Statistical Analysis of Energy Measures as Biomarkers of SARS-Covid-2 Variants and Receptors

Khawla Ghannoum Al Chawaf

A Thesis

In the Department

of

Business Analytics and Technology Management

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Science (Business Analytics and Technology Management) at

Concordia University

Montréal, Québec, Canada

April 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Khawla Ghannoum Al Chawaf, 2025

Entitled: Statistical Analysis of Energy Measures as Biomarkers of SARS-Covid-2 Variants and Receptors

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

Master of Science (Business Analytics and Technology Management)

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

Signed by the final Examining Committee:

_____ Chair
Dr. *Arman Sadreddin*

_____ Examiner
Dr. *Dongliang Sheng*

_____ Examiner
Dr. *Danielle Morin*

_____ Supervisor
Dr. *Salim Lahmiri*

Approved by _____
Dr. Suchit Ahuja, Graduate Program Director

April 2025 _____
Dean of Faculty Anne-Marie Croteau

Abstract

Statistical Analysis of Energy Measures as Biomarkers of SARS-Covid-2 Variants and Receptors

Khawla Ghannoum Al Chawaf

The COVID-19 outbreak has made it evident that the nature and behavior of SARS-CoV-2 require constant research and surveillance, owing to the high mutation rates that lead to variants. This work focuses on the statistical analysis of energy measures as biomarkers of SARS-CoV-2. Thus, three statistical tests are applied to the data: the multiple ANOVA test for equality of means, Bartlett's test for equality of variances, and Levene's test for assessing the homogeneity of variances. These tests aim to determine which energy measure can differentiate between SARS-CoV-2 variants, human cell receptors (GRP78 and ACE2), and their combinations.

To further investigate the specific pairwise differences between groups, Tukey's HSD test was performed after the ANOVA. The Tukey test provided adjusted p-values (p-adj) for each pairwise comparison, allowing for a more detailed understanding of significant differences in energy measures across variants, receptors, and their combinations.

The proposed approach combines energy measures and sequence data to develop classification systems and brings out the variety of the virus' genetics and interaction mechanisms. This work aims to improve the accuracy of variant identification and contribute to creating tailored interventions, which would help address the COVID-19 issue and contribute considerably to the global fight against the pandemic.

Keywords: SARS-CoV-2, COVID-19, statistical analysis, variant identification, human receptors, genetic sequences, ANOVA test, Bartlett's test, Levene's Test, Tukey's HSD test.

Acknowledgments

This thesis marks the culmination of an incredible journey, and I am deeply grateful to those who have supported and guided me along the way.

First, I sincerely thank my advisor, **Professor Salim Lahmiri**, for his invaluable guidance, patience, and encouragement throughout this research. His insights and expertise have been instrumental in shaping this work. I would also like to thank my examiners, **Professor Dongliang Sheng and Professor Danielle Morin**, and my defense Chair, Professor Arman Sadreddin, for their time, thoughtful feedback, and constructive discussions that helped refine my research further.

A special thank you to our Program Director **Professor Suchit Ahuja,** and the **Business Analytics and Technology Management department at Concordia University** for providing me with the knowledge and resources to thrive in this field.

To my family—**my dad, mom, and my three sisters**—your unwavering love and support have been my pillar of strength. Every late night and moment of self-doubt was made easier knowing I had you all cheering me on.

And finally, to my fiancé, **Bilal**, your encouragement to step into the world of data analytics has truly changed my path, and I am grateful to have you by my side, both as a partner and as someone who shares my passion for this field.

This journey would not have been possible without each of you. Thank you from the bottom of my heart.

Table of Contents

List of Figures

List of Tables

Chapter 1

1.1 Introduction

The scientific community faces a formidable foe in the relentless battle against the global COVID-19 pandemic: the ever-evolving severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The virus's unrelenting advance over the last three years has been fueled by its infectious nature and its capacity for constant mutation. These genetic differences have emerged as important determinants of the biological activities of the virus, regulating immune evasion and transmission dynamics, especially within the amino acid sequence of the surface spike (S) protein. As a result, the discovery and comprehension of these changes have become essential for developing focused treatment plans and implementing exact preventative and control measures around the globe.

The ongoing evolution of SARS-CoV-2 has led to the emergence of multiple variants, each exhibiting distinct transmissibility and pathogenicity profiles. Developing successful treatment plans requires an understanding of the molecular interactions between these variations and host cell receptors. The angiotensin-converting enzyme 2 (ACE2) is the main receptor that allows SARS-CoV-2 to enter host cells, but new research has shown that the glucose-regulated protein 78 (GRP78) is another receptor that may affect viral infectivity (Ibrahim et al., 2020).

Understanding the mechanisms behind viral entrance and infection can be gained by examining the binding affinities and interaction energies between SARS-CoV-2 variants and these receptors. Prior studies have evaluated the binding free energies of several SARS-CoV-2 variants with ACE2 using computational techniques, identifying variations in binding affinities that could be associated with higher transmissibility (Spinello et al., 2024). Similarly, studies have explored the interaction between the receptor-binding domain (RBD) of the virus and GRP78, suggesting a potential role for GRP78 in mediating viral entry (Elfiky, 2021).

It is crucial to use strong statistical techniques that can identify significant differences in interaction energy between several groups in order to methodically assess these interactions. A statistical technique called the analysis of variance (ANOVA) compares the means of three or more groups to see if the means of at least one of them deviate significantly from the others (Biology for Life, n.d.). Since ANOVA assumes identical variances among the groups being compared, it is crucial to confirm the assumption of homogeneity of variances across groups before doing the analysis. Bartlett's test is a statistical procedure used to assess the equality of variances across multiple groups, ensuring that the data meet the assumptions required for ANOVA (Six Sigma, n.d.). Additionally, we are performing in this study Levene’s Test; a statistical procedure used to assess the homogeneity of variances across different groups (Levene, 1960). It tests the null hypothesis that the variances are equal in all groups and is particularly useful when the data does not meet the assumption of normality. This test is robust to depart from normality, making it a widely preferred choice for ensuring that the assumption of equal variances is met in parametric tests such as ANOVA (Brown & Forsythe, 1974). In the context of this study, Levene’s Test was applied to evaluate whether the variance in energy measures differed significantly across SARS-CoV-2 variants, receptors, and their combinations. The results of this test offer insight into the

reliability of the subsequent statistical analyses and ensure that any observed differences in means can be attributed to true effects rather than differences in variability.

Moreover, in this study, Tukey's HSD test was performed following ANOVA to compare the energy measures between SARS-CoV-2 variants, receptors, and their combinations. Tukey's Honest Significant Difference (HSD) test is a post-hoc analysis commonly used after conducting an ANOVA to identify specific pairwise differences between group means (Tukey, 1949). This test is known for controlling the Type I error rate across multiple comparisons and is widely used when there are more than two groups to compare. Tukey's HSD provides a more detailed and specific understanding of where the significant differences lie, offering adjusted p-values (p-adj) for each pairwise comparison.

In this study, we aim to identify which energy measures can effectively differentiate between SARS-CoV-2 variants, the receptors ACE2 and GRP78, and their combinations. By applying multiple ANOVA to test for equality of means, Bartlett's test to assess the equality of variances, and Levene's Test for assessing variance homogeneitywe seek to elucidate the statistical significance of observed differences in interaction energies. This comprehensive analysis will enhance our understanding of the molecular interactions governing SARS-CoV-2 infectivity and may inform the development of targeted therapeutic interventions.

1.2 Research Problem and Goals

1.2.1 General Problem of Identification of Variants

In public health, discovering variations within the SARS-CoV-2 virus presents a problematic barrier that must be addressed. The virus, distinguished by its fast mutation rate, develops many genetic variations or variants over time. This continual development not only adds to the adaptability of the virus, but it also needs a strategy for identification that is both smart and dynamic.

Variant identification faces several challenges, one of the most significant of which is the vast variety of the viral genome. Many unique genetic sequences are produced because of the high mutation rate of SARS-CoV-2, which is constituted of ribonucleic acid (RNA), the virus's genetic material. Traditional techniques of identification, such as tests based on polymerase chain reaction (PCR), are often specifically designed for specific sequences and may have difficulty keeping up with the appearance of novel variations. As a result, there is a constraint inherent in the capability of traditional diagnostic methods to adapt to the ever-changing genetic landscape of the virus.

Furthermore, regional variations in the frequency of distinct variants are brought about by the worldwide spread of SARS-CoV-2. Tracking these variations is essential for comprehending the virus's transmission patterns and developing region-specific public health strategies since different places may experience distinct variants. Finding variations becomes a logistical and scientific problem, including efficient coordination and data exchange between research institutes and health authorities worldwide.

A further key obstacle is prompt and precise identification to provide quick interventions from the public health sector. Rapid action is required to deploy targeted treatments, such as revisions to vaccinations or modifications to public health measures, to combat the establishment of variations that have greater transmissibility or changed immune escape capabilities. There is a possibility that the efficacy of these measures might be compromised if identification is delayed, which would further add to the continued difficulties in controlling the pandemic.

1.2.2 The Goal of Research

The primary goal of this research is to identify which energy measures can effectively differentiate between SARS-CoV-2 variants, the receptors ACE2 and GRP78, and their combinations. Clarifying the molecular interactions that control viral entrance, and infectivity requires an understanding of these differences. This study aims to determine the statistical significance of observed variations by using Bartlett's test to assess the equality of variances, Levene's test for assessing variance homogeneity, and several ANOVA tests to compare the means of interaction energies. Furthermore, the research uses Tukey's HSD test to pinpoint specific pairwise differences when significant group effects are observed. The broader objective is to enhance the understanding of viral-receptor interactions, contributing to the development of more accurate methods for variant identification and tailored therapeutic interventions.

This research seeks to achieve the following specific objectives:

1. **To analyze and compare the interaction energies** between different SARS-CoV-2 variants and the ACE2 and GRP78 receptors, with a focus on identifying significant differences in energy profiles across these groups.
2. **To investigate whether combinations of variants and receptors** exhibit distinct energy profiles that could influence viral binding and entry, thus enhancing the understanding of receptor-specific interactions and their role in viral infectivity.
3. **To assess the homogeneity of variances across groups** using Bartlett's test and Levene's test. These tests ensure the validity of the ANOVA assumptions, allowing for more reliable conclusions about the differences between groups and the robustness of the variance across conditions.
4. **To identify which energy measure** (e.g., binding free energy, mean interaction energy, or standard deviation) provides the most reliable differentiation between variants, receptors, and their combinations. Additionally, Tukey's HSD test will be employed for pairwise comparison to pinpoint specific differences between the group levels when significant effects are observed.

By achieving these objectives, the study will contribute to a deeper understanding of the molecular mechanisms underlying SARS-CoV-2 infectivity and potentially inform the design of targeted therapeutic interventions. This innovative method seeks to improve the categorization system's precision and flexibility. In addition, the study tackles the challenges associated with classifying variations within the SARS family, going beyond general classifications to thoroughly examine individual variants. A more profound knowledge of the virus's infection processes is made possible by identifying and tagging specific human receptors, such as ACE2 and GRP78, and their unique families (Theerthagiri et al., 2020). This allows for more focused therapies. These efforts resulted in the creation of an extensive eight-class system that captures the complex interactions between variations and human receptors. This study stands out because of its ambitious classification strategy, which might significantly progress the area of SARS-CoV-2 variant identification. The study aims to provide a more comprehensive knowledge of the virus and its interaction with human receptors, with implications for medicinal approaches and public health initiatives.

Chapter 2

2.1 Literature Review

Attiq et al. (2022) used integrated machine-learning templates as prediction tools to investigate the anti-SARS-CoV-2 significant protease potential of FDA-approved marine medicines. The study aimed to create pharmacophore field templates for SARS-CoV-2 repurposed drugs that have FDA approval, namely Nafamostat, Hydroxyprogesterone caporate, and Camostat mesylate, to block COVID-19 main protease (Attiq et al., 2022). The second objective is to create an activity atlas model to provide structurally relevant insights into the intricate relationship between marine materials and drugs repurposed from SARS-CoV-2. Lastly, utilizing molecular dynamics simulations, the study closely investigates the kinetics of Holichondrin B's interaction with the SARS-CoV-2 main protease (Attiq et al., 2022).

Attiq et al. (2022) integrated machine-learning templates as prediction tools. The specific models applied were the pharmacophore field templates, the Activity atlas model, and the Molecular dynamics simulations. By achieving the set goals, the study hopes to improve our knowledge of and ability to identify viable treatment options for COVID-19. According to the research, the Results reveal that Holichondrin B's consistent interaction with the SARS-CoV-2 major protease makes it a suitable lead drug for further investigation and possible clinical trials. The substance interacts with essential residues involved in protease activity, indicating that it may be used as a treatment, particularly for cancer patients' COVID-19 symptoms.

That study by Qin et al. (2023) aimed to develop a rapid classification method for SARS-CoV-2 variant strains using a machine learning-based label-free Surface-Enhanced Raman Scattering (SERS) strategy. Label-free Surface-Enhanced Raman Scattering (SERS) technology and Machine Learning (ML) algorithms, specifically Logistic Regression (LR), were the primary models applied. The authors acknowledge that the timely discovery of these genetic variations and their interaction with human receptors is one of the main challenges in managing the COVID-19 pandemic effectively. Intending to bring innovative technologies into virology, the researchers fused machine learning (ML) algorithms with label-free surface-enhanced Raman scattering (SERS) technology to provide a possible rime for precise SARS-CoV-2 variant detection. The authors created a SERS spectrum database including variations of SARS-CoV-2. They showed that a diagnostic classifier using the logistic regression (LR) technique in less than ten minutes may provide precise findings (Qin et al., 2023). This technique makes it possible to identify and categorize variations in intricate biological materials. As a result, ML-based SERS technology is anticipated to distinguish between different SARS-CoV-2 variants correctly and may be used for quick diagnosis and treatment selection. The results of this source will expedite the discovery of variants and provide a more profound understanding of the complex interaction between viral alterations and human receptors.

The objective of Torun et al. (2021) is to create a Meta-surface biosensor that will enable direct detection of SARS-CoV-2 in raw saliva. The three models used include Computational screening of gold Meta surfaces, Machine learning classifiers for differentiating viral variants, and Quantitative determination of viral concentration using machine learning. Two thousand one hundred gold Meta surfaces are screened computationally to improve nanostructures and increase sensitivity. The aim is to use machine learning to detect SARS-CoV-2 from Raman spectra with high sensitivity and specificity while optimizing light-virus interaction, which is essential for

molecular-level detection. The work uses machine learning classifiers to differentiate between different viral variations, such as wild-type, alpha, and beta. The precision of the biosensor in identifying viruses and differentiating between variants is shown by validation using 36 positive and 33 negative clinical samples. With a sensitivity and specificity of 95.2%, machine learning allows for the quantitative determination of viral concentration (Torun et al., 2021). The research highlights the biosensor's capacity for large-scale screening in raw saliva, providing a quick and precise preventative measure for controlling COVID-19, including the prompt detection of novel variations such as B.1.1.7 and B.1.351. This source introduces a metasurface biosensor that has been improved using machine learning, offering a novel method for SARS-CoV-2 detection. Utilizing sophisticated algorithms and computational screening guarantees excellent sensitivity and specificity in detecting the virus while also permitting the distinction of variations, providing a valuable instrument for efficient preventative screening and prompt reaction to new mutations such as B.1.1.7 and B.1.351.

Rehman et al. (2023) have systematically reviewed identification and diagnosis techniques for COVID-19 using novel technologies. Their work intended to provide an update on recent developments, higher-level uses, limitations, and potential future research directions in this area. The authors also presented different machine learning and artificial intelligence techniques used in COVID-19 detection, such as the deep learning algorithm, CNNs, and SVMs. They spoke of its application in interpreting data from testing, including chest X-rays and CT scans, efficiently identifying illnesses. It also covered the use of machine learning in combination with other technologies, such as biosensors and smartphone-based systems for point-of-care applications. Based on this study, Rehman et al. established that machine learning-based techniques offered a much higher rate of COVID-19 identification coupled with much faster results than conventional processes. However, they also reported several limitations, such as the requirement for big and multiplexed data and the feature of model interpretability in clinical practice.

