fluconazole	1.12	False
31	L-leucine zwitterion	0.45	fluconazole	1.12	False
31	L-threonine zwitterion	0.48	fluconazole	1.12	False
31	S-methyl-L-methionine zwitterion	0.48	fluconazole	1.12	False
31	glycerol 1-phosphodiester(1-)	0.48	fluconazole	1.12	False
31	benzoate	0.50	fluconazole	1.12	False
32	microcin	0.34	bilirubin(2-)	0.55	False
32	tetracycline zwitterion	0.39	bilirubin(2-)	0.55	False
32	(+)-abscisic acid D-glucopyranosyl ester	0.47	bilirubin(2-)	0.55	False
32	cytidine	0.49	bilirubin(2-)	0.55	False
32	enterobactin(1-)	0.49	bilirubin(2-)	0.55	False
32	coenzyme A(4-)	0.49	bilirubin(2-)	0.55	False
32	acadesine	0.49	bilirubin(2-)	0.55	False
32	ferrienterobactin(3-)	0.50	bilirubin(2-)	0.55	False
32	bilirubin(2-)	0.55	bilirubin(2-)	0.55	True
32	hexuronic acid	0.60	bilirubin(2-)	0.55	False
33	tetracycline zwitterion	0.37	nickel cation	1.89	False
33	coenzyme A(4-)	0.41	nickel cation	1.89	False
33	ferrienterobactin(3-)	0.42	nickel cation	1.89	False
33	microcin	0.43	nickel cation	1.89	False
33	(+)-abscisic acid D-glucopyranosyl ester	0.43	nickel cation	1.89	False
33	enterobactin(1-)	0.51	nickel cation	1.89	False

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
33	acadesine	0.55	nickel cation	1.89	False
33	bilirubin(2-)	0.57	nickel cation	1.89	False
33	cytidine	0.57	nickel cation	1.89	False
33	hexuronic acid	0.62	nickel cation	1.89	False
34	O-acylcarnitine	0.37	methyl beta-D-galactoside	0.53	False
34	L-serine zwitterion	0.44	methyl beta-D-galactoside	0.53	False
34	L-threonine zwitterion	0.45	methyl beta-D-galactoside	0.53	False
34	L-aspartate(1-)	0.45	methyl beta-D-galactoside	0.53	False
34	S-methyl-L-methionine zwitterion	0.45	methyl beta-D-galactoside	0.53	False
34	organic phosphonate	0.47	methyl beta-D-galactoside	0.53	False
34	galactitol	0.47	methyl beta-D-galactoside	0.53	False
34	glycerol 1-phosphodiester(1-)	0.48	methyl beta-D-galactoside	0.53	False
34	methyl beta-D-galactoside	0.53	methyl beta-D-galactoside	0.53	True
34	L-phenylalanine zwitterion	0.55	methyl beta-D-galactoside	0.53	False
35	O-acylcarnitine	0.40	cytidine	0.43	False
35	cytidine	0.43	cytidine	0.43	True
35	methyl beta-D-galactoside	0.47	cytidine	0.43	False
35	galactitol	0.48	cytidine	0.43	False
35	organic phosphonate	0.52	cytidine	0.43	False
35	acadesine	0.56	cytidine	0.43	False
35	L-threonine zwitterion	0.57	cytidine	0.43	False
35	S-methyl-L-methionine zwitterion	0.58	cytidine	0.43	False
35	glycerol 1-phosphodiester(1-)	0.58	cytidine	0.43	False
35	hexuronic acid	0.59	cytidine	0.43	False
36	cytidine	0.34	cytosine	0.99	False
36	methyl beta-D-galactoside	0.43	cytosine	0.99	False
36	O-acylcarnitine	0.43	cytosine	0.99	False
36	acadesine	0.46	cytosine	0.99	False
36	galactitol	0.47	cytosine	0.99	False
36	hexuronic acid	0.51	cytosine	0.99	False
36	organic phosphonate	0.53	cytosine	0.99	False
36	D-galactofuranose	0.57	cytosine	0.99	False
36	thiamine(1+) diphosphate(3-)	0.67	cytosine	0.99	False
36	(+)-abscisic acid D-glucopyranosyl ester	0.68	cytosine	0.99	False
37	nitrite	0.33	2-cis-abscisate	1.54	False
37	oxalate(2-)	0.40	2-cis-abscisate	1.54	False
37	sulfite	0.42	2-cis-abscisate	1.54	False
37	nicotinate	0.48	2-cis-abscisate	1.54	False
37	succinate(2-)	0.48	2-cis-abscisate	1.54	False
37	acetate	0.48	2-cis-abscisate	1.54	False
37	benzoate	0.56	2-cis-abscisate	1.54	False
37	L-serine zwitterion	0.66	2-cis-abscisate	1.54	False
37	creatine zwitterion	0.69	2-cis-abscisate	1.54	False
37	arsenite(1-)	0.70	2-cis-abscisate	1.54	False
38	L-tyrosine zwitterion	0.32	L-glutamine zwitterion	0.34	False
38	L-phenylalanine zwitterion	0.34	L-glutamine zwitterion	0.34	False
38	L-glutamine zwitterion	0.34	L-glutamine zwitterion	0.34	True

