

**COMPARATIVE ANALYSIS OF ALPHAFOLD2 &
ESMFOLD ALGORITHMS TO SUPPORT
BIOSIMILARITY & DRUG DISCOVERY**



BY

Zamara Mariam

Fall-2021-MSBI-00000364054

Supervised by
Dr. Rehan Zafar Paracha

Co-Supervisor
Dr. Sarfaraz Khan Niazi

in

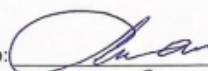
MS Bioinformatics

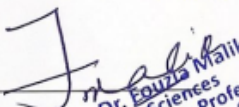
June 2023

**School of Interdisciplinary Engineering & Science (SINES)
National University of Sciences & Technology (NUST)**

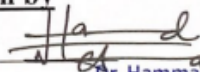
THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Mr/Ms Zamara Mariam Registration No. 00000364054 of SINES has been vetted by undersigned, found complete in all aspects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature with stamp: 
Name of Supervisor: Dr. Renar Zafar
Date: 03/07/23 Associate Professor
SINES - NUST, Sector H-12
Islamabad

Signature of HoD with stamp: 
Date: 11-7-2022 Dr. Fouzia Malik
HoD Sciences
Associate Professor
SINES - NUST, Sector H-12
Islamabad

Countersign by

Signature (Dean/Principal): 
Date: 11.1 JUL 2023 Dr. Hammat M. Cheema
Principal & Dean
SINES - NUST, Sector H-12
Islamabad

*I would like to dedicate my thesis to my beloved parents, my mentors,
my brothers and my friends, who have played instrumental roles in my
journey with their unwavering belief and support.*

DECLARATION

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes the outcome of the work done.

Zamara Mariam

June 2023

Acknowledgement

All praise for **Almighty ALLAH** Who is the ultimate source of all knowledge. It is through the boundless grace and blessings of Allah that I have attained this current level of knowledge. I humbly extend my profound respect to the Holy Prophet Hazrat Muhammad (PBUH); his teachings and wisdom continue to illuminate our path and enrich our understanding.

I earnestly thank my supervisor Dr.Rehan Zafar Paracha, and my co-supervisor Dr. Sarafarz Niazi for their keen interest, unwavering guidance, encouragement and wisdom that illuminated my path. I am grateful for their thought provoking discussions, valuable suggestions, and dedication to my growth that have been a constant source of inspiration. I am truly honored to have been mentored by both of them, and I dedicate this work to them as a token of my utmost respect and gratitude. I am thankful to my GEC committee members, Dr. Zartasha Mustansar and Dr. Uzma Habib for their valuable suggestions and concise comments. Their contributions have enhanced the quality and credibility of this thesis, and I am honored to have benefited from their wisdom and knowledge.

No words can express and no deeds can return the support and inspiration that my parents and brothers permeated in me during the course of my research work. This work stands as a testament to the boundless love and sacrifices of my mother and her unwavering dedication that has nurtured and propelled me forward, giving me the strength to pursue my dreams. My dear brothers, Muneeb, Faris, and Talha, have been a constant source of companionship, trust, and encouragement throughout this journey and served as my confidants and cheerleaders. To all those who have contributed to my personal and academic growth, I am deeply grateful for your influence. May the respect and dedication embedded within these words resonate with each of you, for without your unwavering support, this journey would not have been possible.

Contents

Contents	v
List of Tables	viii
List of Figures	x
ABSTARCT	1
1 INTRODUCTION	3
1.1 Therapeutic proteins	3
1.2 Biosimilars	5
1.3 Unraveling the complexity of Protein Folding	6
1.4 Protein Structure Prediction	9
1.4.1 Shifting the Paradigm: From Experimental Structure Prediction to In-silico Computer-Based Approaches	10
1.4.2 Structure prediction in Machine Learning era	12
1.5 Orthogonality: AlphaFold2 and ESMFold	15
1.5.1 Predicted Local Distance Difference Test	16
1.5.2 Predicted Template Modeling	17
1.5.3 Predicted Aligned Error	17
2 LITERATURE REVIEW	19
2.1 Structure Prediction - AlphaFold2	19
2.2 Impact of Mutations on proteins stricture	21
2.3 AlphaFold2 scores interpretation	22
2.4 Study Rationale	24
2.5 Objectives	24
3 MATERIALS AND METHODS	26
3.1 Data Collection	26
3.2 Structure prediction tools and scores	27
3.3 Physicochemical properties	28
3.4 Protein-target interaction	29
3.5 Protein domains & pTM score	30
3.6 Investigating relationship between Predictions, Structural and Sequential Data	31
3.7 Randomization & Mutation	32
3.8 Affinity Maturation: Trastuzumab	33
3.8.1 Trastuzumab & Tyrosine-protein kinase erbB-2 (HER2)	33
3.8.2 Alanine Scanning	33
3.8.3 mCSM-PPI2	33
3.8.4 Trastuzumab & HER2 affinity maturation	34