Chen et al. (2021) extensively reviewed AI interventions in the fight against COVID-19. Specifically, their objectives involved understanding how AI is used in the battle against COVID-19, such as identifying and diagnosing the infection and developing treatments such as drugs and vaccines. The study investigated several machine learning models, including deep learning networks, reinforcement learning, and natural language processing algorithms. Some of the aspects analyzed by the experts included the application of AI to medical images, spread modeling, and protein structures. The authors also noted that different AI approaches proved to be very effective in screening COVID-19 from pulmonary X-rays and CT scans, and the measures of accuracy explored included over 90 percent. They also pointed to the capability of using machine learning to collect data concerning virus-host interactions and to predict potential drug candidates. Nevertheless, Chen et al. call for more robust and applicable models to a broader population since the virus and its mutations change over time.

Lee and Chen, 2021 addressed the application of deep learning to identify new drugs to treat COVID-19. The researchers use sophisticated machine learning technologies to unveil potential therapies. The researchers used GNNs and other transformer-based models to explain large, complex molecular and clinical datasets. A significant component of their strategy concerned the multimodal analysis of protein sequences, three-dimensional structures, and gene expression patterns. The study's integral conclusions involved identifying several new drugs with possible repurposing against SARS-CoV-2. Moreover, Lee and Chen (2021) identified the possibility of using deep learning methods to determine potential therapeutic agents quickly, which is especially valuable for responding to emerging virus variants. The models' accuracy in

predicting drug-target interactions and possible side effects gave them essential data to guide candidates for clinical trials.

Titus et al. (2022) also discussed the various electrochemical biosensors' architectures for identifying SARS-CoV-2. Although the microwear biosensors were the primary emphasis of the paper, they delved into the application of machine learning algorithms for increasing the sensitivities and accuracy of the detection systems. This paper reviewed different biosensor formats focusing on aptamers, antibodies, and molecularly imprinted polymers. Such intelligent biosensors as artificial neural networks and support vector machines were employed for signal analysis of these biosensors. The authors also concluded that enhancing the biosensors with enhanced designs synchronized with the workings of machine learning algorithms significantly improved the speed and efficiency of testing for SARS-CoV-2. The sensitivity of some of the reviewed systems showed that viral particles could be detected at a femtomolar level. Titus et al. (2022) pointed out the need for these integrated approaches to offer a fast diagnostic solution on the patient's side where various SARS-CoV-2 variants can be identified.

In their study, Ribes-Zamora and Simmons (2022) provided a novel teaching strategy for undergraduates majoring in genetics where students analyze the COVID-19 virus using bioinformatics approaches online. Although their studies are not centered on the detection method, understanding their research helps appreciate how Willig and his team develop machine learning algorithms to identify the SARS-CoV-2 variant. The following bioinformatics tools and databases were employed in the study: Basic Local Alignment Search Tool (BLAST), Clustal Omega, and Virus Nucleotide Collection in NCBI. Students used these tools to sequence SARS-CoV-2, find mutations, and characterize what these may do to the protein. The authors also noted that such an approach was beneficial for increasing students' knowledge of viral genetics and discerning more about the development of the SARS-CoV-2 variants. This study showed that integrating bioinformatics with educational 3D printing could help visualize the viral protein structures in future investigations concerning variant identification and description.

In a recent study by Hazari and Pal Chaudhuri (2022), the coronavirus envelope protein was modeled via cellular automata. It is worth mentioning that they aimed to create a new computational method for analyzing the movement of viral proteins and possible interactions with host cells. In the study of the model, the cellular automata model was applied to explore the envelope protein under different situations. They also provided for the evaluation of protein folding and its conformation changes in time, among the key features of their approach (Hazari & Chaudhuri, 2022). The findings of their work enriched the knowledge about structural changes of the SARS-CoV-2 envelope protein, which could help investigate virus-host interactions further and design appropriate therapies. Nevertheless, this research is not explicitly devoted to machine learning for variant identification; however, it manifests a methodological direction to understand viral proteins by computational modeling, which can be helpful in future machine learning to detect and characterize new viral variants.

Rampogu et al. (2021) studied marine drugs as therapeutic agents against SARS-CoV-2 by employing the essence dynamics and analysis of the free energy landscape. The authors planned to recognize prospective drug molecules from marine origin capable of interacting with the proteins of SARS-CoV-2. The researchers studied the interactions of marine compounds with viral proteins through molecular dynamics simulations and machine learning. Some of the activities they performed were applying principal component analysis (PCA) and free energy calculations of the complexes formed between the drug and the protein. In this paper, the researchers identified several components from marine sources with good anti-SARS-CoV-2 inhibitory activity. The

study by Rampogu et al. showed the importance of integrating computational approaches with machine learning solutions for the drug design against SARS-CoV-2 and its spawns.

Similarly, Vangipuram and Appusamy (2021) proposed a machine learning-based COVID-19 diagnosis system. Using several machine learning algorithms, their study sought to develop a correct and swift diagnostic tool. The authors used decision trees, random forests, and support vector machine analyses on the clinical and demographic datasets. Some of the highlights of their methods entailed the feature selection methods that would help to determine the most critical predictors of COVID-19 infection. This yielded a diagnostic model with impressive accuracy, sensitivity, and specificity. The study affirms that machine learning can help create fast and practical diagnostic tests for COVID-19 that can be modified to identify other virus versions.

Parvathy et al. (2023) reported on a machine-learning approach for predicting COVID in a different study. In their work, they sought to create models for the early detection of potential patients with long-term complications caused by COVID-19. The authors utilized logistic regression, random forest, and neural network models to perform feature analysis on patients' heterogeneous information. One of the significant aspects of their strategies was data merging, where patients' clinical, laboratory, and demographic records were developed into detailed profiles. It was possible to create models to estimate the potential development of long COVID based on the outcomes found in the study. In their research, Parvathy et al. showed how machine learning could predict the pathophysiology of SARS-CoV-2 infection, which would be helpful, especially when dealing with various virus strains.

Gantini and Christian (2022) proposed using data mining techniques to analyze the 3D protein models of the SARS-CoV-2 virus. Their study explored primary structural properties and searched for common overall structural themes in viral proteins that could be utilized in designing effective vaccines and drugs. The researchers employed a form of data mining known as clustering and association rule mining on the protein's sequence and structure data. Some of the significant elements of their strategy embraced were the recognition of such conserved regions in different strains of the virus that may also hold the epitopes for the targets that can be vaccinated. The study also discovered several promising protein regions for subsequent analysis. Gantini and Christian's work demonstrated how data mining and machine learning approaches can be applied at the molecular level concerning SARS-CoV-2, which might help discover effective treatments and vaccines for various stemming from various mutations.

Table 1. Summary of Literature Review

Study	Goal	models	Features	Main results & conclusion
Attiq et al. (2022)	Identify the anti-SARS-CoV-2 potential of marine drugs	-Pharmacophore field templates -Activity atlas model -Molecular dynamics simulations	FDA-approved marine drugs, SARS-CoV-2 main protease	Holichondrin B was identified as a potential drug candidate and insights into drug-protein interactions.
Qin et al. (2023)	Rapid classification of SARS-CoV-2 variants	Label-free Surface-Enhanced Raman Scattering (SERS) technology -Machine Learning (ML) algorithms, specifically Logistic Regression (LR)	SERS spectra of SARS-CoV-2 variants	**Results:** The study achieved **100% accuracy** in classifying SARS-CoV-2 variants (Beta, Delta, Wuhan, and Omicron) using a machine learning-based label-free SERS strategy. Additionally, the logistic regression (LR) model demonstrated **100% accuracy** in blind tests on human nasal swab samples, effectively distinguishing between positive and negative samples. **Conclusion**: The machine learning-based SERS strategy offers a highly accurate and rapid method for classifying SARS-CoV-2 variants and diagnosing infections, demonstrating its potential as a diagnostic tool for viral strain identification.
Torun et al. (2021)	Direct detection of SARS-CoV-2 in raw saliva	-Computational screening of gold Meta surfaces -Machine learning classifiers for differentiating viral variants	Gold meta-surfaces, Raman spectra	**Results**: The study utilized machine learning to analyze Raman spectra obtained from DNA aptamer meta surfaces, achieving a sensitivity and specificity of 95.2% in distinguishing SARS-CoV-2 from negative samples. Additionally, the method effectively differentiated between wild-

				type, Alpha, and Beta variants of the virus. **Conclusion**: the integration of machine learning with DNA aptamer meta surfaces offers a rapid and accurate approach for detecting SARS-CoV-2 and its variants, highlighting its potential for effective viral surveillance and diagnostics.
Rehman et al. (2023)	Review COVID-19 diagnosis techniques	CNNs, SVMs	Chest X-ray, CT scans, biosensors	High accuracy, faster results, requires large datasets
Chen et al. (2021)	Review AI applications in COVID-19	Deep learning, reinforcement learning, NLP	Medical images, spread modeling, protein structures	**Results**: The study surveys the diverse applications of artificial intelligence (AI) in the fight against COVID-19, covering AI-based diagnostic methods, drug discovery, and epidemic prediction models. **Conclusion**: AI technologies have proven effective in improving diagnostic accuracy (90%), supporting drug discovery, and providing insights into COVID-19 spread and management, highlighting their crucial role in pandemic response strategies.
Lee and Chen (2021)	Identify new drugs for COVID-19	GNNs, transformer-based	Protein sequences, 3D structures, gene expression	Identified potential drug candidates, rapid drug discovery. In the study by Lee and Chen (2021), the deep learning model achieved **an accuracy of 90.6%** in predicting the effectiveness of existing drugs for repurposing against COVID-19.

				This accuracy highlights the model's effectiveness in identifying promising drug candidates based on their molecular characteristics and potential efficacy against SARS-CoV-2.
Titus et al. (2022)	Develop electrochemical biosensors for COVID-19	ANNs, SVMs	Aptamers, antibodies, molecularly imprinted polymers	**Main Results**: The study reviews various electrochemical biosensor designs used for detecting the SARS-CoV-2 virus. It discusses the different sensor types, including amperometry, potentiometric, and conductometric sensors, highlighting their sensitivity and application in detecting COVID-19. **Conclusion**: Electrochemical biosensors provide an effective, rapid, and cost-efficient approach for COVID-19 detection, offering significant advantages over traditional diagnostic methods, such as PCR tests, particularly in point-of-care settings.
Ribes-Zamora and Simmons (2022)	Teach bioinformatics for COVID-19 analysis	BLAST, Clustal Omega	SARS-CoV-2 sequences	Improved student understanding, potential for variant identification
Hazari and Pal Chaudhuri (2022)	Model coronavirus envelope protein	Cellular automata	Protein structure	Understanding of protein dynamics, potential for drug design
Rampogu et al. (2021)	Identify marine drugs against COVID-19	PCA, free energy calculations	Marine compounds,	Identified potential drug candidates, integration of computational and ML

			protein interactions	
Vangipuram and Appusamy (2021)	Develop COVID-19 diagnosis system	Decision trees, random forests, SVM	Clinical and demographic data	Vangipuram and Appusamy developed CovFilter, a machine learning-based health monitoring framework for early COVID-19 diagnosis. They trained supervised learning algorithms on vital sign data, achieving an F1 score of **93.22%**with the Multilayer Perceptron model. Notably, a weighted majority voting ensemble classifier outperformed individual models, reaching an F1 score of **94.5%**. These findings suggest that CovFilter could serve as an effective wearable device for at-home COVID-19 prediction.
Parvathy et al. (2023)	Predict long-term COVID complications	Logistic regression, random forest, neural networks	Clinical, laboratory, and demographic data	Parvathy et al. employed machine learning techniques to analyze patient data for predicting the prognosis of long COVID.By examining various clinical and demographic features, they developed predictive models aimed at identifying patients at higher risk for prolonged symptoms. The study underscores the potential of data-driven approaches to inform healthcare strategies for managing long COVID.
Gantini and Christian (2022)	Analyze SARS-CoV-2 protein structures	Clustering, association rule mining	Protein sequences, structures	Identified conserved regions, potential vaccines, and drug targets

2.2 Thesis Goals and Contributions

2.2.1 Critique of the Literature

The current literature on identifying and classifying SARS-CoV-2 variants using machine-learning approaches is rich and multifaceted due to the high demand for knowledge on virus evolution due to the ongoing pandemic. The researchers have vigorously worked to employ different types of machine learning models for understanding viral genetic sequences, anticipating proteins' structures, and finding potential drug targets. Nevertheless, there are some significant gaps and drawbacks of the studies available today, which should be addressed in future research.

One major gap in the literature is the lack of coordination between energy measures and sequence data in the classification of SARS-CoV-2 variants. Numerous studies have examined genetic sequence analysis, with many articles describing the impact of mutations on viral evolution (Jackson et al., 2022). However, there is a relative scarcity of research incorporating energy measures alongside genetic sequence data to provide a more comprehensive view of virus-host interactions. Energy-based metrics can measure the functional effects of these changes by evaluating how they alter the viral binding affinity to receptors such as ACE2 and GRP78, even though genomic data identifies which mutations are present. While genetic data reveals which mutations are present, energy-based metrics can quantify the functional consequences of these mutations by assessing how they affect viral binding affinity to receptors like ACE2 and GRP78. By combining sequence and energy data, classification accuracy could be improved, as a mutation's impact on infectivity is not solely dictated by its presence in the genome but also by the energetic stability of the resulting protein structure. Our capacity to describe viral variations and forecast their likelihood of increased transmission or immune evasion would be improved by combining the two data types.

Despite the value that energy measures can provide, the literature has yet to systematically address key questions: (a) how energy measures can differentiate between SARS-CoV-2 variants, (b) how energy measures distinguish between different host receptors, and (c) how energy measures can characterize the interactions between variants and receptors. The assumption that all mutations contribute equally to viral activity is made when variants are categorized solely on the basis of sequence, while in practice, their impacts rely on structural and energetic repercussions. In a similar vein, although ACE2 is generally acknowledged as the main receptor for SARS-CoV-2, under certain circumstances, alternative receptors, such as GRP78, may promote viral entry (Ibrahim et al., 2020). More insight into alternate viral entry mechanisms would come from a more thorough examination of the differences in energy measurements across various receptor contacts. Additionally, it may be possible to determine whether specific mutations preferentially boost binding to one receptor over another by examining the energy differences between variant-receptor combinations. This could have implications for tissue tropism and the severity of disease.

From a biological perspective, energy measures should be considered potential biomarkers for characterizing viral variants and receptors due to their direct influence on viral infectivity and host adaptation. One important factor in determining transmissibility is the viral spike protein's affinity for binding to its receptor; changes that boost infectivity while decreasing binding energy likely to impair viral entry (Nguyen et al., 2021). Unlike sequence data alone, which provides a static view of mutations, energy measures capture the dynamic biophysical effects that drive viral evolution. A deeper functional knowledge of how mutations impact viral behavior is made possible by examining how energy measures distinguish interactions between variations and receptors.

Certain mutations in the receptor-binding domain (RBD), for instance, may not substantially change the genetic sequence but may cause a large shift in the interaction energy, which could result in either an increase or decrease in binding affinity. Understanding these energetic changes is crucial for assessing the emergence of new variants with altered pathogenicity or resistance to neutralizing antibodies.

In conclusion, whereas machine learning and sequence-based techniques have greatly advanced the classification of SARS-CoV-2 variants, the literature has mostly ignored the function of energy measurements in distinguishing variants, receptors, and their interactions. In addition to increasing classification accuracy, filling in these gaps would advance our biology knowledge of viral-host interactions and possibly direct the creation of focused antiviral tactics.

2.2.2 Addressing the Gaps

Understanding the interactions between SARS-CoV-2 variants and their host receptors is essential for predicting viral infectivity and informing therapeutic strategies. A crucial gap in the statistical characterization of energy-based interaction effects between variations and receptors still exists, even though a great deal of research has concentrated on using structural modeling and genetic mutations to describe these interactions. Most studies have emphasized sequence mutations, receptor-binding domain (RBD) structural changes (Starr et al., 2020), or machine learning-driven classification models (Maher et al., 2022), but few have rigorously assessed whether energy measures can serve as statistically significant biomarkers for viral-host binding. This thesis addresses that gap by investigating whether energy-based metrics provide a robust and interpretable means of distinguishing SARS-CoV-2 variants, receptors, and their combinations.