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
38	L-serine zwitterion	0.37	L-glutamine zwitterion	0.34	False
38	S-methyl-L-methionine zwitterion	0.41	L-glutamine zwitterion	0.34	False
38	L-threonine zwitterion	0.41	L-glutamine zwitterion	0.34	False
38	L-aspartate(1-)	0.44	L-glutamine zwitterion	0.34	False
38	L-leucine zwitterion	0.46	L-glutamine zwitterion	0.34	False
38	spermine(4+)	0.55	L-glutamine zwitterion	0.34	False
38	O-acylcarnitine	0.56	L-glutamine zwitterion	0.34	False
39	nitrite	0.36	silver cation	0.47	False
39	sulfite	0.46	silver cation	0.47	False
39	silver cation	0.47	silver cation	0.47	True
39	benzoate	0.48	silver cation	0.47	False
39	oxalate(2-)	0.49	silver cation	0.47	False
39	mevalonate	0.50	silver cation	0.47	False
39	phytochelatin	0.50	silver cation	0.47	False
39	acetate	0.50	silver cation	0.47	False
39	nicotinate	0.51	silver cation	0.47	False
39	D-ribose	0.51	silver cation	0.47	False
40	benzoate	0.46	sulfite	0.52	False
40	nicotinate	0.47	sulfite	0.52	False
40	oxalate(2-)	0.51	sulfite	0.52	False
40	sulfite	0.52	sulfite	0.52	True
40	nitrite	0.52	sulfite	0.52	False
40	succinate(2-)	0.54	sulfite	0.52	False
40	L-serine zwitterion	0.58	sulfite	0.52	False
40	acetate	0.60	sulfite	0.52	False
40	L-phenylalanine zwitterion	0.62	sulfite	0.52	False
40	silver cation	0.64	sulfite	0.52	False
41	L-aspartate(1-)	0.19	allantoate	0.52	False
41	L-serine zwitterion	0.26	allantoate	0.52	False
41	L-glutamine zwitterion	0.36	allantoate	0.52	False
41	L-phenylalanine zwitterion	0.37	allantoate	0.52	False
41	L-tyrosine zwitterion	0.39	allantoate	0.52	False
41	L-leucine zwitterion	0.40	allantoate	0.52	False
41	L-threonine zwitterion	0.42	allantoate	0.52	False
41	S-methyl-L-methionine zwitterion	0.43	allantoate	0.52	False
41	creatine zwitterion	0.46	allantoate	0.52	False
41	O-acylcarnitine	0.47	allantoate	0.52	False
42	acadesine	0.49	oxalate(2-)	1.33	False
42	tetracycline zwitterion	0.55	oxalate(2-)	1.33	False
42	(+)-abscisic acid D-glucopyranosyl ester	0.57	oxalate(2-)	1.33	False
42	cytidine	0.58	oxalate(2-)	1.33	False
42	enterobactin(1-)	0.59	oxalate(2-)	1.33	False
42	coenzyme A(4-)	0.65	oxalate(2-)	1.33	False
42	hexuronic acid	0.67	oxalate(2-)	1.33	False
42	microcin	0.70	oxalate(2-)	1.33	False
42	methyl beta-D-galactoside	0.71	oxalate(2-)	1.33	False
42	O-acylcarnitine	0.72	oxalate(2-)	1.33	False