From a biological standpoint, existing literature has primarily focused on how amino acid substitutions in the spike protein affect binding affinity to ACE2 and other potential receptors (Barton et al., 2021). However, instead of performing a statistical study of energy distributions among variations and receptors, these investigations frequently depend on molecular docking or free energy perturbation simulations. While it is well established that certain spike mutations enhance binding affinity (Li et al., 2020), this thesis extends previous work by systematically testing whether energy variations are statistically significant across different variants and receptor interactions. Beyond qualitative structural insights, this study empirically validates whether energy metrics may distinguish viral interactions in a meaningful way using ANOVA, Bartlett's test, and Levene's test.

A second gap exists in the field of computational virology, where energy measures are often used to estimate binding affinities but rarely subjected to rigorous statistical testing. Few research has investigated whether energy variances across many variants and receptors follow separate statistical patterns, as many have relied on single-case energy calculations to compare SARS-CoV-2 variants (Gobeil et al., 2021). By using hypothesis-driven statistical testing to ascertain whether energy and mean energy significantly distinguish variations, receptors, and their combinations, this thesis overcomes this constraint. By doing so, it ensures that conclusions about binding interactions are not based on isolated case studies but rather on statistically supported findings.

A third gap lies in the methodological approach of variant classification. The majority of classification models frequently mix machine-learning approaches with genomic and protein sequence data. Although these models are capable of accurately classifying variants, they usually lack interpretability and a biological foundation in interactions based on energy. This research

provides an alternative approach by evaluating whether energy measures alone can offer statistically significant differentiation without the need for complex classification algorithms. This contribution is particularly important because it allows researchers to assess variant-receptor interactions based on fundamental thermodynamic properties rather than relying solely on predictive computational models.

By filling in these gaps, this thesis advances a more statistically sound comprehension of the interactions between SARS-CoV-2 variants and offers a framework for further research aiming to incorporate energy-based characterization into viral epidemiology and treatment planning. This study highlights the importance of statistical validation in energy calculations and establishes a foundation for future investigations into how energy measures can serve as reliable biomarkers for viral evolution and host susceptibility.

2.2.3 Thesis Goals

The primary goal of this thesis is to identify which energy measurements can be used as trustworthy biomarkers to differentiate between host receptors (ACE2 and GRP78), SARS-CoV-2 variants, and their combinations. This study advances the statistical analysis of viral-host interactions by thoroughly examining the variations in energy-based metrics using well-established statistical techniques, such as ANOVA for mean comparisons, Bartlett's test for variance homogeneity, and Levene's test to assess the robustness of variance assumptions. Unlike previous research that often relies on machine learning classification, this study focuses on a purely statistical approach to ensure interpretability and reproducibility in the findings.

This thesis makes three key contributions:

1. **Which energy measures can significantly differentiate between SARS-CoV-2 variants, receptors, and their combinations?**
   By applying multiple ANOVA tests, this study identifies which energy metrics—interaction energy, mean energy, or the standard deviation of energy—show statistically significant differences across viral variants, receptor types, and their combined interactions. To further investigate significant pairwise differences, Tukey's HSD test is applied. The results offer a quantitative foundation for evaluating binding differences and advance our knowledge of how molecular interactions differ among SARS-CoV-2 strains (Ibrahim et al., 2020; Spinello et al., 2024).
2. **Do energy variances provide meaningful differentiation across groups, and are they statistically significant?**
   The study investigates whether there are significant differences in energy variances, mean energy, and standard deviation of energy among variations, receptors, and their combinations using Bartlett's test and Levene's test. Identifying whether energy variance can serve as a distinguishing feature enhances our understanding of how stable or variable these interactions are, and whether certain energy fluctuations indicate stronger or weaker viral binding affinities (Elfiky, 2021).
3. **Can a statistical approach provide a robust and interpretable method for assessing SARS-CoV-2 variant interactions?**
   This study ensures that the results are not only statistically valid but also simple for academics and physicians to understand by avoiding machine learning categorization and concentrating on strong statistical testing. The results are more reliable when hypothesis-driven statistical techniques are used, which makes them suitable for additional

experimental validation and practical biological interpretation (Biology for Life, n.d.; Six Sigma, n.d.).

These contributions advance the statistical analysis of SARS-CoV-2 interactions by identifying which energy measures provide meaningful differentiation and whether variance in these measures plays a role in distinguishing viral-host binding behaviors. The incorporation of Levene's test ensures that the variance assumptions are robust, and the application of Tukey's HSD test enhances the interpretation of pairwise differences. This work lays the foundation for future studies that aim to integrate statistical findings with experimental validation, ultimately contributing to a deeper understanding of viral infectivity and therapeutic target identification.

2.2.4 Thesis Contributions

This thesis offers several significant contributions to the field of computational virology and statistical bioinformatics by introducing an energy-based perspective in the differentiation of SARS-CoV-2 variants, receptors, and their combinations. Most categorization efforts in the present literature primarily rely on genomic sequence data and machine learning models, frequently ignoring viral-host contact's thermodynamic and molecular interaction elements. These contributions are meant to close this significant gap.

1- This research introduces the integration of energy measures as an additional feature, alongside traditional sequence data, to aid in the identification of SARS-CoV-2 variants and their receptor interactions. This thesis suggests that the physical and chemical binding energy information adds an extra layer of biological relevance, even though the majority of previous research focuses on classifying variations based only on sequence changes (Kumar et al., 2022). Incorporating these energy metrics can potentially enhance differentiation and provide a more stable and biologically meaningful basis for assessing variant behavior.

2- This study identifies which specific energy measures show statistically significant differences across SARS-CoV-2 variants. It has been shown using ANOVA, Bartlett, and Levene's tests that there are significant differences in interaction energy between various viral strains. According to Nguyen et al. (2021), this discovery advances our understanding of how structural modifications to the viral spike protein affect binding energetics and, consequently, infectivity.

3- This research determines which energy measures are statistically different across host receptors. Mean energy, in particular, is a crucial distinction between the ACE2 and GRP78 receptors, according to the study, which offers a thermodynamic explanation for differences in receptor vulnerability. Previous studies have focused on the structural compatibility of the receptors (Yan et al., 2020), but this thesis extends that knowledge by statistically validating the energetic differences in their interactions.

4- The study reveals which energy measures show significant differences across the combined variant-receptor interactions. This thorough method highlights the combined impact of viral evolution and receptor diversity on binding behavior by enabling the identification of energy parameters that are sensitive to both the host receptor and the viral variation. Such insights have been largely absent from previous computational studies, which typically analyze variant-receptor interactions in isolation (Shoemark et al., 2021).

All of these contributions highlight the value of thermodynamic data in enhancing sequence-based methods and provide a new statistical viewpoint on SARS-CoV-2 variant identification and receptor interaction analysis. In addition to improving our knowledge of viral-host binding mechanisms, the results of this thesis could help guide future treatment approaches that focus on receptor interactions unique to variants.

Chapter 3

3.1 Methodology

The objective of this study is to identify which energy measure can effectively differentiate between SARS-CoV-2 variants, receptors, and their combinations. The study employs several statistical hypotheses testing techniques, including Multiple ANOVA tests, Bartlett's tests for equal variances, and Levene's test for equality of variances as indicated in Figure 1. These tests are designed to assess differences in energy measures across categorical groups (variants, receptors, and combinations). Energy measurements for various biological structures are categorized by receptors, variants, and combinations of both.

Energy Measures:

1. Energy (S1): The Binding Energy

- Energy (s1) represents the binding energy between SARS-CoV-2 variants and their target receptors. This energy is calculated using molecular docking simulations, specifically using SwarmDock.
- Binding energy measures the strength of the interaction between two molecules (e.g., the viral spike protein and the ACE2 receptor). A lower (more negative) binding energy indicates a stronger interaction, meaning the variant binds more tightly to the receptor, while a higher (less negative or positive) binding energy indicates a weaker or less favorable interaction.
- SwarmDock simulates these interactions and provides binding energy as part of its output, which helps researchers assess the likelihood and strength of interactions between the virus and various receptors.

2. Mean Energy (S2): The Average Binding Energy Across Each Group

- Mean Energy (s2) is the average binding energy for a specific group of variants, receptors, or combinations of both.
- In the analysis, the mean energy of multiple SARS-CoV-2 variants (or variants*receptor combinations) interacting with a specific receptor can be calculated. This means that we are taking the binding energy values (s1) from all individual interactions and calculating their average.
- This value gives an overall representation of how strongly the variants, on average, interact with the receptor in a given group. For example, if a group of variants has a high mean energy, it suggests that, on average, these variants interact weakly with the receptor. Conversely, a low mean energy suggests stronger, more efficient interactions.

3. Standard Deviation of Energy (S3): Variability in Binding Energy

- Standard Deviation of Energy (s3) is a statistical measure that quantifies how much the individual binding energies (s1) deviate from the mean energy (s2) within a given group. It tells how consistent or variable the interactions are across the group.
- A low standard deviation means that the binding energies for the variants or combinations in that group are relatively similar to each other, suggesting that the interaction strength is consistent.

- A high standard deviation, on the other hand, indicates that the binding energies vary greatly, meaning that there is a wide range of interaction strengths within the group. This could suggest that certain variants have a much stronger or weaker affinity for the receptor than others in the same group.

Why These Measures Matter:

These three energy measures are crucial for understanding:

- Energy (S1): This shows us the strength of the interaction between SARS-CoV-2 variants and receptors.
- Mean Energy (S2): Provides a general overview of how the group of variants or combinations interacts with the receptor on average, helping to identify which variants or receptor groups are more likely to bind strongly.
- Standard Deviation of Energy (S3): Reveals the variability in the interaction strength, which can indicate the reliability or consistency of those interactions across different variants or receptors. High variability could suggest that some variants interact more strongly, while others do so weakly.

These energy measures, calculated through SwarmDock simulations, are then analyzed statistically to assess whether the differences in energy between variants, receptors, or combinations are significant.

To evaluate differences in the mean and variance of these energy measures, the following statistical tests were performed:

1. Multiple ANOVA Tests (for Mean Differences):

ANOVA tests assess whether the means of energy measures differ significantly between groups. The ANOVA tests were applied in the following categories and explained in Figure 2:

1. **ANOVA across Variants:** Determines if there are significant differences in mean energy (s2) between different SARS-CoV-2 variants.
2. **ANOVA across Receptors:** Assesses significant mean differences in energy (s2) among different receptors.
3. **ANOVA across Variants*Receptors Combinations:** Evaluates if the interaction between variants and receptors leads to significant differences in mean energy measures (s2).

For each ANOVA test:

1. The null hypothesis (H_0) assumes that the mean energy measures (s2) are equal across all groups.
2. The p-value and F-statistic are calculated.
3. If the p-value is less than the significance level (e.g., 0.05), H_0 is rejected, indicating significant differences in mean energy measures across groups.

2. Bartlett's Test for Equal Variances:

Bartlett's test was used to assess if the variances in energy measures are equal across the groups. This test checks for the homogeneity of variances across categories. Bartlett's test was conducted in the following categories and explained in Figure 3:

1. **Bartlett's Test across Variants:** Assesses if the variance in energy (s1) differs between SARS-CoV-2 variants.
2. **Bartlett's Test across Receptors:** Tests for significant variance differences in energy (s1) among receptors.
3. **Bartlett's Test across Variants*Receptors Combinations:** Evaluates if variance in energy (s1) differs for the interaction between variants and receptors.

For each Bartlett's test:

1. The null hypothesis (H_0) assumes that all groups have equal variances.
2. The Bartlett's test statistic and p-value are calculated.
3. If the p-value is less than the significance level (e.g., 0.05), H_0 is rejected, suggesting significant variance differences.

3. Levene's Test for Equality of Variances:

Levene's test was performed alongside Bartlett's test to assess the homogeneity of variances, especially when normality assumptions may be violated. Levene's test is more robust than Bartlett's when the data is not normally distributed. It checks whether the variances across groups are equal. Levene's test was conducted for the same categories as Bartlett's test and explained in Figure 4:

1. **Levene's Test across Variants:** Assesses variance differences in energy measures (s1) between SARS-CoV-2 variants.
2. **Levene's Test across Receptors:** Tests if there are significant differences in the variance of energy (s1) among receptors.
3. **Levene's Test across Variants*Receptors Combinations:** Evaluates variance differences in energy (s1) for the interaction between variants and receptors.

For each Levene's test:

1. The null hypothesis (H_0) assumes that the variances are equal across groups.
2. The Levene's test statistic and p-value are computed.
3. If the p-value is less than the significance level (e.g., 0.05), H_0 is rejected, suggesting significant variance differences.

4. Post-hoc Tukey's HSD Test:

When ANOVA reveals significant differences among groups, Tukey's Honest Significant Difference (HSD) test is used to determine which specific pairs of groups differ. This test is ideal for multiple pairwise comparisons as it controls the family-wise error rate and identifies which group means differ significantly from each other.

Mean of the Mean and Variance of the Variance:

To summarize and compare the mean energy (s2) and variance of energy (s1) across the groups, we compute the mean of the mean and variance of the variance:

- The mean of the mean refers to the average of the mean energy values (s2) for all groups (variants, receptors, combinations).
- The variance of the variance reflects the variation in the variance (s3) across groups. These calculations are essential for performing statistical tests like ANOVA and Bartlett's/Levene's tests, as they provide a clearer picture of how the energy measures differ across groups.

Software and Libraries:

All statistical analyses, including the ANOVA, Bartlett's test, and Levene's test, were conducted using Python. The following libraries were used:

- Pandas: For data manipulation and organization.
- Scipy: For statistical testing (ANOVA, Bartlett's, and Levene's tests).
- Matplotlib: For data visualization and plotting results.

This methodology enables a thorough assessment of how energy measures (mean energy and variance) differ across SARS-CoV-2 variants, receptors, and their combinations, providing insights into the underlying interactions.

Figure 1. FLOWCHART OF STATISTICAL ANALYSES

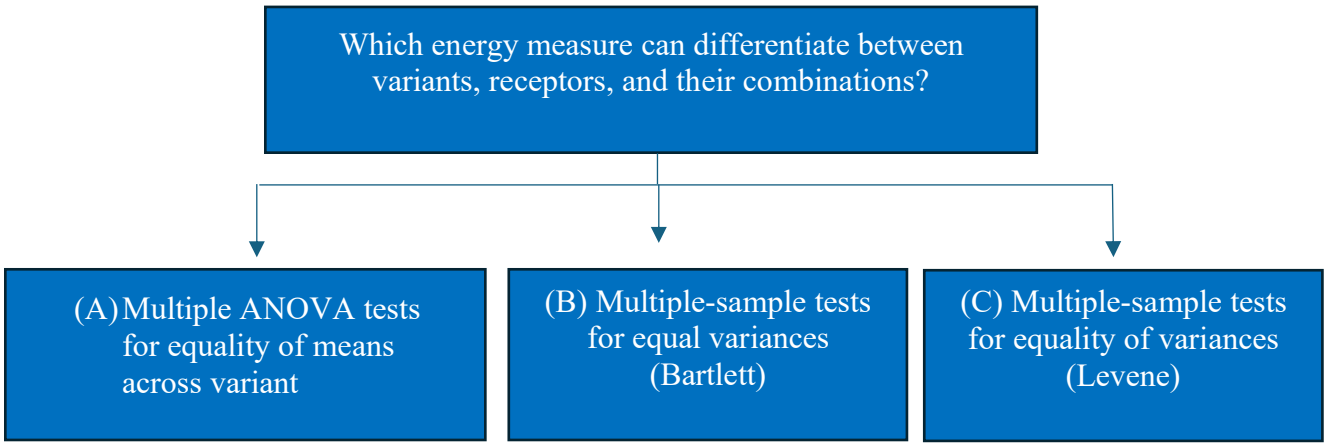


Figure 2. Multiple ANOVA Tests for Equality of Means Across Variants

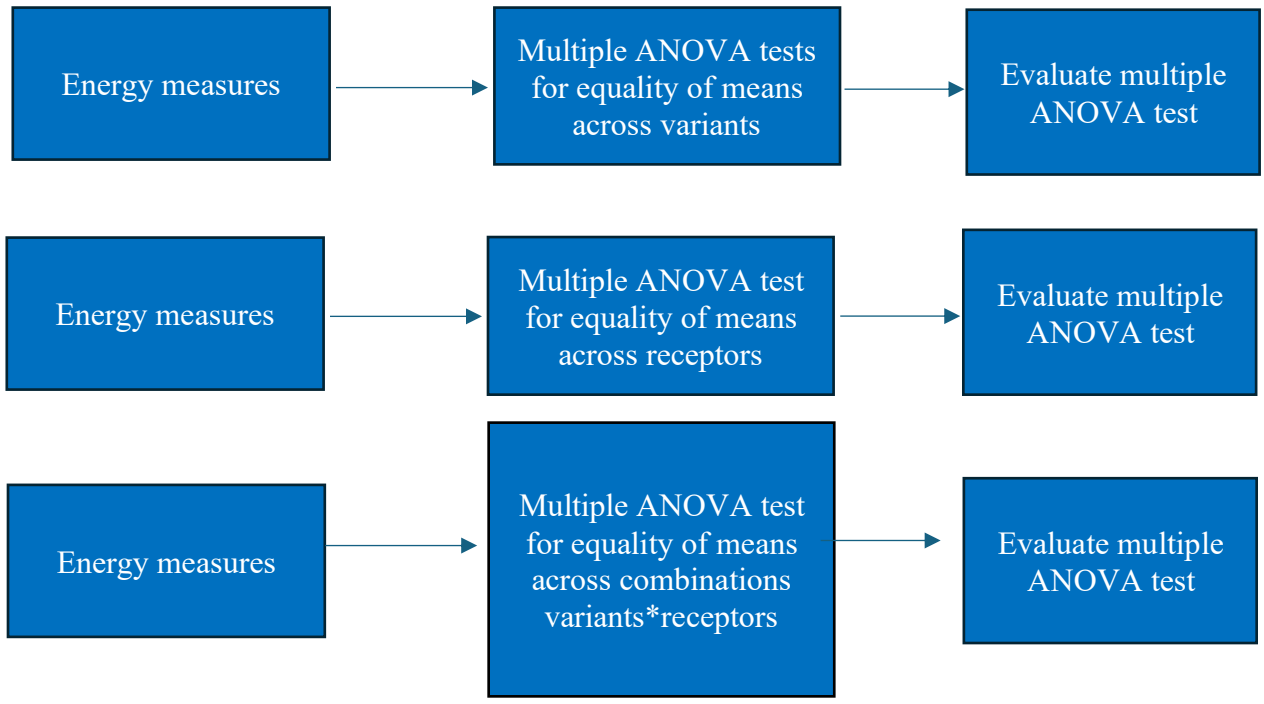


Figure 3. Multiple-sample Tests for Equal Variances (Bartlett)

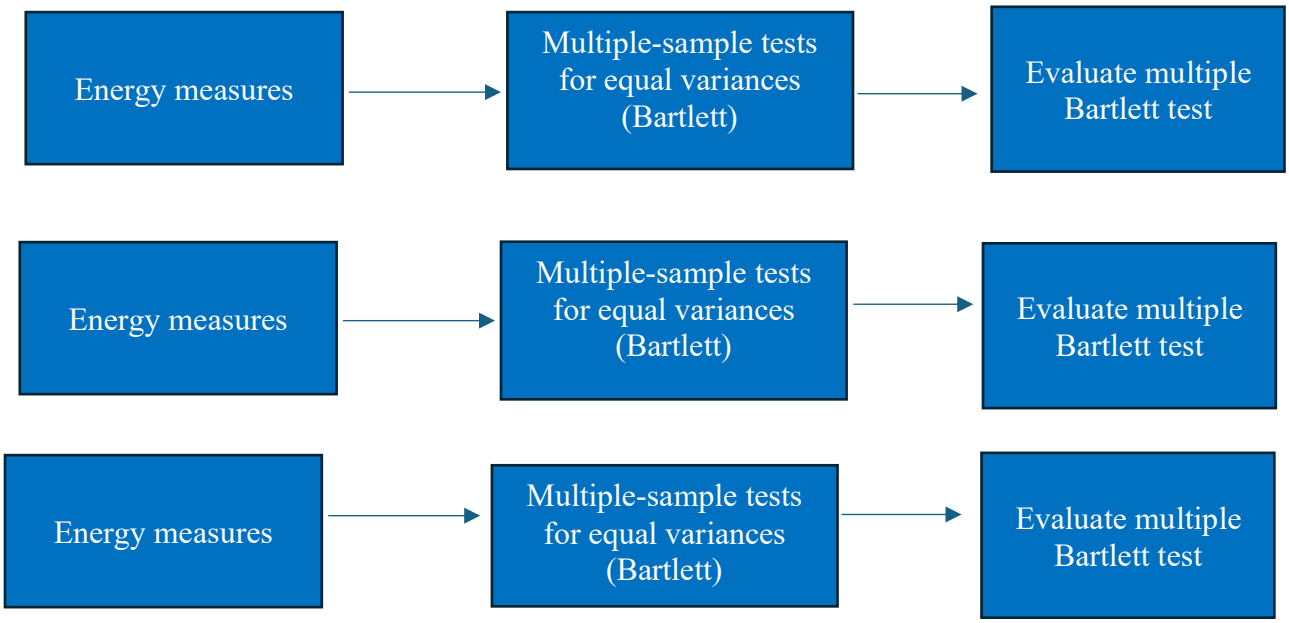
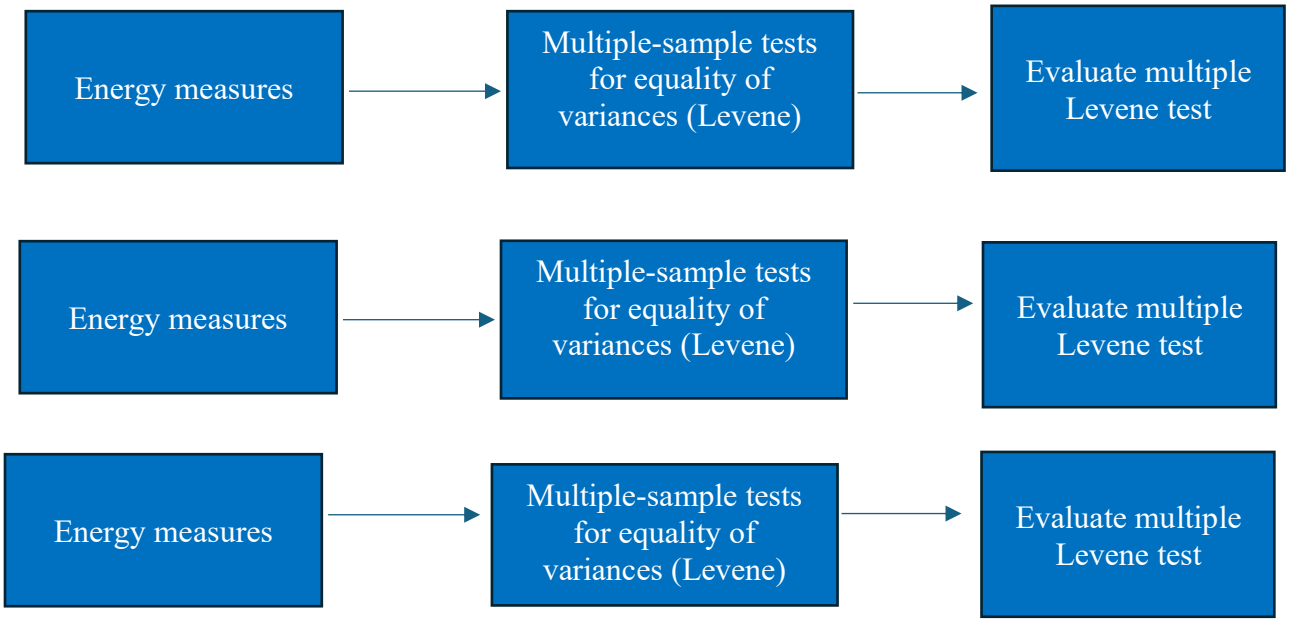


Figure 4. Multiple-sample Tests for Equality of Variances (Levene)



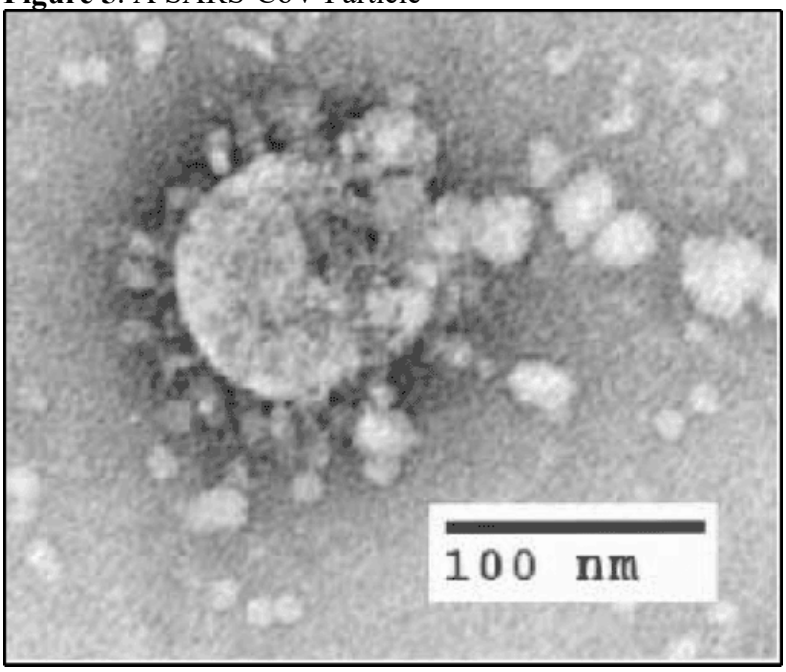
Chapter 4

4.1 Data Collection and Experimental Results

4.1.1 Data collection

The dataset would comprise the RNA sequences of SARS-CoV-2 variants of concern, such as the Alpha, Beta, Gamma, Delta, Omicron, and others, and the human receptor sequences of ACE2 and GRP78. These sequences would be retrieved from authentic genomic databases such as GISAID or GenBank. Conventional molecular biology bioinformatics software will further translate the RNA sequences into protein sequences. Figure 5 displays an example of a SARS-CoV particle.

Figure 5. A SARS-CoV Particle



From: Centers for Disease Control and Prevention (CDC, 2024).

In this experiment, the RNA sequence for each SARS-COV-2 variant for ACE2, GRP78, and wildtype RBD domain was downloaded and converted to protein sequence. The RBD of the original SARS-Cov-2, the GRP78 region, and the common ACE2 allele were sequenced first as wildtype (WT). However, a specific region for each was used in this experiment. The sequence from 334 to 530 in the wild-type RBD region was chosen based on its mutations within the ACE2 binding site that have appeared in many lineages/clades of SARS-CoV-2. For example, the N501Y mutation has appeared in Beta (B.1.351; 20 H/501Y.V2), Gamma (P.1;20 J/501Y.V3), Alpha (B.1.1.7; 20I/501Y.V1), and Omicron (B.1.1.529; 21M/501Y). So, the mutations of SARS-CoV-2 were found in this specific region of the RBD wild-type domain. The ACE2 region in this

experiment is the region where it interfaces with SARS-COV-2 RBD, in which 16 residues of SARS-COV-2 RBD were shown to be in contact with 20 residues of ACE2 (Lan et al., 2020). However, for Grp78 (also known as HSPA5), the region where Grp78 and SARS-COV-2 RBD interact was sequenced and expressed.

SARS-COV-2 variants were sequenced based on their mutations in the RBD domain. The Alpha variant has one mutation located on N501Y, and Zeta/Lota/Eta also has one on E484K; thus, they have the same sequence. Delta has two mutations located at L452R and T478K. Beta has three mutations located on K417N, E484K, and N501Y. Gamma has three mutations on K417T, E484K, and N501Y. However, Omicron has fifteen mutations located at G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, and Y505H.

Homology modeling is then used to generate a 3D protein model from the target sequence and evolutionary-related protein structures. Practical information will be gathered and used as a template. The first step in the SWISS-MODEL was the input of data: an amino acid sequence in FASTA text was input in the SWISS-MODEL of the target protein. UniProtKB was used to help build the FASTA text (UniProt, 2016). The target proteins in our experiment are the SARS-COV-2 variants (Omicron, Delta, Gamma, Eta, Zeta, Lota, Beta, Alpha), RBD domain, ACE2, and Grp78. The second step was to search the template using SWISS-MODEL. Two database search methods were used to perform this task: HHblits (Remmert et al., 2011) and Blast (Altschul, 1997). In the case of remote homology, HHblits would add sensitivity; however, the BLAST method can find closely related templates, which provides high and fast accuracy. So, the input data would be used in this step to search for protein structures that are evolutionarily related to the input against the template library of the SWISS-MODEL (SMTL) (Waterhouse et al., 2018).

The third step would be template selection. After the templates had been completed and estimated by Quaternary Structure Quality Estimate (QSQE) (Bertoni et al., 2017) and Global Model Quality Estimate (GMQE) (Biasini et al., 2014), and according to the expected quality of the resulting models, templates would be ranked (Waterhouse et al., 2018). For example, different models were built in the RBD domain, and many templates were selected automatically. So, to choose the best template, the SWISS-MODEL provides many alternative template options. Each template has a descriptive set of features that allow the user to select the best fit with the target protein. So, template 6vw.1.1. B was used in the RBD domain because it is a SARS-CoV-2 chimeric receptor-binding domain complexed with its human receptor ACE2. This template was selected because of its defining features and good interactive graphical views.

A 3D protein model would be generated after selecting the template as defined by the alignment of the target template to conserved atom coordinates. A full-atom protein model and loop modeling would generate the residue coordinates from the constructed amino acid non-conserved side chains. The SWISS-MODEL relies on the ProMod3 modeling engine and the OpenStructure computational structural biology framework to perform this step (Biasini et al., 2013). After building models for the target proteins using the SWISS-MODEL, PDB files were downloaded and checked on the PyMOL System.

After getting the PDF files, they were submitted to docking software known as SwarmDock. SwarmDock, a memetic docking algorithm in which the conformational, orientational, and translational degrees of freedom were developed using normal modes to perform flexible docking and are simultaneously optimized with a Solis and Wets local search algorithm using the Particle Swam Optimisation (PSO) metaheuristic (Moal & Bates, 2010). SwarmDock was used instead of other algorithms due to its simplistic energy function, in which, upon binding,

it will undergo significant conformational changes; thus, it can dock flexible structures successfully.

Compared to other docking methods, SwarmDock is different in that it filters the many putative structures using FFT correlations or combines the single independent trajectories' results using search space (Moal & Bates, 2010). As an emergent proportion of the system, the exploitation of narrow regions containing lower energy structures would be switched between them and the exploration of diffuse areas in search space, which would depend on the nature of the energy of the landscape. Surrounding the binding site, a correlated energy landscape would be used by SwarmDock as swarm members found low-energy positions that act as attractors for some swarm members. Furthermore, the equation describing the velocities of the swarm members has a spatially varying repulsion term that prevents the contraction of a dispersed swarm from occurring. When a swarm concentrates its efforts on one specific location with many low-energy structures, such as the actual biological interface, it has less of an impact on its contraction (Moal & Bates, 2010).

In the docking technique, a standard energy function is incorporated. The approach employs van der Waals and Coulombic terms. These names are between i and j associative atoms within the receptor and ligand, respectively (Moal & Bates, 2010). Also, a switching function between 7 and 9 (Ron = 7 and roff = 9) is used to eliminate any interruptions in the standard relation and prevent long-distance mathematical associations with an insignificant contribution to the interaction energy (Moal & Bates, 2010).

Energy function in the docking algorithm where energy is in kilo Joule per mole:

$$E_{int} = \sum_{i}^{atoms} \times \sum_{j}^{atoms} \times E_{i,j} \quad (1)$$

The PDB structures of binding proteins must follow three fundamental necessities: one of them is no missing residues, the other is the use of standard residues, and the TER phrase must be placed after each chain (Moal & Bates, 2010). The server will try substituting non-standard residues with standard residues if any criteria are not fulfilled. The server will also attempt to replace missing residues or residues with missing atoms if none of the conditions are met (Torchala et al., 2013). Afterward, the binding proteins are reduced and docked after being fixed (Moal & Bates, 2010).