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
43	O-acylcarnitine	0.36	nicotinate	0.60	False
43	L-aspartate(1-)	0.37	nicotinate	0.60	False
43	L-serine zwitterion	0.37	nicotinate	0.60	False
43	S-methyl-L-methionine zwitterion	0.41	nicotinate	0.60	False
43	L-threonine zwitterion	0.42	nicotinate	0.60	False
43	glycerol 1-phosphodiester(1-)	0.44	nicotinate	0.60	False
43	L-leucine zwitterion	0.44	nicotinate	0.60	False
43	succinate(2-)	0.45	nicotinate	0.60	False
43	oxalate(2-)	0.46	nicotinate	0.60	False
43	arsenite(1-)	0.47	nicotinate	0.60	False
44	O-acylcarnitine	0.50	microcin	0.53	False
44	microcin	0.53	microcin	0.53	True
44	ferrienterobactin(3-)	0.54	microcin	0.53	False
44	tetracycline zwitterion	0.54	microcin	0.53	False
44	cytidine	0.56	microcin	0.53	False
44	hexuronic acid	0.57	microcin	0.53	False
44	organic phosphonate	0.58	microcin	0.53	False
44	(+)-abscisic acid D-glucopyranosyl ester	0.61	microcin	0.53	False
44	acadesine	0.62	microcin	0.53	False
44	methyl beta-D-galactoside	0.65	microcin	0.53	False
45	silver cation	0.46	L-serine zwitterion	0.67	False
45	nitrite	0.52	L-serine zwitterion	0.67	False
45	manganese cation	0.56	L-serine zwitterion	0.67	False
45	sulfite	0.57	L-serine zwitterion	0.67	False
45	phytochelatin	0.59	L-serine zwitterion	0.67	False
45	nickel cation	0.60	L-serine zwitterion	0.67	False
45	benzoate	0.63	L-serine zwitterion	0.67	False
45	mevalonate	0.63	L-serine zwitterion	0.67	False
45	D-ribose	0.63	L-serine zwitterion	0.67	False
45	nicotinate	0.64	L-serine zwitterion	0.67	False
46	sulfite	0.52	silicic acid	0.62	False
46	arsenite(1-)	0.54	silicic acid	0.62	False
46	succinate(2-)	0.60	silicic acid	0.62	False
46	nitrite	0.61	silicic acid	0.62	False
46	silicic acid	0.62	silicic acid	0.62	True
46	oxalate(2-)	0.63	silicic acid	0.62	False
46	glycerol 1-phosphodiester(1-)	0.63	silicic acid	0.62	False
46	L-serine zwitterion	0.64	silicic acid	0.62	False
46	phytochelatin	0.71	silicic acid	0.62	False
46	L-aspartate(1-)	0.71	silicic acid	0.62	False
47	mevalonate	0.13	D-galactofuranose	1.05	False
47	phytochelatin	0.17	D-galactofuranose	1.05	False
47	D-ribose	0.18	D-galactofuranose	1.05	False
47	poly-beta-1-6-N-acetyl-D-glucosamine	0.20	D-galactofuranose	1.05	False
47	manganese cation	0.38	D-galactofuranose	1.05	False
47	nickel cation	0.40	D-galactofuranose	1.05	False

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
47	sulfite	0.61	D-galactofuranose	1.05	False
47	acetate	0.63	D-galactofuranose	1.05	False
47	nitrite	0.64	D-galactofuranose	1.05	False
47	succinate(2-)	0.66	D-galactofuranose	1.05	False
48	coenzyme A(4-)	0.34	thiamine(1+) diphosphate(3-)	1.10	False
48	tetracycline zwitterion	0.36	thiamine(1+) diphosphate(3-)	1.10	False
48	(+)-abscisic acid D-glucopyranosyl ester	0.40	thiamine(1+) diphosphate(3-)	1.10	False
48	microcin	0.41	thiamine(1+) diphosphate(3-)	1.10	False
48	ferrienterobactin(3-)	0.50	thiamine(1+) diphosphate(3-)	1.10	False
48	enterobactin(1-)	0.53	thiamine(1+) diphosphate(3-)	1.10	False
48	bilirubin(2-)	0.56	thiamine(1+) diphosphate(3-)	1.10	False
48	acadesine	0.57	thiamine(1+) diphosphate(3-)	1.10	False
48	cytidine	0.59	thiamine(1+) diphosphate(3-)	1.10	False
48	hexuronic acid	0.65	thiamine(1+) diphosphate(3-)	1.10	False
49	oxalate(2-)	0.42	organic phosphonate	0.61	False
49	sulfite	0.43	organic phosphonate	0.61	False
49	O-acylcarnitine	0.45	organic phosphonate	0.61	False
49	L-serine zwitterion	0.46	organic phosphonate	0.61	False
49	L-aspartate(1-)	0.46	organic phosphonate	0.61	False
49	succinate(2-)	0.47	organic phosphonate	0.61	False
49	arsenite(1-)	0.47	organic phosphonate	0.61	False
49	creatine zwitterion	0.50	organic phosphonate	0.61	False
49	glycerol 1-phosphodiester(1-)	0.51	organic phosphonate	0.61	False
49	S-methyl-L-methionine zwitterion	0.52	organic phosphonate	0.61	False
50	L-serine zwitterion	0.39	benzoate	0.54	False
50	sulfite	0.43	benzoate	0.54	False
50	oxalate(2-)	0.44	benzoate	0.54	False
50	succinate(2-)	0.45	benzoate	0.54	False
50	L-aspartate(1-)	0.45	benzoate	0.54	False
50	O-acylcarnitine	0.46	benzoate	0.54	False
50	L-phenylalanine zwitterion	0.47	benzoate	0.54	False
50	glycerol 1-phosphodiester(1-)	0.49	benzoate	0.54	False
50	arsenite(1-)	0.49	benzoate	0.54	False
50	nicotinate	0.50	benzoate	0.54	False
51	cytidine	0.30	acadesine	0.41	False
51	acadesine	0.41	acadesine	0.41	True
51	methyl beta-D-galactoside	0.45	acadesine	0.41	False
51	hexuronic acid	0.52	acadesine	0.41	False
51	galactitol	0.53	acadesine	0.41	False
51	O-acylcarnitine	0.54	acadesine	0.41	False
51	D-galactofuranose	0.57	acadesine	0.41	False
51	(+)-abscisic acid D-glucopyranosyl ester	0.57	acadesine	0.41	False
51	tetracycline zwitterion	0.62	acadesine	0.41	False
51	microcin	0.64	acadesine	0.41	False
52	mevalonate	0.13	methyl beta-D-galactoside	1.19	False
52	phytochelatins	0.17	methyl beta-D-galactoside	1.19	False