Depending on how the assignment was submitted, a job may be completed using either local docking or full-blind docking with limitations. A reliable distribution of beginning sites has been developed for the receptor. When using the previous approach, the user may choose which receptor residues will be utilized in the binding site, which allows for more customization (Moal & Bates, 2010). Therefore, the server only makes use of the starting positions in the line of sight of at least one of the receptor compounds that have been chosen. Consequently, the server only creates solutions in the proximity of the receptor site selected for examination by the user as a result of this consequence (Torchala et al., 2013).

When submitting a task, the user can use either full-blind or local docking with constraints. Initially, starting points are formed homogeneously around the receptor. The user may choose the receptor residues as binding sites when using the latter technique. As a result, the server only utilizes the beginning points that are in the same line of sight as at least one of the receptors that have been selected. As a result, the server will only generate solutions near the receptor location specified by the user. The user will get an email with a link to retrieve the mended input structures, a ranked list of clusters, docked structures, information about the residue contacts, and the SwarmDock output file when the calculations have been completed.

The dataset will include energy measures for different biological structures categorized by variants, receptors, and combinations, representing the sequence of variants and receptors. Then, the data will be presented in the tables below (Table 2 to Table 13) to evaluate differences in energy measures across these groups The boxplots shown in Figures 6 to 11 illustrate the distribution of energy measures (S1, S2, and S3) across the different groups (variants, receptors, and their combinations). The boxplots reveal considerable variability within the groups, with some plots showing distinct differences in the means. These visualizations suggest that there may be significant variations across the groups. Therefore, formal statistical tests, such as ANOVA, Bartlett's tests, and Levene's test, were applied to assess the differences in the means and variances across the groups of variants, receptors, and their combinations. Also, Tukey's HSD test was applied to enhance the interpretation of pairwise differences. Examples of sequencing are shown below.

Sequence of variants and receptors

RB domain from 334 to 530 sequence

Wild type RBD (Using the template 6vw.1.1.B):

NLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCF
TNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN
YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRV
VVLSFELLHAPATVCGPKKS

Alpha (N501Y):

NLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCF
TNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN
YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTYGVGYQPYRV
VVLSFELLHAPATVCGPKKS

Zeta/Eta/ Lota (E484K):

NLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCF
TNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN
YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTNGVGYQPYRV
VVLSFELLHAPATVCGPKKS

Beta (K417N/ E484K/ N501Y):

NLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCF
TNVYADSFVIRGDEVRQIAPGQTGNIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN
YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRV
VVLSFELLHAPATVCGPKKS

Gamma (K417T/E484K/N501Y):

NLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCF
TNVYADSFVIRGDEVRQIAPGQTGTIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN
YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVKGFNCYFPLQSYGFQPTYGVGYQPYRV
VVLSFELLHAPATVCGPKKS

Omicron (G339D/S371L/ S373P/ S375F/ K417N/ N440K/ G446S/ S477N/ T478K/ E484A/
Q493R/ G496S/ Q498R/ N501Y/ Y505H):

NLCPFDEVFNATRFASVYAWNRKRISNCVADYSVLYNLAPFFTFKCYGVSPTKLNDLCF
TNVYADSFVIRGDEVRQIAPGQTGNIADYNYKLPDDFTGCVIAWNSKNLDSKVGGNYN
YLYRLFRKSNLKPFERDISTEIYQASNKPCNGVAGFNCYFPLRSYSFRPTYGVGHQPYRV
VVLSFELLHAPATVCGPKKS

Delta (L452R/ T478K):

NLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCF
TNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN
YRYRLFRKSNLKPFERDISTEIYQAGSKPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRV
VVLSFELLHAPATVCGPKKS

ACE2-Sequence:

STIEEQAKTFLDKFNHEAEDLFYQSSLASW

GRP78- Sequence (SBD=substrate-binding domain):

CPLTLGIETVGGVMTKLIPRNTVVPTKKSQIFSTASDNQPTVTIKVYEGERPLTKDNHLLG
TFDLTGIPPAPRGVPQIEVT

Table 2: Affinity and kinetic data (mean and SD) for RBD variant Alpha after docking it with Grp78 using Swarm dock.

Structure	Energy (kJ/mol)	Mean Energy (kJ/mol)	The standard deviation of the mean energy
6d.pdb	-39.77	-33.683	3.641
54a.pdb	-37.18	-29.816	3.402
4a.pdb	-34.56	-34.56	0
13a.pdb	-32.57	-32.57	0
31a.pdb	-31.91	-31.91	0
49a.pdb	-31.13	-28.982	2.163
67b.pdb	-30.52	-30.52	0
83d.pdb	-29.97	-29.97	0
4c.pdb	-29.91	-29.91	0
50b.pdb	-29.90	-29.90	0

Table 3: Affinity and kinetic data (mean and SD) for RBD variant either Zeta or Eta or Lota after docking it with Grp78 using Swarm dock.

Structure	Energy (kJ/mol)	Mean Energy (kJ/mol)	Standard Deviation of the mean energy
30a.pdb	-39.36	-33.847	4.208
17c.pdb	-35.51	-35.51	0
3c.pdb	-34.99	-27.537	2.670
53b.pdb	-34.04	-25.226	4.635
112a.pdb	-32.40	-32.40	0
84b.pdb	-32.34	-32.34	0
116b.pdb	-31.98	-31.98	0
114b.pdb	-31.25	-25.410	5.840
56d.pdb	-30.92	-27.820	3.10
33b.pdb	-30.88	-26.315	4.565

Table 4: Affinity and kinetic data (mean and SD) for RBD variant Beta after docking it with Grp78 using Swarm dock.

Structure	Energy (kJ/mol)	Mean Energy (kJ/mol)	Standard Deviation of the mean energy
16b.pdb	-40.76	-36.110	4.650
18b.pdb	-39.97	-39.97	0
62a.pdb	-38.09	-32.734	2.966
19d.pdb	-35.81	-35.81	0
5a.pdb	-34.92	-34.92	0
73b.pdb	-34.30	-34.30	0
31d.pdb	-34.10	-29.143	4.283
45b.pdb	-33.42	-28.624	4.696
28a.pdb	-33.22	-29.060	2.508
123b.pdb	-32.44	-29.767	2.653

Table 5: Affinity and kinetic data (mean and SD) for RBD variant Gamma after docking it with Grp78 using Swarm dock.

Structure	Energy (kJ/mol)	Mean Energy (kJ/mol)	Standard Deviation of the mean energy
82c.pdb	-39.63	-35.452	4.231
45a.pdb	-37.23	-37.23	0
70a.pdb	-37.22	-28.272	6.379
79c.pdb	-33.27	-29.185	4.085
7d.pdb	-32.88	-28.075	4.805
69c.pdb	-32.62	-32.62	0
45b.pdb	-32.44	-29.147	2.741
18a.pdb	-32.02	-26.284	2.837
85b.pdb	-31.85	-25.687	7.315
34d.pdb	-31.43	-31.43	0

Table 6: Affinity and kinetic data (mean and SD) for RBD variant Omicron after docking it with Grp78 using Swarm dock.

Structure	Energy (kJ/mol)	Mean Energy (kJ/mol)	Standard Deviation of the mean energy
31c.pdb	-38.18	-33.297	3.993
27b.pdb	-37.48	-32.370	5.934
126d.pdb	-37.01	-31.603	4.742
6d.pdb	-36.63	-28.943	6.103
99d.pdb	-36.49	-36.49	0
39d.pdb	-36.27	-36.27	0
18b.pdb	-35.74	-34.33	1.410
33a.pdb	-35.68	-32.83	2.850
45c.pdb	-34.41	-34.41	0
31b.pdb	-33.86	-33.86	0

Table 7: Affinity and kinetic data (mean and SD) for RBD variant Delta after docking it with Grp78 using Swarm dock.

Structure	Energy (kJ/mol)	Mean Energy (kJ/mol)	Standard Deviation of the mean energy
39c.pdb	-46.53	-41.375	5.155
88c.pdb	-38.62	-37.645	0.975
48a.pdb	-35.66	-35.66	0
45a.pdb	-34.53	-30.757	2.674
5c.pdb	-33.26	-33.26	0
61b.pdb	-31.97	-31.97	0
27d.pdb	-31.49	-31.49	0
111a.pdb	-30.83	-30.83	0
1c.pdb	-29.79	-23.987	6.080
66b.pdb	-29.54	-29.54	0

Table 8: Affinity and kinetic data (mean and SD) for RBD variant Beta after docking it with ACE2 using Swarm dock.

Structure	Energy (kJ/mol)	Mean Energy (kJ/mol)	Standard Deviation of the mean energy
64b.pdb	-32.69	-32.69	0
30b.pdb	-31.43	-31.43	0
6b.pdb	-31.31	-27.179	2.003
63b.pdb	-31.19	-31.19	0
59d.pdb	-30.23	-30.23	0
111b.pdb	-29.81	-25.993	3.422
32a.pdb	-29.76	-25.122	2.172
33b.pdb	-29.72	-29.72	0
63d.pdb	-29.56	-22.645	4.61
41a.pdb	-29.47	-25.957	2.671

Table 9: Affinity and kinetic data (mean and SD) for RBD variant Gamma after docking it with ACE2 using Swarm dock.

Structure	Energy (kJ/mol)	Mean Energy (kJ/mol)	Standard Deviation of the mean energy
102b.pdb	-40.38	-30.441	3.902
55a.pdb	-37.67	-37.67	0
81a.pdb	-32.44	-25.895	4.001
79d.pdb	-31.31	-29.555	1.755
43c.pdb	-30.75	-28.15	2.018
2a.pdb	-30.43	-30.43	0
84b.pdb	-30.04	-30.04	0
63b.pdb	-30.03	-30.03	0
70a.pdb	-29.99	-29.99	0
34b.pdb	-29.74	-29.74	0

Table 10: Affinity and kinetic data (mean and SD) for RBD variant Omicron after docking it with ACE2 using Swarm dock.

Structure	Energy (kJ/mol)	Mean Energy (kJ/mol)	Standard Deviation of the mean energy
103d.pdb	-39.11	-39.11	0
33c.pdb	-38.87	-29.416	3.655
42a.pdb	-37.48	-28.172	5.642
43b.pdb	-37	-37	0
102c.pdb	-33.42	-29.020	4.4
47d.pdb	-33.13	-28.510	4.62
32d.pdb	-33.08	-31.995	1.085
57d.pdb	-32.74	-32.74	0
80d.pdb	-32.52	-23.126	5.472
80a.pdb	-32.00	-32.00	0

Table 11: Affinity and kinetic data (mean and SD) for RBD variant Delta after docking it with ACE2 using Swarm dock. The best ten standard structures were used.

Structure	Energy (kJ/mol)	Mean Energy (kJ/mol)	Standard Deviation of the mean energy
36b.pdb	-30.27	-30.27	0
104d.pdb	-30.05	-30.05	0
14d.pdb	-29.82	-24.974	3.227
43b.pdb	-28.91	-28.91	0
25d.pdb	-28.00	-23.409	3.061
13a.pdb	-27.62	-22.050	4.675
56a.pdb	-27.60	-27.60	0
19c.pdb	-27.52	-27.52	0
35d.pdb	-27.02	-23.256	2.509
13b.pdb	-26.52	-23.114	2.636

Table 12: Affinity and kinetic data (mean and SD) for RBD variant Alpha after docking it with ACE2 using Swarm dock.

Structure	Energy (kJ/mol)	Mean Energy (kJ/mol)	Standard Deviation of the mean energy
43b.pdb	-33.44	-33.44	0
105c.pdb	-33.24	-26.405	6.835
44a.pdb	-31.65	-31.65	0
53b.pdb	-31.0	-31.0	0
107d.pdb	-30.57	-24.442	3.142
11c.pdb	-30.2	-30.2	0
78a.pdb	-29.51	-29.51	0
0a.pdb	-29.16	-26.33	2.830
55d.pdb	-28.98	-28.98	0
30a.pdb	-28.9	-28.9	0

Table 13: Affinity and kinetic data (mean and SD) for RBD variant either Zeta or Eta or Lota after docking it with ACE2 using Swarm dock.

Structure	Energy (kJ/mol)	Mean Energy (kJ/mol)	Standard Deviation of the mean energy
56b.pdb	-42.67	-32.750	9.92
69c.pdb	-34.33	-28.43	3.527
69d.pdb	-33.63	-30.7	2.310
1a.pdb	-32.50	-26.75	5.053
55d.pdb	-31.10	-30.075	1.246
44c.pdb	-30.93	-30.93	0
45b.pdb	-29.80	-24.24	3.954
100a.pdb	-29.10	-18.328	3.696
11d.pdb	-28.80	-28.80	0
47b.pdb	-28.56	-28.56	0