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
52	D-ribose	0.18	methyl beta-D-galactoside	1.19	False
52	poly-beta-1-6-N-acetyl-D-glucosamine	0.20	methyl beta-D-galactoside	1.19	False
52	manganese cation	0.38	methyl beta-D-galactoside	1.19	False
52	nickel cation	0.40	methyl beta-D-galactoside	1.19	False
52	sulfite	0.64	methyl beta-D-galactoside	1.19	False
52	acetate	0.65	methyl beta-D-galactoside	1.19	False
52	succinate(2-)	0.67	methyl beta-D-galactoside	1.19	False
52	nitrite	0.68	methyl beta-D-galactoside	1.19	False
53	oxalate(2-)	0.38	tetracycline zwitterion	1.55	False
53	nitrite	0.40	tetracycline zwitterion	1.55	False
53	succinate(2-)	0.43	tetracycline zwitterion	1.55	False
53	sulfite	0.43	tetracycline zwitterion	1.55	False
53	benzoate	0.43	tetracycline zwitterion	1.55	False
53	acetate	0.44	tetracycline zwitterion	1.55	False
53	nicotinate	0.44	tetracycline zwitterion	1.55	False
53	L-serine zwitterion	0.44	tetracycline zwitterion	1.55	False
53	silver cation	0.45	tetracycline zwitterion	1.55	False
53	phytochelatin	0.46	tetracycline zwitterion	1.55	False
54	coenzyme A(4-)	0.35	thiamine(1+) diphosphate(3-)	1.12	False
54	tetracycline zwitterion	0.36	thiamine(1+) diphosphate(3-)	1.12	False
54	microcin	0.40	thiamine(1+) diphosphate(3-)	1.12	False
54	(+)-abscisic acid D-glucopyranosyl ester	0.42	thiamine(1+) diphosphate(3-)	1.12	False
54	enterobactin(1-)	0.48	thiamine(1+) diphosphate(3-)	1.12	False
54	ferrienterobactin(3-)	0.49	thiamine(1+) diphosphate(3-)	1.12	False
54	acadesine	0.53	thiamine(1+) diphosphate(3-)	1.12	False
54	GDP-mannose	0.57	thiamine(1+) diphosphate(3-)	1.12	False
54	cytidine	0.58	thiamine(1+) diphosphate(3-)	1.12	False
54	bilirubin(2-)	0.59	thiamine(1+) diphosphate(3-)	1.12	False
55	ferrienterobactin(3-)	0.50	nickel cation	1.50	False
55	tetracycline zwitterion	0.52	nickel cation	1.50	False
55	microcin	0.54	nickel cation	1.50	False
55	coenzyme A(4-)	0.55	nickel cation	1.50	False
55	(+)-abscisic acid D-glucopyranosyl ester	0.57	nickel cation	1.50	False
55	O-acylcarnitine	0.59	nickel cation	1.50	False
55	cytidine	0.60	nickel cation	1.50	False
55	hexuronic acid	0.61	nickel cation	1.50	False
55	methyl beta-D-galactoside	0.65	nickel cation	1.50	False
55	bilirubin(2-)	0.66	nickel cation	1.50	False
56	nitrite	0.38	galactitol	1.00	False
56	acetate	0.50	galactitol	1.00	False
56	creatine zwitterion	0.52	galactitol	1.00	False
56	allantoate	0.52	galactitol	1.00	False
56	oxalate(2-)	0.53	galactitol	1.00	False
56	nicotinate	0.55	galactitol	1.00	False
56	succinate(2-)	0.56	galactitol	1.00	False