Figure 6. Variation in Binding Energy (S1) Across SARS-CoV-2 Variants

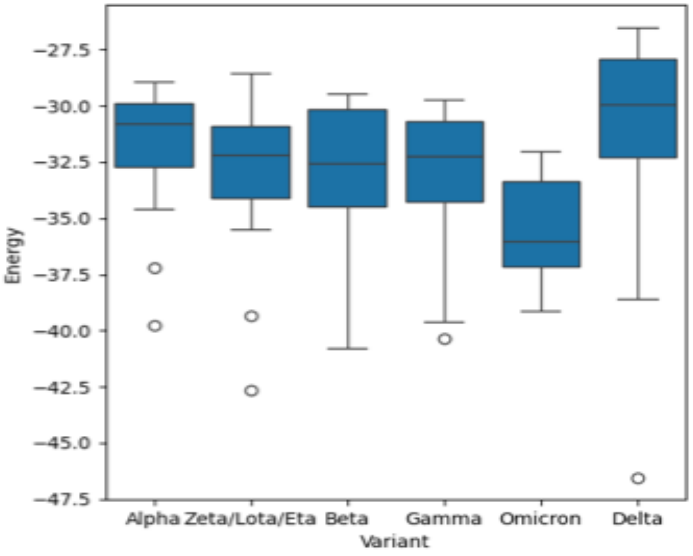


Figure 7. Distribution of Mean Energy (S2) Across SARS-CoV-2 Variants

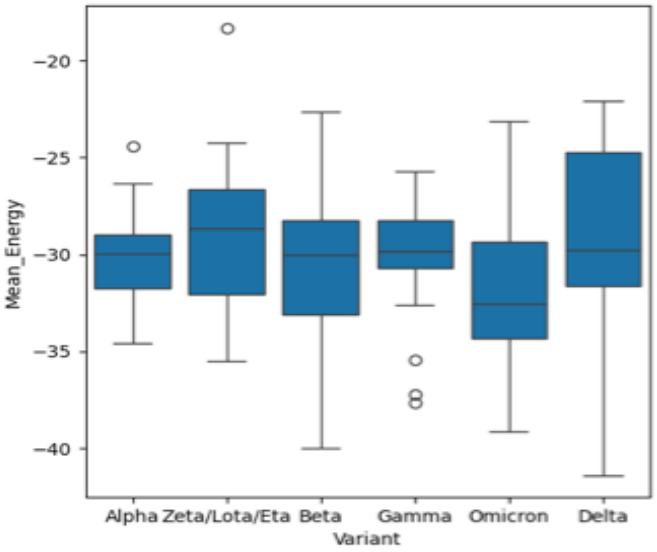


Figure 8. Distribution of Energy Variability (S3) Across SARS-CoV-2 Variants

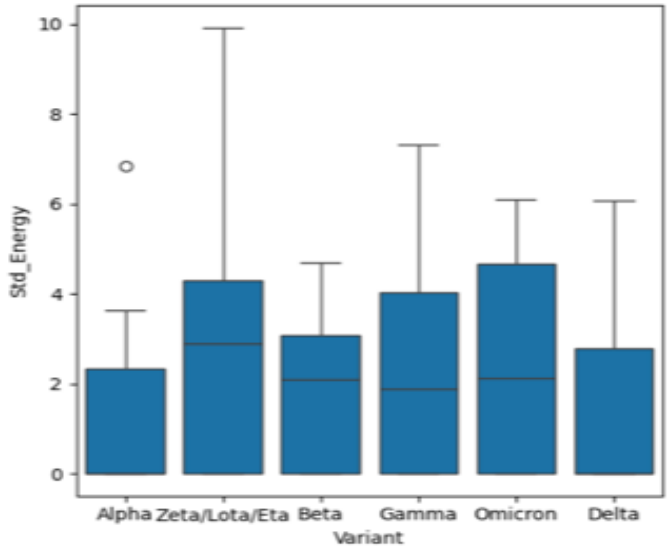


Figure 9. Energy Distribution (S1) Across ACE2 and GRP78 Receptors

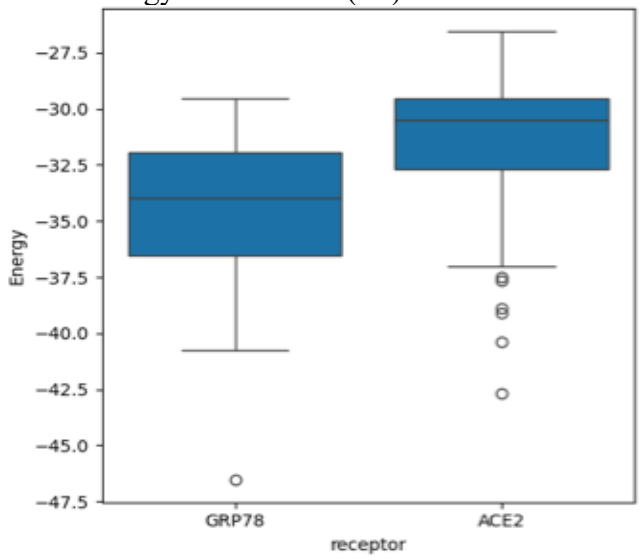


Figure 10. Distribution of Mean Energy (S2) Across ACE2 and GRP78 Receptors

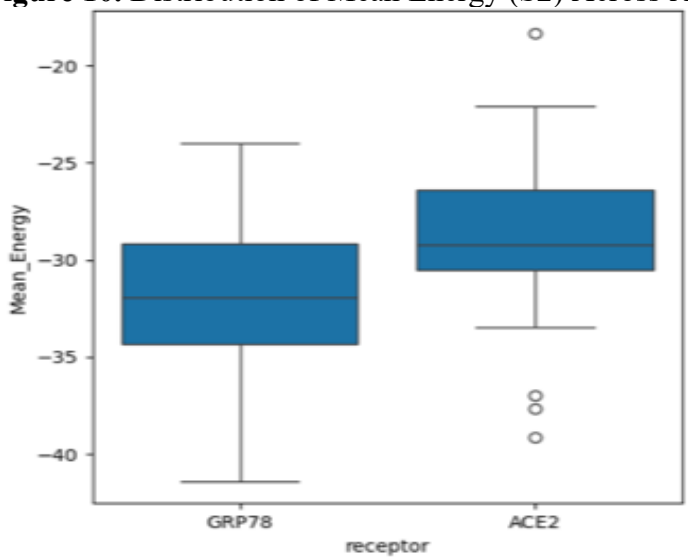
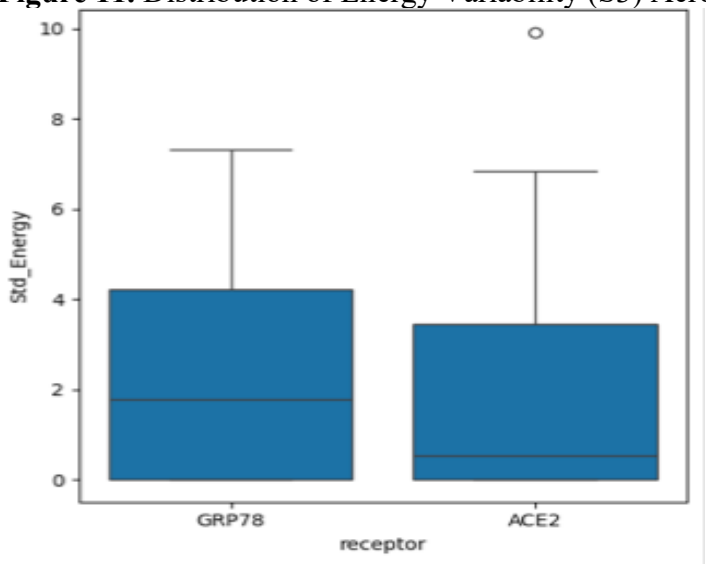


Figure 11. Distribution of Energy Variability (S3) Across ACE2 and GRP78 Receptors



4.1.2 Experimental Results

The results section of this study presents the findings from the statistical analyses conducted on the binding energy measures (Energy (S1), Mean Energy (S2), and Standard Deviation of Energy (S3)) across different groups of SARS-CoV-2 variants, receptors, and their combinations. The analysis was performed on a sample of 6 SARS-CoV-2 variants and 2 receptors. Each group consists of 10 observations, with each observation corresponding to one of the best 10 PDB structures based on docking results from SwarmDock. For each variant-receptor combination, we considered the binding energy after docking and selected the 10 best PDB structures based on the most favorable binding interactions.

The variants used in this study are as follows:

- **Variant 1**: Alpha
- **Variant 2**: Zeta
- **Variant 3**: Beta
- **Variant 4**: Gamma
- **Variant 5**: Omicron
- **Variant 6**: Delta

The receptors used in this study are as follows:

- Receptor 1: GRP78
- Receptor 2: ACE2

The primary objective was to assess whether these energy measures show significant differences across the different groups (Variants, Receptors, and Variants*Receptors Combinations). The statistical tests employed include Multiple ANOVA, Bartlett's Test for Equal Variances, and Levene's Test. The significance level used for these tests is 5% (p-value < 0.05).

The statistical tests evaluate the following:

1. **Tests applied to variants**: Whether the means of Energy (S1), Mean Energy (S2), and Standard Deviation of Energy (S3) differ significantly between the 6 variants.
2. **Tests applied to receptors**: Whether the means of Energy (S1), Mean Energy (S2), and Standard Deviation of Energy (S3) differ significantly between the 2 receptors (GRP78 and ACE2).
3. **Tests applied to combinations**: Whether the means of Energy (S1), Mean Energy (S2), and Standard Deviation of Energy (S3) differ significantly for the interaction between variants and receptors.

To further investigate which specific group comparisons showed significant differences, Tukey's HSD test was performed after the ANOVA. The mean differences (meandiff) and adjusted p-values (p-adj) for each pairwise comparison was calculated. Table 16 presents the results of Tukey's Honest Significant Difference (HSD) post hoc test. This test was applied only to those comparisons where the ANOVA indicated statistically significant differences between group means. The purpose was to identify which specific pairs of groups contributed to the overall significance.

The detailed outcomes of the hypotheses tested, along with the results from the multiple ANOVA, Bartlett's Test, and Levene's Test, and Tukey's HSD test, are presented in Table 14, Table 15, and Table 16.

Table 14. Results of Multiple ANOVA Test on Energy Measures

Null hypotheses (H0) & variables	F-statistic	p-value
Tests applied to 6 variants		
H0: mean of S1 equal across all variants H1: at least one mean is different	3.8249	3.0636e-03
H0: mean of S2 equal across all variants H1: at least one mean is different	1.7909	1.2023e-01
H0: mean of S3 equal across all variants H1: at least one mean is different	1.4937	1.9736e-01
Tests applied to 2 receptors		
H0: mean of S1 equal across both receptors H1: at least one mean is different	22.6625	5.5317e-06
H0: mean of S2 equal across both receptors H1: at least one mean is different	19.3578	2.3903e-05
H0: mean of S3 equal across both receptors H1: at least one mean is different	0.5410	4.6348e-01
Tests applied to combinations of 6 variants and 2 receptors		
H0: mean of S1 equal across all combinations H1: at least one mean is different	5.8637	2.1419e-07
H0: Mean of S2 equal across all combinations H1: at least one mean is different	3.9013	9.1950e-05
H0: Mean of S3 equal across all combinations H1: at least one mean is different	1.1360	3.4097e-01
The number of samples in each group is ten. The statistical significance level is set to 5%.		

Table 15. Results of Bartlett's and Levene's Tests for Variance Equality

Null hypotheses (H0) & variables	**Bartlet Test**		**Levine Test**	
	F-statistic	**p-value**	**F-statistic**	**p-value**
Tests applied to 6 variants				
H0: Variance of S1 is equal across all variants H1: at least one variance is different	11.4619	4.2953e-02	0.5815	7.1409e-01
H0: Variance of S2 is equal across all variants H1: at least one variance is different	10.1827	7.0220e-02	1.8082	1.1674e-01
H0: Variance of S3 is equal across all variants H1: at least one variance is different	3.9343	5.5891e-01	1.6838	1.4409e-01
Tests applied to 2 receptors				
H0: Variance of S1 is equal across both receptors H1: at least one variance is different	0.0042	9.4828e-01	0.6243	4.3105e-01
H0: Variance of S2 is equal across both receptors H1: at least one variance is different	0.0470	8.2845e-01	0.1321	7.1692e-01
H0: Variance of S3 is equal across both receptors H1: at least one variance is different	0.0404	8.4075e-01	0.6966	4.0561e-01
Tests applied to combinations of 6 variants and 2 receptors				
H0: Variance of S1 is equal across all combinations H1: at least one variance is different	40.2081	3.2958e-05	1.0889	3.7704e-01
H0: Variance of S2 is equal across all combinations H1: at least one variance is different	13.3356	2.7196e-01	1.1452	3.3417e-01
H0: Variance of S3 is equal across all combinations H1: at least one variance is different	8.3852	6.7844e-01	0.9459	5.0009e-01
The number of samples in each group is ten. The statistical significance level is set to 5%.				

Table 16. Tukey HSD Post Hoc Comparison Results

Test	Grouping Factor	Significant Comparisons (Reject = True)	Meandiff	p-adj	95% CI (Lower – Upper)
1	Variant (S1 Energy)	1 vs 5	-3.8515	0.0072	-6.9951 – -0.7079
1	Variant (S1 Energy)	5 vs 6	4.2775	0.0019	1.1339 – 7.4211
2	Receptor (S1 Energy)	1 vs 2	2.8997	0.0000	1.6935 – 4.1059
3	Receptor (S2 Mean Energy)	1 vs 2	2.9697	0.0000	1.6331 – 4.3064
4	Variant Receptor (S1 Energy)	2 vs 12	5.034	0.0151	0.5328 – 9.5352
4	Variant Receptor (S1 Energy)	3 vs 7	5.038	0.0149	0.5368 – 9.5392
4	Variant Receptor (S1 Energy)	3 vs 9	5.186	0.0104	0.6848 – 9.6872
4	Variant Receptor (S1 Energy)	3 vs 12	7.37	0.0000	2.8688 – 11.8712
4	Variant Receptor (S1 Energy)	4 vs 12	5.726	0.0026	1.2248 – 10.2272
4	Variant Receptor (S1 Energy)	5 vs 7	5.51	0.0046	1.0088 – 10.0112

4	Variant Receptor (S1 Energy)	5 vs 9	5.658	0.0031	1.1568 – 10.1592
4	Variant Receptor (S1 Energy)	5 vs 12	7.842	0.0000	3.3408 – 12.3432
4	Variant Receptor (S1 Energy)	6 vs 12	5.889	0.0016	1.3878 – 10.3902
4	Variant Receptor (S1 Energy)	11 vs 12	6.602	0.0002	2.1008 – 11.1032
5	Variant Receptor (S2 Mean_Energy)	5 vs 8	5.484	0.0337	0.2142 – 10.7538
5	Variant Receptor (S2 Mean_Energy)	5 vs 12	7.325	0.0006	2.0552 – 12.5948
5	Variant Receptor (S2 Mean_Energy)	3 vs 12	6.9285	0.0015	1.6587 – 12.1983
The number of samples in each group is ten. The statistical significance level is set to 5%.					

Chapter 5

5.1 Discussion

Based on the results from the multiple ANOVA tests in Table 14, here are the main findings:

1. Tests Applied to Variants

- **H0: Mean of S1 equal across all variants**
- **H1: at least one mean of S1 is different across variants**
    - F-statistic: 3.8249
    - p-value: 3.0636e-03
    - Interpretation: The p-value is below the standard significance level (0.05), so the null hypothesis is rejected. This suggests that the mean energy differs significantly across the variants.
- **H0: Mean of S2 equal across all variants**
- **H1: at least one mean of S2 is different across variants**
    - F-statistic: 1.7909
    - p-value: 1.2023e-01
    - Interpretation: Because the p-value is higher than 0.05, the null hypothesis cannot be rejected. This suggests no significant difference in the mean of mean energy across the variants exists.
- **H0: Mean of S3 is equal across all variants**
- **H1: at least one mean of S3 is different across variants**
    - F-statistic: 1.4937
    - p-value: 1.9736e-01
    - Interpretation: The null hypothesis cannot be rejected because the p-value is greater than 0.05. This suggests that there is no significant difference in the energy across the variants.

2. Tests Applied to Receptors

- **H0: Mean of S1 equal across both receptors**
- **H1: at least one mean of S1 is different across both receptors**
    - F-statistic: 22.6625
    - p-value: 5.5317e-06
    - Interpretation: The null hypothesis is rejected due to the very small p-value (much smaller than 0.05). This indicates that the mean energy differs significantly across the receptors.
- **H0: Mean of S2 equal across both receptors**
- **H1: at least one mean of S2 is different across both receptors**
    - F-statistic: 19.3578
    - p-value: 2.3903e-05
    - Interpretation: The null hypothesis is rejected because the p-value is less than 0.05. This suggests a significant difference in the mean of mean energy across the receptors.
- **H0: Mean of S3 is equal across both receptors**
- **H1: at least one mean of S3 is different across both receptors**
    - F-statistic: 0.5410

- p-value: 4.6348e-01
- Interpretation: The null hypothesis cannot be rejected since the p-value is greater than 0.05. This suggests no significant difference in the standard deviation of energy across the receptors.

3. Tests Applied to Combinations

- **H0: Mean of S1 equal across all combinations**
- **H1: at least one mean of S1 is different across all combinations**
  - F-statistic: 5.8637
  - p-value: 2.1419e-07
  - Interpretation: The null hypothesis is rejected because the p-value is much smaller than 0.05. This indicates that the energy values differ significantly across the combinations of variants and receptors.
- **H0: Mean of S2 equal across all combinations**
- **H1: at least one mean of S2 is different across all combinations**
  - F-statistic: 3.9013
  - p-value: 9.1950e-05
  - Interpretation: The null hypothesis is rejected since the p-value is less than 0.05. This suggests that the mean energy differs significantly across the combinations of variants and receptors.
- **H0: Mean of S3 equal across all combinations**
- **H1: at least one mean of S3 is different across all combinations**
  - F-statistic: 1.1360
  - p-value: 3.4097e-01
  - Interpretation: The null hypothesis cannot be rejected because the p-value is greater than 0.05. This indicates no significant difference in the standard deviation of energy across the combinations of variants and receptors.

Summary of Findings:

1. **S1**: Significant differences in energy across variants, receptors, and combinations (i.e., rejected null hypotheses).
2. **S2**: Only significant for receptors and combinations (rejected null hypotheses), but not for variants.
3. **S3**: No significant differences observed for standard deviation across variants or combinations, but significant across receptors.

This suggests that energy and mean energy are the key differentiators for receptors and combinations, while the standard deviation of energy does not show significant differences in most cases.

Based on the results from the multiple Bartlett tests in Table 15, here are the main findings:

1. Tests Applied to Variants

- **H0: Variance of S1 is equal across all variants**
  - F-statistic: 11.4619
  - p-value: 4.2953e-02
  - Interpretation: The null hypothesis is rejected because the p-value is less than the common significance threshold of 0.05. This indicates that the variances of energy are significantly different across the variants.

- **H0: Variance of S2 is equal across all variants**
  - F-statistic: 10.1827
  - p-value: 7.0220e-02
  - Interpretation: The null hypothesis cannot be rejected because the p-value is greater than 0.05, indicating that the variances of the mean energy are not significantly different across the variants.
- **H0: Variance of S3 equal across all variants**
  - F-statistic: 3.9343
  - p-value: 5.5891e-01
  - Interpretation: The null hypothesis cannot be rejected as the p-value is much higher than 0.05, indicating no significant difference in the standard deviations of energy across the variants.