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
56	sulfite	0.57	galactitol	1.00	False
56	L-serine zwitterion	0.57	galactitol	1.00	False
56	benzoate	0.58	galactitol	1.00	False
57	L-serine zwitterion	0.31	arsenite(1-)	0.52	False
57	oxalate(2-)	0.32	arsenite(1-)	0.52	False
57	succinate(2-)	0.33	arsenite(1-)	0.52	False
57	L-aspartate(1-)	0.37	arsenite(1-)	0.52	False
57	L-phenylalanine zwitterion	0.40	arsenite(1-)	0.52	False
57	L-glutamine zwitterion	0.42	arsenite(1-)	0.52	False
57	nicotinate	0.42	arsenite(1-)	0.52	False
57	creatine zwitterion	0.44	arsenite(1-)	0.52	False
57	allantoate	0.45	arsenite(1-)	0.52	False
57	O-acylcarnitine	0.46	arsenite(1-)	0.52	False
58	O-acylcarnitine	0.38	uracil	0.90	False
58	galactitol	0.44	uracil	0.90	False
58	methyl beta-D-galactoside	0.50	uracil	0.90	False
58	organic phosphonate	0.52	uracil	0.90	False
58	glycerol 1-phosphodiester(1-)	0.54	uracil	0.90	False
58	S-methyl-L-methionine zwitterion	0.56	uracil	0.90	False
58	L-threonine zwitterion	0.57	uracil	0.90	False
58	L-aspartate(1-)	0.57	uracil	0.90	False
58	L-serine zwitterion	0.58	uracil	0.90	False
58	D-galactofuranose	0.62	uracil	0.90	False
59	oxalate(2-)	0.31	L-tyrosine zwitterion	0.50	False
59	L-serine zwitterion	0.33	L-tyrosine zwitterion	0.50	False
59	succinate(2-)	0.33	L-tyrosine zwitterion	0.50	False
59	L-aspartate(1-)	0.39	L-tyrosine zwitterion	0.50	False
59	nicotinate	0.41	L-tyrosine zwitterion	0.50	False
59	L-phenylalanine zwitterion	0.42	L-tyrosine zwitterion	0.50	False
59	L-glutamine zwitterion	0.43	L-tyrosine zwitterion	0.50	False
59	creatine zwitterion	0.43	L-tyrosine zwitterion	0.50	False
59	sulfite	0.44	L-tyrosine zwitterion	0.50	False
59	acetate	0.45	L-tyrosine zwitterion	0.50	False
60	mevalonate	0.13	D-ribose	0.18	False
60	phytochelatin	0.17	D-ribose	0.18	False
60	D-ribose	0.18	D-ribose	0.18	True
60	poly-beta-1-6-N-acetyl-D-glucosamine	0.20	D-ribose	0.18	False
60	manganese cation	0.37	D-ribose	0.18	False
60	nickel cation	0.40	D-ribose	0.18	False
60	sulfite	0.64	D-ribose	0.18	False
60	acetate	0.65	D-ribose	0.18	False
60	succinate(2-)	0.66	D-ribose	0.18	False
60	nitrite	0.68	D-ribose	0.18	False
61	L-serine zwitterion	0.32	L-serine zwitterion	0.32	True
61	L-glutamine zwitterion	0.33	L-serine zwitterion	0.32	False
61	S-methyl-L-methionine zwitterion	0.34	L-serine zwitterion	0.32	False
61	L-threonine zwitterion	0.34	L-serine zwitterion	0.32	False

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
61	L-aspartate(1-)	0.39	L-serine zwitterion	0.32	False
61	L-tyrosine zwitterion	0.39	L-serine zwitterion	0.32	False
61	L-phenylalanine zwitterion	0.40	L-serine zwitterion	0.32	False
61	L-leucine zwitterion	0.41	L-serine zwitterion	0.32	False
61	O-acylcarnitine	0.50	L-serine zwitterion	0.32	False
61	creatine zwitterion	0.52	L-serine zwitterion	0.32	False
62	microcin	0.35	bilirubin(2-)	0.65	False
62	tetracycline zwitterion	0.44	bilirubin(2-)	0.65	False
62	cytidine	0.45	bilirubin(2-)	0.65	False
62	acadesine	0.46	bilirubin(2-)	0.65	False
62	coenzyme A(4-)	0.49	bilirubin(2-)	0.65	False
62	(+)-abscisic acid D-glucopyranosyl ester	0.50	bilirubin(2-)	0.65	False
62	enterobactin(1-)	0.52	bilirubin(2-)	0.65	False
62	ferrienterobactin(3-)	0.58	bilirubin(2-)	0.65	False
62	hexuronic acid	0.58	bilirubin(2-)	0.65	False
62	O-acylcarnitine	0.61	bilirubin(2-)	0.65	False
63	L-glutamine zwitterion	0.42	L-tyrosine zwitterion	0.54	False
63	S-methyl-L-methionine zwitterion	0.42	L-tyrosine zwitterion	0.54	False
63	L-aspartate(1-)	0.45	L-tyrosine zwitterion	0.54	False
63	agmatinium(2+)	0.45	L-tyrosine zwitterion	0.54	False
63	L-serine zwitterion	0.47	L-tyrosine zwitterion	0.54	False
63	O-acylcarnitine	0.48	L-tyrosine zwitterion	0.54	False
63	creatine zwitterion	0.50	L-tyrosine zwitterion	0.54	False
63	L-threonine zwitterion	0.51	L-tyrosine zwitterion	0.54	False
63	allantoate	0.52	L-tyrosine zwitterion	0.54	False
63	L-tyrosine zwitterion	0.54	L-tyrosine zwitterion	0.54	True
64	silver cation	0.46	manganese cation	0.56	False
64	nitrite	0.52	manganese cation	0.56	False
64	manganese cation	0.56	manganese cation	0.56	True
64	sulfite	0.57	manganese cation	0.56	False
64	phytochelatin	0.59	manganese cation	0.56	False
64	nickel cation	0.61	manganese cation	0.56	False
64	benzoate	0.63	manganese cation	0.56	False
64	mevalonate	0.64	manganese cation	0.56	False
64	D-ribose	0.64	manganese cation	0.56	False
64	nicotinate	0.64	manganese cation	0.56	False
65	tetracycline zwitterion	0.41	phytochelatin	1.77	False
65	bilirubin(2-)	0.43	phytochelatin	1.77	False
65	ferrienterobactin(3-)	0.46	phytochelatin	1.77	False
65	microcin	0.49	phytochelatin	1.77	False
65	(+)-abscisic acid D-glucopyranosyl ester	0.52	phytochelatin	1.77	False
65	enterobactin(1-)	0.56	phytochelatin	1.77	False
65	coenzyme A(4-)	0.67	phytochelatin	1.77	False
65	acadesine	0.77	phytochelatin	1.77	False
65	cytidine	0.78	phytochelatin	1.77	False
65	hexuronic acid	0.78	phytochelatin	1.77	False