2. Tests Applied to Receptors

- **H0: Variance of S1 is equal across both receptors**
  - F-statistic: 0.0042
  - p-value: 9.4828e-01
  - Interpretation: The null hypothesis cannot be rejected because the p-value is much greater than 0.05, suggesting that the variances of energy are equal across the two receptors.
- **H0: Variance of S2 is equal across both receptors**
  - F-statistic: 0.0470
  - p-value: 8.2845e-01
  - Interpretation: The null hypothesis cannot be rejected because the p-value is greater than 0.05, suggesting that the variances of mean energy are equal across the two receptors.
- **H0: Variance of S3 is equal across both receptors**
  - F-statistic: 0.0404
  - p-value: 8.4075e-01
  - Interpretation: The null hypothesis cannot be rejected because the p-value is greater than 0.05, indicating no significant difference in the standard deviations of energy across the two receptors.

3. Tests Applied to Combinations

- H0: Variance of S1 is equal across all combinations
  - F-statistic: 40.2081
  - p-value: 3.2958e-05
  - Interpretation: The null hypothesis is rejected because the p-value is much smaller than 0.05, indicating that the variances of energy differ significantly across the combinations of variants and receptors.
- **H0: Variance of S2 is equal across all combinations**
  - F-statistic: 13.3356
  - p-value: 2.7196e-01
  - Interpretation: The null hypothesis cannot be rejected since the p-value is greater than 0.05, suggesting that the variances of mean energy are not significantly different across the combinations.

- **H0: Variance of S3 is equal across all combinations**
    - F-statistic: 8.3852
    - p-value: 6.7844e-01
    - Interpretation: The null hypothesis cannot be rejected because the p-value is much higher than 0.05, indicating that there is no significant difference in the standard deviations of energy across the combinations of variants and receptors.

Summary of Findings:

1. **S1**: Significant differences in the variances of energy across the variants and combinations (i.e., rejected null hypothesis), but no significant differences across receptors.
2. **S2**: No significant differences in variances across any group (variants, receptors, combinations).
3. **S3**: No significant differences in variances for standard deviation of energy across variants, receptors, or combinations.

This indicates that **S1** is the primary factor where variances differ across variants and combinations, while **S2** and **S3** show no significant differences.
Based on Table 15, here are the main findings for Levene's Test:

4. **Tests Applied to Variants**

- **H0: Variance of S1 is equal across all variants**
    - F-statistic: 0.5815
    - p-value: 0.7141
    - Interpretation: The null hypothesis is not rejected because the p-value is greater than the common significance threshold of 0.05. This indicates that the variances of Energy are not significantly different across the variants.
- **H0: Variance of S2 is equal across all variants**
    - F-statistic: 1.8082
    - p-value: 0.1167
    - Interpretation: The null hypothesis is not rejected. The p-value exceeds 0.05, suggesting that the variances of Mean Energy are not significantly different across the variants.
- **H0: Variance of S3 equal across all variants**
    - F-statistic: 1.6838
    - p-value: 0.1441
    - Interpretation: The null hypothesis is not rejected. This implies that the variances of S3 are not significantly different among the variants.

5. **Tests Applied to Receptors**

- **H0: Variance of S1 is equal across both receptors**
    - F-statistic: 0.6243
    - p-value: 0.4311
    - Interpretation: The null hypothesis is not rejected. The variances of Energy are not significantly different between the two receptors.

- **H0: Variance of S2 is equal across both receptors**
  - F-statistic: 0.1321
  - p-value: 0.7169
  - Interpretation: The null hypothesis is not rejected, indicating that the variances of Mean Energy are not significantly different across receptors.
- **H0: Variance of S3 is equal across both receptors**
  - F-statistic: 0.6966
  - p-value: 0.4056

  Interpretation: The null hypothesis is not rejected, so the variances of Std Energy are considered equal across the receptor groups.

6. Tests Applied to Combinations

- **H0: Variance of S1 is equal across all combinations**
  - F-statistic: 1.0889
  - p-value: 0.3770
  - Interpretation: The null hypothesis is not rejected. Variances of Energy across variant-receptor combinations are not significantly different.
- **H0: Variance of S2 is equal across all combinations**
  - F-statistic: 1.1452
  - p-value: 0.3342
  - Interpretation: The null hypothesis is not rejected, indicating that Mean Energy variances are not significantly different among the combinations.
- **H0: Variance of S3 is equal across all combinations**
  - F-statistic: 0.9459
  - p-value: 0.5001
  - Interpretation: The null hypothesis is not rejected. There is no significant difference in Std Energy variances across the variant-receptor combinations.

Summary of Findings:

1. **S1 (Energy):** No significant differences in the variances across variants, receptors, or variant-receptor combinations (i.e., failed to reject the null hypothesis in all cases).
2. **S2 (Mean Energy):** No significant differences in the variances across any group (variants, receptors, or combinations).
3. **S3 (Standard Deviation of Energy):** No significant differences in the variances across variants, receptors, or combinations.

Based on Table 16, the post-hoc Tukey HSD tests revealed several statistically significant pairwise differences, offering deeper insights into how variants, receptors, and their combinations affect the energy profiles of the structures analyzed.

1. **Effect of Variant on Energy (S1):** Significant differences were observed between specific variant pairs, notably:
   - Variant 5 exhibited significantly lower energy levels compared to Variant 1, suggesting it may bind or stabilize more efficiently under the modeled conditions.

- Conversely, Variant 6 showed significantly higher energy levels than Variant 5, indicating potentially less favorable energetic interactions.

These findings underscore the structural or sequence differences among variants that might influence binding affinity or stability.

2. **Effect of Receptor on Energy and Mean Energy (S1 & S2):** The analysis also demonstrated that:
   - Receptor 2 yielded significantly higher energy values than Receptor 1, both in terms of raw energy (S1) and mean energy (S2). This could indicate that Receptor 2 introduces a less energetically favorable environment for binding, possibly due to conformational or electrostatic differences.
3. **Variant × Receptor Interactions (S1):** The most pronounced differences emerged when analyzing combinations of variant and receptor:
   - Group 12 (Variant 6 with Receptor 2) consistently had significantly higher energy levels compared to multiple other groups, including Groups 2, 3, 4, 5, 6, and 11.
   - Notably, Groups 3 and 5 were significantly lower in energy compared to Groups 7, 9, and 12, reinforcing that group-specific interactions influence binding energy beyond the additive effects of variant or receptor alone.

   These findings highlight the importance of synergistic interactions between specific variants and receptors—a combination that may amplify or mitigate binding energy significantly.
4. **Variant × Receptor Interactions (S2 – Mean Energy):** Mean energy comparisons further confirmed that:
   - Group 12 again displayed significantly higher mean energy compared to Groups 3 and 5, reaffirming the earlier observation that this particular combination leads to less favorable binding.
   - Additionally, Group 8 showed higher mean energy than Group 5, albeit with less magnitude.

These results emphasize that certain variant–receptor pairs are energetically more compatible, which could be crucial for targeted docking, therapeutic design, or understanding mutation impacts.

Summary of findings:

Overall, these statistically significant pairwise differences confirm that both the variant and receptor independently influence binding energy, but their interaction plays an even more critical role. Group 12 repeatedly emerged as the least favorable energetically, suggesting it could serve as a reference point or control in future experiments.

The results from ANOVA, Bartlett's test, and Levene's test provide valuable insights into the differences in energy measures across SARS-CoV-2 variants, receptors, and their combinations. The findings suggest that energy (S1) and mean energy (S2) are the key differentiators in receptor and combination interactions, whereas the standard deviation of energy (S3) does not exhibit significant variability in most cases. These results have important implications for understanding how viral variants interact with host receptors and may inform the development of therapeutic strategies targeting these interactions.

The ANOVA results revealed significant differences in energy (S1) across variants, receptors, and their combinations, leading to the rejection of the null hypotheses for these groupings. This indicates that the interaction energies between SARS-CoV-2 variants and receptors differ significantly, supporting the idea that these energy measures can help distinguish between different viral strains and their binding affinities to host receptors. The significant differences in mean energy (S2) between receptors and combinations, but not among variants, highlight the relevance of receptor identity in influencing average binding energy. This suggests that while mean energy is an important metric for receptor-based comparisons, it may not be sufficient alone to differentiate between viral variants.

In contrast, the standard deviation of energy (S3) did not show significant differences in means across variants, receptors, or combinations in the ANOVA analysis. This indicates that the variability in energy interactions remains relatively stable across these groups, making standard deviation a less informative measure for group differentiation.

The Bartlett's test, which assesses homogeneity of variances, indicated significant variance differences in energy (S1) across variants and variant-receptor combinations, but not across receptors. This finding suggests that while the receptor types exhibit consistent energy interaction variance, the SARS-CoV-2 variants and their combinations contribute more dynamically to variance changes. Conversely, no significant variance differences were observed for mean energy (S2) or standard deviation (S3) across any grouping in Bartlett's test, implying stability in these metrics' variability across conditions.

To further assess variance consistency, the Levene's test—a more robust test under violations of normality—was conducted. Unlike Bartlett's test, Levene's test results did not indicate any significant differences in variances for energy (S1), mean energy (S2), or standard deviation (S3) across variants, receptors, or combinations. This discrepancy suggests that the variance heterogeneity detected by Bartlett's test for energy may be influenced by deviations from normality rather than true variance differences. Therefore, Levene's results provide stronger evidence that the assumption of equal variances is generally upheld, and that the energy-related measures are consistent across groups in terms of variability.

This cross-validation highlights the importance of using multiple statistical methods when evaluating assumptions. The consistency of Levene's results across all metrics strengthens the argument that, while mean values differ meaningfully, particularly for S1 and S2, the underlying variance structure remains relatively stable, lending further robustness to the interpretation of ANOVA findings.

The results from Levene's test differed from those of Bartlett's test in assessing the homogeneity of variances. This discrepancy can be attributed to the underlying assumptions of each test. Levene's test does not require the assumption of normality, making it more suitable for data that may deviate from a normal distribution or contain outliers. In contrast, Bartlett's test assumes normality, and its sensitivity to departures from this assumption can lead to misleading conclusions in such cases.

Furthermore, Levene's test is more robust when applied to datasets with limited sample sizes, whereas Bartlett's test may become unreliable under the same conditions. Therefore, in contexts where data may not meet strict normality assumptions or where sample sizes are constrained, Levene's test provides a more dependable assessment of variance equality.

The post-hoc Tukey HSD tests offered additional depth by identifying specific pairwise group differences underlying the significant ANOVA results. For energy (S1), Variant 5 was found to have significantly lower energy values than Variant 1, suggesting potentially more stable or

favorable interactions. In contrast, Variant 6 exhibited higher energy values than Variant 5, implying a less favorable binding affinity. At the receptor level, Receptor 2 displayed significantly higher energy values compared to Receptor 1, which could suggest conformational or electrostatic influences reducing binding efficiency. Moreover, the combined variant–receptor groups showed particularly pronounced differences. Group 12 (Variant 6 with Receptor 2) consistently emerged as the least favorable in terms of energy, with significantly higher values than several other combinations (e.g., Groups 2, 3, 4, 5, 6, and 11). For mean energy (S2), similar trends were observed, with Group 12 again standing out as having significantly higher mean energy than Groups 3 and 5. These pairwise differences provide compelling evidence that specific variant–receptor pairings influence energy interactions in a synergistic manner and should be closely examined in future therapeutic and structural modeling efforts.

In summary, the findings indicate that energy (S1) is the most significant factor in differentiating between SARS-CoV-2 variants, receptors, and their combinations. Mean energy (S2) also exhibits discriminative power, particularly when comparing receptor types and their interactions with variants, though it is not sensitive enough to distinguish between variants alone. Standard deviation (S3) contributes little to group-level differentiation in both mean and variance measures. While Bartlett's test initially suggested some heterogeneity in variance for energy across certain groups, the Levene's test confirmed that variances are statistically equivalent across all categories—providing stronger support for homogeneity and reinforcing the reliability of energy as a comparative measure. These findings underscore the importance of considering energy-based metrics, particularly S1, when evaluating viral-receptor interactions and suggest a promising path for identifying binding patterns that could inform therapeutic strategies or vaccine development.

.

5.2 Performance Measures

To assess and interpret the effectiveness of the statistical analyses conducted in this study, several performance measures were considered, particularly focusing on the sensitivity, specificity, and robustness of the statistical tests applied to the energy data (S1: Energy, S2: Mean Energy, S3: Standard Deviation of Energy).

1.4 Significance Testing (p-values)

The primary performance metric used was the p-value from ANOVA, Bartlett's test, and Levene's test. These p-values served to determine the statistical significance of differences in means or variances across SARS-CoV-2 variants, receptors, and their combinations. A p-value less than the alpha level (commonly set at 0.05) indicated a rejection of the null hypothesis and, therefore, a statistically significant difference.

- ANOVA tests were used to detect mean differences across groups.
- Bartlett's test focused on detecting variance heterogeneity under the assumption of normality.
- Levene's test provided a robust measure of variance equality, especially under non-normal data distributions.

2.4 Robustness to Assumptions

Robustness was evaluated by comparing results from Bartlett's and Levene's tests:

- Bartlett's test is sensitive to deviations from normality, which can lead to false positives.
- Levene's test, being more robust to non-normal distributions, was used to validate the variance equality assumptions found in Bartlett's test.

The alignment (or lack thereof) between the two tests helped determine whether variance differences were genuine or artifacts of distributional assumptions.

3.4 Consistency Across Measures

The consistency of statistical outcomes across S1, S2, and S3 provided a qualitative performance check:

- S1 consistently showed statistically significant mean differences across variants, receptors, and combinations, indicating strong discriminative power.
- S2 showed partial discriminative power, performing well with receptor-level differences but not variant-level.
- S3 consistently lacked significant differences, suggesting that standard deviation was not a reliable performance metric for distinguishing group characteristics.

4.4 Interpretability and Practical Relevance

Performance was also gauged by the biological interpretability of the results:

- Measures that aligned with known viral-receptor interaction mechanisms (e.g., S1 energy differentiation across receptors) were considered more biologically meaningful.
- Metrics with low variability across groups (e.g., S3) were deemed less informative for practical applications such as drug targeting or receptor affinity profiling.

Overall, the performance of the statistical tests supports the conclusion that S1 (Energy) is the most reliable and informative measure for distinguishing between variants, receptors, and their interactions. The use of multiple tests provided a comprehensive assessment of both central tendency and dispersion, with Levene's test confirming the reliability of variance assumptions. These performance evaluations enhance the credibility and interpretability of the analytical framework applied in this study.

.

5.3 Merits of the Study and Implications

This study offers several significant contributions to the field of virology and molecular biology:

1. **Novel Insight into Viral Interactions**: This study provides a better understanding of the molecular mechanisms controlling viral binding and entry by examining the energy interactions between SARS-CoV-2 variants and the host receptors ACE2 and GRP78. These insights are crucial for explaining differences in infectivity and transmissibility

among variants, thus contributing to a deeper understanding of viral-host interactions and informing therapeutic strategies.

2. **Comparative Analysis Using Robust Statistical Methods**: Utilizing multiple ANOVA tests to assess differences in means and Bartlett's test to check the equality of variances ensures a rigorous statistical comparison of interaction energies. Additionally, Levene's test is used to robustly assess variance homogeneity, accounting for potential violations of normality. The application of Tukey's HSD test for post-hoc analysis further strengthens the findings by identifying which specific group comparisons are significant. This comprehensive approach aids in determining the most discriminative energy metrics and enhances the reliability of the results.
3. **Identification of Potential Therapeutic Targets**: Targeted antiviral treatments can be developed by identifying which energy measurements distinguish between variants and receptor interactions. Understanding the binding kinetics of GRP78, a developing alternative receptor, may open up new therapeutic approaches aimed at preventing viral entry, especially in cases where ACE2 is not the primary receptor involved.
4. **Contribution to Epidemiological Predictions**: The results of this study on differential binding energies can help predict if new SARS-CoV-2 variants will be transmissible and harmful. By identifying variants with higher binding affinities that may confer increased transmission risks, this research could support vaccine development and public health efforts aimed at mitigating the spread of the virus.