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
66	L-tyrosine zwitterion	0.31	L-phenylalanine zwitterion	0.40	False
66	L-phenylalanine zwitterion	0.40	L-phenylalanine zwitterion	0.40	True
66	L-glutamine zwitterion	0.46	L-phenylalanine zwitterion	0.40	False
66	L-serine zwitterion	0.47	L-phenylalanine zwitterion	0.40	False
66	L-aspartate(1-)	0.54	L-phenylalanine zwitterion	0.40	False
66	L-leucine zwitterion	0.55	L-phenylalanine zwitterion	0.40	False
66	L-threonine zwitterion	0.55	L-phenylalanine zwitterion	0.40	False
66	S-methyl-L-methionine zwitterion	0.56	L-phenylalanine zwitterion	0.40	False
66	O-acylcarnitine	0.72	L-phenylalanine zwitterion	0.40	False
66	creatine zwitterion	0.77	L-phenylalanine zwitterion	0.40	False
67	allantoate	0.60	adenine	0.72	False
67	O-acylcarnitine	0.62	adenine	0.72	False
67	L-aspartate(1-)	0.67	adenine	0.72	False
67	S-methyl-L-methionine zwitterion	0.68	adenine	0.72	False
67	L-serine zwitterion	0.68	adenine	0.72	False
67	cytosine	0.69	adenine	0.72	False
67	creatine zwitterion	0.70	adenine	0.72	False
67	adenine	0.72	adenine	0.72	True
67	L-leucine zwitterion	0.73	adenine	0.72	False
67	organic phosphonate	0.74	adenine	0.72	False
68	microcin	0.35	(+)-abscisic acid D-glucopyranosyl ester	0.51	False
68	tetracycline zwitterion	0.45	(+)-abscisic acid D-glucopyranosyl ester	0.51	False
68	cytidine	0.46	(+)-abscisic acid D-glucopyranosyl ester	0.51	False
68	acadesine	0.46	(+)-abscisic acid D-glucopyranosyl ester	0.51	False
68	coenzyme A(4-)	0.49	(+)-abscisic acid D-glucopyranosyl ester	0.51	False
68	(+)-abscisic acid D-glucopyranosyl ester	0.51	(+)-abscisic acid D-glucopyranosyl ester	0.51	True
68	enterobactin(1-)	0.52	(+)-abscisic acid D-glucopyranosyl ester	0.51	False
68	hexuronic acid	0.58	(+)-abscisic acid D-glucopyranosyl ester	0.51	False
68	ferrienterobactin(3-)	0.59	(+)-abscisic acid D-glucopyranosyl ester	0.51	False
68	O-acylcarnitine	0.61	(+)-abscisic acid D-glucopyranosyl ester	0.51	False
69	L-aspartate(1-)	0.30	allantoate	0.58	False
69	L-serine zwitterion	0.32	allantoate	0.58	False
69	oxalate(2-)	0.34	allantoate	0.58	False
69	succinate(2-)	0.35	allantoate	0.58	False
69	sulfite	0.35	allantoate	0.58	False
69	arsenite(1-)	0.40	allantoate	0.58	False
69	glycerol 1-phosphodiester(1-)	0.43	allantoate	0.58	False
69	L-phenylalanine zwitterion	0.45	allantoate	0.58	False

Continued on next page

Sample #	Predicted Label Name	Distance	True Label Name	Distance To True Label	Is prediction correct
69	creatine zwitterion	0.49	allantoate	0.58	False
69	L-glutamine zwitterion	0.50	allantoate	0.58	False

This table lists 10 substrate candidates predicted by TooTranslator for each test sample in TS2. The third column shows the distance between each candidate substrate and the test sample. The “True Label Name” column displays the actual label of the test sample in TS2, with its corresponding distance from the test sample provided in the fifth column. The last column indicates whether the candidate substrate predicted by TooTranslator matches the true label of the test sample.