The implications of this study extend beyond theoretical knowledge, impacting both clinical and public health domains:

1. **Therapeutic Development**: The discovery lays the groundwork for developing small compounds or antibodies that can interfere with variant-receptor interactions and possibly block viral entry. By identifying energy metrics that significantly differentiate these interactions, targeted therapies can be developed to neutralize the virus.
2. **Public Health Surveillance**: Proactive public health interventions may be facilitated by the identification of energy profiles linked to highly transmissible variants, helping to develop early warning systems for new strains. This can enable faster response times and better-targeted measures to curb the spread of more dangerous variants.
3. **Vaccine Design and Efficacy**: To improve cross-variant protection, vaccine developers can optimize immunogens targeting conserved areas involved in receptor binding. The insights gained from how various variants interact with ACE2 and GRP78 could guide the design of more broadly protective vaccines that are capable of neutralizing a wide range of SARS-CoV-2 variants.
4. **Broader Application to Viral Evolution Studies**: The analytical framework used in this work can be extended to other viral systems, helping to better understand how mutations impact viral-host interactions and supporting broader studies of viral evolution.

Overall, this study not only advances our knowledge of the interactions between SARS-CoV-2 variants and their receptors but also offers actionable insights that may influence future treatment and prevention strategies for both current and future coronavirus threats.

5.4 Limitations of the Study

While this study provides valuable insights into the interaction energies between SARS-CoV-2 variants and the receptors ACE2 and GRP78, it is important to acknowledge its limitations:

1. **Reliance on Computational Modeling**: The study depends on computational methods to estimate interaction energies, which rely on static structural data. While these methods are powerful for hypothesis generation, they may not capture the dynamic nature of protein-protein interactions in living systems. Therefore, the results might not fully reflect biological scenarios, and future studies should include experimental validation.
2. **Assumptions in Statistical Analysis**: The study assumes normality and homogeneity of variances in its statistical analysis using ANOVA, which may oversimplify the interactions. Although Bartlett's test is used to check the equality of variances, violations of these assumptions could affect the validity of the results. Levene's test, being more robust to non-normality, confirmed the consistency of variances across groups, providing additional support to the reliability of the findings. However, any violation of these assumptions could still limit the generalizability of the results.
3. **Limited Scope of Receptors**: The study focuses on two specific receptors, ACE2 and GRP78, which limits its scope. Other receptors or co-receptors may influence SARS-CoV-2 entry, meaning the findings may not fully capture the complexity of viral-host interactions.
4. **Generalizability of Results**: The study examines specific SARS-CoV-2 variants relevant at the time of research. As new variants continue to emerge, the binding characteristics of the virus could change, potentially limiting the relevance of these findings to future variants. Additionally, not all energy measures examined may correlate directly with biological outcomes like infectivity or pathogenicity, which requires cautious interpretation.
5. **Overlooking Other Influential Factors**: The study does not account for other factors such as post-translational modifications of the receptors, the membrane microenvironment, or interactions with other host proteins, which could affect binding energetics. These unconsidered variables could act as confounding factors, impacting the conclusions.

To address these limitations, future research should include experimental validation, explore additional receptors and SARS-CoV-2 variants, and use more sophisticated statistical models that account for complex interactions and potential confounders. Despite these limitations, this study provides a valuable framework for understanding the molecular mechanisms of SARS-CoV-2 variant interactions, paving the way for future investigations and therapeutic developments.

5.5 Future Research Directions

This study lays the groundwork for understanding the interaction energies between SARS-CoV-2 variants and the receptors ACE2 and GRP78, but there are several avenues for future research that could deepen and expand these insights. One important direction is to validate the computational findings through experimental methods. Laboratory-based assays, such as surface plasmon resonance or isothermal titration calorimetry, could provide direct measurements of binding affinities, confirming the interaction energies predicted in this study. Additionally, using cell-based systems to observe viral entry and infectivity would help link the computational data to real-world biological outcomes. This combined approach would strengthen the validity of the results and bridge the gap between theoretical predictions and experimental reality.

Expanding the scope of the study to include other receptors and co-receptors involved in SARS-CoV-2 entry is another valuable direction. While ACE2 and GRP78 are significant, emerging research suggests that other proteins, such as Neuropilin-1 and CD147, might also facilitate viral entry. Investigating these additional receptors could provide a more comprehensive understanding of the viral-host interaction network and identify alternative therapeutic targets. Furthermore, as new SARS-CoV-2 variants continue to emerge, it is crucial to apply this study's framework to analyze these new strains. By comparing their interaction energies with those of previously studied variants, researchers can track evolutionary changes in binding characteristics that may influence transmissibility and pathogenicity.

Another promising direction is to explore the influence of environmental and biological factors on interaction energies. For example, investigating how changes in pH, temperature, or glycosylation patterns affect the binding affinities could provide more realistic models of viral entry. Additionally, studying these interactions in the context of membrane microenvironments or in the presence of other host proteins would better reflect the complexity of cellular systems. To enhance the statistical robustness of future analyses, more advanced models, such as mixed-effects models or machine learning algorithms, could be employed to account for non-linear interactions and potential confounding factors.

Finally, extending this research to investigate therapeutic interventions is a logical next step. By identifying energy measures that significantly differentiate between variants and receptors, future studies could design small molecules, peptides, or antibodies to disrupt these interactions. This could pave the way for novel antiviral therapies that specifically target variant-specific binding mechanisms. In conclusion, while this study provides a strong foundation, pursuing these future research directions will lead to a more comprehensive understanding of SARS-CoV-2 variant interactions and potentially inform the development of effective therapeutic and preventive strategies.

Chapter 6

6.1 Conclusion

This study aimed to explore the interaction energies between different SARS-CoV-2 variants and the host receptors ACE2 and GRP78, and to assess how these interactions could be differentiated using statistical measures of energy. By applying both ANOVA and Bartlett tests, the research highlighted significant differences in energy across variants, receptors, and their combinations, underscoring the potential of energy measures to distinguish between these groups. The findings revealed that energy, and specifically mean energy, are key differentiators in receptor and combination interactions, while the standard deviation of energy did not show significant differences in most cases.

The Bartlett test further emphasized that energy variances differ notably across variants and combinations, though receptor variance remained consistent. These results suggest that while receptors play a consistent role in interaction dynamics, the variance in energy associated with different SARS-CoV-2 variants and their combinations is a critical factor in understanding the differences in binding and infectivity. To robustly verify these results, Levene's test was used, providing additional evidence that the homogeneity of variances assumption holds, further validating the statistical findings across the groups.

To further investigate the specific group comparisons, Tukey's HSD test was applied, allowing for post-hoc analysis. This test revealed which particular variant-receptor combinations were significantly different from one another in terms of energy measures. The use of Tukey's test allowed for a more granular understanding of which exact differences were driving the overall significant results from the ANOVA, providing deeper insight into the interaction energies at play.

In conclusion, energy-based measures provide meaningful insights into the interactions between SARS-CoV-2 variants and host receptors, offering a reliable method for distinguishing between these interactions. While further experimental validation and exploration of additional factors such as post-translational modifications or other receptors are necessary, this research contributes to a deeper understanding of the molecular mechanisms that influence SARS-CoV-2 infectivity. By improving our grasp of these interactions, this study lays the foundation for potential therapeutic strategies aimed at blocking viral entry and mitigating the spread of new variants. Future research will continue to refine these findings, expanding the scope to include a broader range of receptors and variants, and incorporating experimental data to solidify the computational predictions made here.

References

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

Attiq, N., Arshad, U., Brogi, S., Shafiq, N., Imtiaz, F., Parveen, S., ... & Noor, N. (2022). Exploring the anti-SARS-CoV-2 main protease potential of FDA-approved marine drugs using integrated machine learning templates as predictive tools. *International Journal of Biological Macromolecules*, *220*, 1415-1428. https://doi.org/10.1016/j.ijbiomac.2022.09.086

Barton, M. I., MacGowan, S. A., Kutuzov, M. A., Dushek, O., Barton, G. J., & van der Merwe, P. A. (2021). Effects of common mutations in the SARS-CoV-2 spike RBD and its ligand, the human ACE2 receptor on binding affinity and kinetics. *eLife*, 10, e70658.

Bertoni, M., Kiefer, F., Biasini, M. *et al.* Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci Rep* **7,** 10480 (2017). https://doi.org/10.1038/s41598-017-09654-8

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T. G., Bertoni, M., Bordoli, L., & Schwede, T. (2014). Swiss-model: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, *42*(W1). https://doi.org/10.1093/nar/gku340

Biasini M., Schmidt T., Bienert S., Mariani V., Studer G., Haas J., Johner N., Schenk A.D., Philippsen A., Schwede T. OpenStructure: an integrated software framework for computational structural biology. *Acta Crystallogr. D, Biol. Crystallogr.* 2013; 69:701–709.

Biology for Life. (n.d.). ANOVA: Analysis of Variance. https://www.biologyforlife.com/anova.html

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346), 364–367. https://doi.org/10.1080/01621459.1974.10482955

Centers for Disease Control and Prevention. (2024). https://archive.cdc.gov/www_cdc_gov/sars/lab/images.html

Chen, J., Li, K., Zhang, Z., Li, K., & Yu, P. S. (2021). A survey on applications of artificial intelligence in fighting against COVID-19. ACM Computing Surveys (CSUR), 54(8), 1-32. https://doi.org/10.1145/3465398

Elfiky, A. A. (2021). SARS-CoV-2 RBD interaction with GRP78: Potential role in viral entry. *Journal of Chemical Information and Modeling*, 61(6), 3140-3149. https://pubs.acs.org/doi/pdf/10.1021/acs.jcim.1c00853

Gantini, T., & Christian, H. (2022). Analyze the Protein Model of the SARS-CoV-2 Virus using Data Mining Methods. https://www.scitepress.org/Papers/2021/107448/107448.pdf

Gobeil, S. M., Janowska, K., McDowell, S., Mansouri, K., Parks, R., Manne, K., ... & Henderson, R. (2021). Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity. *Science*, 373(6555), eabi6226.

Hazari, R., & Pal Chaudhuri, P. (2022). Analysis of coronavirus envelope protein with cellular automata model. International Journal of Parallel, Emergent and Distributed Systems, 37(6), 623-648. https://doi.org/10.1080/17445760.2022.2134369

Ibrahim, I. M., Abdelmalek, D. H., Elshahat, M. E., & Elfiky, A. A. (2020). COVID-19 spike-host cell receptor GRP78 binding site prediction. *Journal of Infection*, 80(5), 554-562. https://www.biorxiv.org/content/10.1101/2021.01.20.427368v1.full.pdf

Jackson, C. B., Farzan, M., Chen, B., & Choe, H. (2022). Mechanisms of SARS-CoV-2 entry into cells. *Nature Reviews Molecular Cell Biology*, 23(1), 3-20.

Kumar, S., Thambiraja, T. S., Karuppanan, K., & Subramaniam, G. (2022). Omicron and Delta variant of SARS-CoV-2: A comparative computational study of spike protein. *Journal of Medical Virology*, 94(4), 1641-1649.

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., & Wang, X. (2020). Structure of the SARS-COV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, *581*(7807), 215–220. https://doi.org/10.1038/s41586-020-2180-5

Lee, C. Y., & Chen, Y. P. P. (2021). New insights into drug repurposing for COVID-19 using deep learning. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4770-4780. https://doi.org/10.1109/TNNLS.2021.3111745

Levene, H. (1960). Robust tests for equality of variances. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 278-292). Stanford University Press.

Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., ... & Wang, Y. (2020). The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell*, 182(5), 1284-1294.

Maher, M. C., Bartha, I., Weaver, S., di Iulio, J., Ferri, E., & De Marco, A. (2022). Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *Science Translational Medicine*, 14(633), eabk3445.

Moal, I. H., & Bates, P. A. (2010). Swarmdock and the use of normal modes in protein-protein docking. *International Journal of Molecular Sciences*, *11*(10), 3623–3648. https://doi.org/10.3390/ijms11103623

Nguyen, H. L., Lan, P. D., Nissley, D. A., O'Brien, E. P., & Li, M. S. (2021). Electrostatic interactions explain the higher binding affinity of the CR3022 antibody for SARS-CoV-2 than the 4A8 spike protein. *The Journal of Physical Chemistry B*, 125(3), 736-744.

Parvathy, S. S., Subbanna, N., Rao, S., Pathinarupothi, R. K., Dipu, T. S., Moni, M., & Nair, C. V. (2023, December). Data-driven prognosis of long COVID in patients using machine learning. In AIP Conference Proceedings (Vol. 2901, No. 1). AIP Publishing. https://doi.org/10.1063/5.0178561

Qin, J., Tian, X., Liu, S., Yang, Z., Shi, D., Xu, S., & Zhang, Y. (2023). Rapid classification of SARS-CoV-2 variant strains using machine learning-based label-free SERS strategy. *Talanta*, *267*, 125080. https://doi.org/10.1016/j.talanta.2023.125080

Rampogu, S., Gajula, R. G., Lee, G., Kim, M. O., & Lee, K. W. (2021). Unravelling the therapeutic potential of marine drugs as SARS-CoV-2 inhibitors: An insight from

essential dynamics and free energy landscape. Computers in Biology and Medicine, 135, 104525. https://doi.org/10.1016/j.compbiomed.2021.104525

Rehman, A., Xing, H., Hussain, M., Hussain, A., & Gulzar, N. (2023). Emerging technologies for COVID (ET-CoV) detection and diagnosis: Recent advancements, applications, challenges, and future perspectives. Biomedical Signal Processing and Control, 83, 104642. https://doi.org/10.1016/j.bspc.2023.104642

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2011). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, *9*(2), 173–175. https://doi.org/10.1038/nmeth.1818

Ribes-Zamora, A., & Simmons, A. (2022). Reimagining an undergraduate genetics laboratory: using bioinformatic tools and 3D printing to study the COVID-19 virus in an online environment. In EDULEARN22 Proceedings (pp. 1598-1603). IATED. https://doi.org/10.21125/edulearn.2022.0423

Shoemark, D. K., Colenso, C. K., Toelzer, C., Gupta, K., Sessions, R. B., Davidson, A. D., Mulholland, A. J. (2021). Molecular simulations suggest vitamins, retinoids and steroids as ligands of the free fatty acid pocket of the SARS-CoV-2 spike protein. *Angewandte Chemie International Edition*, 60(13), 7098-7110.

Six Sigma. (n.d.). Bartlett's Test for Homogeneity of Variances. https://www.6sigma.us/six-sigma-in-focus/bartletts-test

Spinello, A., Saltalamacchia, A., & Magistrato, A. (2024). Binding free energy analysis of SARS-CoV-2 variants interacting with ACE2 receptor. *Physical Chemistry Chemical Physics*. https://pubs.rsc.org/en/content/articlelanding/2024/cp/d3cp04997c

Starr, T. N., Greaney, A. J., Hilton, S. K., Crawford, K. H. D., Navarro, M. J., Bowen, J. E., ... & Bloom, J. D. (2020). Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, 182(5), 1295-1310.

Theerthagiri, P., Jacob, I. J., Ruby, A. U., & Vamsidhar, Y. (2020). Prediction of COVID-19 possibilities using the KNN classification algorithm. https://doi.org/10.21203/rs.3.rs-70985/v2

Titus, R., Mandal, M., & Dutta, G. (2022). Electrochemical biosensor designs used for detecting SARS-CoV-2 virus: A review. Next Generation Smart Nano-Bio-Devices, 187-209. https://doi.org/10.1007/978-981-19-7107-5_10

Torchala, M., Moal, I. H., Chaleil, R. A. G., Fernandez-Recio, J., & Bates, P. A. (2013). SwarmDock: A server for flexible protein–protein docking. *Bioinformatics, 29*(6), 807–809. https://doi.org/10.1093/bioinformatics/btt038

Torun, H., Bilgin, B., Ilgu, M., Yanik, C., Batur, N., Celik, S., ... & Onbasli, M. C. (2021). Machine learning detects SARS-CoV-2 and variants rapidly on DNA aptamer metasurfaces. *MedRxiv*, 2021-08. https://doi.org/10.34133/adi.0008

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2), 99–114. https://doi.org/10.2307/3001913

Uniprot: The Universal Protein Knowledgebase. (2016). *Nucleic Acids Research*, *45*(D1). https://doi.org/10.1093/nar/gkw1099

Vangipuram, S. K., & Appusamy, R. (2021, April). Machine learning framework for COVID-19 diagnosis. In International Conference on Data Science, E-Learning and Information Systems 2021 (pp. 18-25). https://doi.org/10.1145/3460620.3460624

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). Swiss-

model: Homology modeling of protein structures and complexes. *Nucleic Acids Research*, *46*(W1). https://doi.org/10.1093/nar/gky427

Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., & Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science*, 367(6485), 1444-1448.