F.2 Inorganic ion transporter prototypes

Table 71: Results of TooTranslator model trained on UniProt-ICAT-100 with label definition

Substrate	Trainset	Testset	TP	FP	FN	TN	Acc.	Prec.	Rec.	F1	MCC
proton	837	261	249	8	12	554	0.976	0.969	0.954	0.961	0.944
calcium(2+)	555	174	165	4	9	645	0.984	0.976	0.948	0.962	0.952
potassium(1+)	465	145	139	3	6	675	0.989	0.979	0.959	0.969	0.962
chloride	293	92	87	5	5	726	0.99	0.95	0.95	0.95	0.94
sodium(1+)	274	86	79	8	7	729	0.98	0.91	0.92	0.91	0.90
sulfate	62	20	18	6	2	797	0.99	0.75	0.90	0.82	0.82
zinc(2+)	51	16	16	2	0	805	1.00	0.89	1.00	0.94	0.94
ammonium	35	11	10	2	1	810	1.00	0.83	0.91	0.87	0.87
nitrate	18	6	5	4	1	813	1.0	0.6	0.8	0.7	0.7
iron(2+)	15	5	3	3	2	815	1.0	0.5	0.6	0.5	0.5
phosphate ion	15	4	4	0	0	819	1.0	1.0	1.0	1.0	1.0
copper(1+)	11	3	3	0	0	820	1.0	1.0	1.0	1.0	1.0
Overall	2631	823	778	-	-	-	0.991	0.945	0.914	0.883	0.932

The table presents the results of the UniProt-ICAT-100 hold-out test set using a TooTranslator model trained on the training set of the same dataset, containing inorganic cation and anion transporters. The wMSE loss function is used for training. The labels are represented by the label definitions from ChEBI encoded using SciBERT language model. The model is referred to as prototype-1 in Section 6.3.5.

Table 72: Results of TooTranslator model trained on UniProt-ICAT-100 with label SMILES string vector

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	837	261	255	8	6	554	0.983	0.970	0.977	0.973	0.961
calcium(2+)	555	174	168	8	6	641	0.983	0.955	0.966	0.960	0.949
potassium(1+)	465	145	141	7	4	671	0.987	0.953	0.972	0.962	0.954
chloride	293	92	89	0	3	731	1.00	1.00	0.97	0.98	0.98
sodium(1+)	274	86	82	4	4	733	0.99	0.95	0.95	0.95	0.95
sulfate	62	20	18	0	2	803	1.00	1.00	0.90	0.95	0.95
zinc(2+)	51	16	16	0	0	807	1.00	1.00	1.00	1.00	1.00

Continued on next page

Table 72: Results of TooTranslator model trained on UniProt-ICAT-100 with label SMILES string vector

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
ammonium	35	11	11	0	0	812	1.00	1.00	1.00	1.00	1.00
nitrate	18	6	6	1	0	816	1.0	0.9	1.0	0.9	0.9
iron(2+)	15	5	2	0	3	818	1.0	1.0	0.4	0.6	0.6
phosphate ion	15	4	4	0	0	819	1.0	1.0	1.0	1.0	1.0
copper(1+)	11	3	3	0	0	820	1.0	1.0	1.0	1.0	1.0
Overall	2631	823	795	-	-	-	0.994	0.966	0.928	0.94	0.957

The table presents the results of the UniProt-ICAT-100 hold-out test set using a TooTranslator model trained on the training set of the same dataset, containing inorganic cation and anion transporters. The wMSE loss function is used for training. The labels are represented by the compound SMILES strings encoded by ChemBERTa-2. The model is referred to as prototype-2 in Section 6.3.5.

Table 73: Results of TooTranslator model trained on UniProt-ICAT-100 with label definition and SMILES string summation vector

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	837	261	252	6	9	556	0.982	0.977	0.966	0.971	0.958
calcium(2+)	555	174	166	2	8	647	0.988	0.988	0.954	0.971	0.963
potassium(1+)	465	145	139	0	6	678	0.993	1.000	0.959	0.979	0.975
chloride	293	92	89	9	3	722	0.98	0.91	0.97	0.94	0.93
sodium(1+)	274	86	83	15	3	722	0.98	0.85	0.96	0.90	0.89
sulfate	62	20	18	1	2	802	1.00	0.95	0.90	0.92	0.92
zinc(2+)	51	16	16	0	0	807	1.00	1.00	1.00	1.00	1.00
ammonium	35	11	11	0	0	812	1.00	1.00	1.00	1.00	1.00
nitrate	18	6	5	1	1	816	1.0	0.8	0.8	0.8	0.8
iron(2+)	15	5	3	0	2	818	1.0	1.0	0.6	0.8	0.8
phosphate ion	15	4	4	0	0	819	1.0	1.0	1.0	1.0	1.0
copper(1+)	11	3	3	0	0	820	1.0	1.0	1.0	1.0	1.0
Overall	2631	823	789	34	34	9019	0.993	0.959	0.929	0.939	0.949

The table presents the results of the UniProt-ICAT-100 hold-out test set using a TooTranslator model trained on the training set of the same dataset, containing inorganic cation and anion transporters. The wMSE loss function is used for training. The labels are represented by a summation of label definitions from ChEBI encoded using SciBERT language model and SMILES strings encoded by ChemBERTa-2. The model is referred to as prototype-3 in Section 6.3.5.

Table 74: Results of TooTranslator model trained on UniProt-ICAT-100 with label definition and SMILES string concatenation vector

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	837	261	253	6	8	556	0.983	0.977	0.969	0.973	0.961
calcium(2+)	555	174	168	2	6	647	0.990	0.988	0.966	0.977	0.971
potassium(1+)	465	145	139	4	6	674	0.988	0.972	0.959	0.965	0.958
chloride	293	92	90	8	2	723	0.99	0.92	0.98	0.95	0.94
sodium(1+)	274	86	83	3	3	734	0.99	0.96	0.96	0.96	0.96
sulfate	62	20	18	3	2	800	0.99	0.86	0.90	0.88	0.88

Continued on next page

Table 74: Results of TooTranslator model trained on UniProt-ICAT-100 with label definition and SMILES string concatenation vector

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
zinc(2+)	51	16	16	0	0	807	1.00	1.00	1.00	1.00	1.00
ammonium	35	11	11	0	0	812	1.00	1.00	1.00	1.00	1.00
nitrate	18	6	5	4	1	813	1.0	0.6	0.8	0.7	0.7
iron(2+)	15	5	3	0	2	818	1.0	1.0	0.6	0.8	0.8
phosphate ion	15	4	4	0	0	819	1.0	1.0	1.0	1.0	1.0
copper(1+)	11	3	3	0	0	820	1.0	1.0	1.0	1.0	1.0
Overall	2631	823	793	30	30	9023	0.994	0.964	0.931	0.927	0.955

The table presents the results of the UniProt-ICAT-100 hold-out test set using a TooTranslator model trained on the training set of the same dataset, containing inorganic cation and anion transporters. The wMSE loss function is used for training. The labels are represented by a concatenation of label definitions from **ChEBI** encoded using the SciBERT language model and SMILES strings encoded by ChemBERTa-2. The model is referred to as prototype-4 in Section 6.3.5.

Table 75: Results of TooTranslator model with frozen weights label definition and SMILES string concatenation vector

Substrate	Trainset	Testset	TP	FP	FN	TN	Accuracy	Precision	Recall	F1-Score	MCC
proton	837	261	0	0	261	562	0.683	NaN	0.000	0.000	NaN
calcium(2+)	555	174	0	0	174	649	0.789	NaN	0.000	0.000	NaN
potassium(1+)	465	145	0	0	145	678	0.824	NaN	0.000	0.000	NaN
chloride	293	92	0	0	92	731	0.89	NaN	0.00	0.00	NaN
sodium(1+)	274	86	0	0	86	737	0.90	NaN	0.00	0.00	NaN
sulfate	62	20	0	0	20	803	0.98	NaN	0.00	0.00	NaN
zinc(2+)	51	16	0	0	16	807	0.98	NaN	0.00	0.00	NaN
ammonium	35	11	0	0	11	812	0.99	NaN	0.00	0.00	NaN
nitrate	18	6	0	0	6	817	1.0	NaN	0.0	0.0	NaN
iron(2+)	15	5	0	0	5	818	1.0	NaN	0.0	0.0	NaN
phosphate ion	15	4	4	819	0	0	0.0	0.0	1.0	0.0	NaN
copper(1+)	11	3	0	0	3	820	1.0	NaN	0.0	0.0	NaN
Overall	2631	823	4	-	-	-	0.834	0.005	0.083	0.001	-

The table presents the results of the UniProt-ICAT-100 hold-out test set using a TooTranslator model with Frozen weights without any training. The labels are represented by a summation of label definitions from **ChEBI** encoded using the SciBERT language model and SMILES strings encoded by ChemBERTa-2. The model is referred to as prototype-4 in Section 6.3.5.