4	RESULTS	35
4.1	Sequence Length vs Molecular Weight	35
4.2	Proteins & Peptides Correlation	35
4.3	Orthogonal comparison between AF2 vs. ESMF	37
4.4	pLDDT & pTM Rank ordering Biosimilars	37
4.5	Physicochemical Attributes	39
4.6	Proteins Interactions: Effects of Structural folds	47
4.7	TrRosetta: Domains-based analysis	53
4.8	Analyzing the Learning of AF2 & ESMF based on available data	55
4.9	Randomization & Mutation: Novel molecules predictions	57
4.10	Trastuzumab: Alanine scanning results	58
5	DISCUSSION	61
6	CONCLUSION AND FUTURE PERSPECTIVES	65
	REFERENCES	68
	Appendix A Proteins and Peptides Complete Data - Github URL	77
	A.1 Proteins Data	77
	A.2 Peptides Data	77
	A.3 Alanine Scanning Results	78
	Appendix B Biosimilars - Rank order	79
	Appendix C Physicochemical Attributes computation Source Code	85

Nomenclature

Acronyms / Abbreviations

ΔG	Gibbs Free Energy
AA	Amino Acids
AF2	AlphaFold2
BLA	Biologics License Application
CAPRI	Critical Assessment of PRedicted Interactions
CASP	Critical Assessment of Structure Prediction
ESMF	ESMFold
FDA	The Food and Drug Administration
GRAVY	Grand Average of Hydropathicity Index
ICs	Interfacial Contacts
IDPRS	Intrinsically Disordered Protein Regions
mAbs	Monoclonal Antibodies
MSA	Multiple Sequence Alignment
MW	Molecular Weight
NCBI	National Center for Biotechnology Information
NDA	New Drug Application
NIS	Non-Interacting Surfaces
NMR	Nuclear Magnetic Resonance
PAE	Predicted aligned error
PDB	Protein Data Bank
pLDDT	Predicted Local Distance Difference Test
pTM	Predicted Template Modelling
TM	Transmembrane

List of Tables

1.1	Biosimilars of Epoetin (EPO)	6
1.2	Structure prediction methods are the respective tools that can be employed for the prediction of structure	11
4.1	Confidence scores of approved biosimilars using AF2 and ESMF	39
4.2	Correlations between structural and physicochemical parameters of 188 therapeutic proteins with strong correlations (>0.5)	41
4.3	Correlations between structural and physicochemical parameters of 16 therapeutic peptides with strong correlations (>0.5)	42
4.4	Average extinction coefficients (reduced cystine) for all types of molecules with the average amino acid number and molecular weight	45
4.5	Interacting PTH-PTHr1 residues pLDDT from AF2 and ESMF, although they have different pLDDT scores but produced similar binding interactions and scores	48
4.6	AF and ESM predicted cytokines, hormones and fusion proteins binding affinity values calculated from PRODIGY server	49
4.7	PRODIGY results of AF2 and ESMF predicted structures docked with their respective targets	51
4.8	Single and multi-domain molecules with low pTM (<0.5) predicted from AF2 and ESMF	53
4.9	Single and multi-domain molecules with low pTM scores predicted from AF2, ESMF, and trRosetta	55
4.10	BLAST PDB and UniProt compared with AF2 pLDDT scores	57
4.11	AF2 and ESMF prediction scores comparison for mutated single and multiple domains	58
4.12	Alanine Scanning results for Trastuzumab-HER2	60

B.1 Rank order from pLDDT score AF2	79
B.2 Rank order from pTM score AF2	82

List of Figures

1.1	Proteins level of folding	9
1.2	PAE plot by AF2	18
4.1	pLDDT plots of proteins from AF2 and ESMF	36
4.2	pTm plots of proteins from AF2 and ESMF	36
4.3	pLDDT plots of peptides from AF2 and ESMF	36
4.4	pTM plots of peptides from AF2 and ESMF	37
4.5	Orthogonal comparison of AF2 and ESMF	38
4.6	Misfolding of PTH from AF2 - comparison	46
4.7	Alanine Scanning Affinity Chart	59

ABSTRACT

The elucidation of three-dimensional protein structure plays a pivotal role in comprehending biological phenomena. It directly governs protein function and hence aids in drug discovery. Development of protein prediction algorithms, AlphaFold2 and ESMFold, have the potential to shift the paradigm of protein-based therapeutic discovery. Turning an amino acid chain into 3D domains and docking them can aid in unlocking a protein's full potential. Besides this, the effects of mutations on the domain structure can be studied meticulously. Prediction scores from extensive studies were examined in the hope of searching for newer modalities of transforming protein therapeutics. Most of these studies failed to find any utility of these algorithms, and a few suggested, despite their dismal findings, that their utility can be found. The inventors of the algorithms cautioned that the predicted structures and scores have no utility except regurgitate known structures from the known structure databases. A few possible applications, as considered in this study, are to predict pre-translation variations, mutations, and structural changes. A potential correlation of repeatedly manufactured batches of therapeutic protein is correlated with the structure prediction score as a measure of thermodynamic instability. 204 unmodified FDA-approved therapeutic proteins were correlated with their prediction scores and available physicochemical and functional properties. Slight residual differences among the commercial therapeutic proteins and structures reported in the PDB were found. The potential impact of mutations on the prediction scores is also studied. No observed correlation was found between the prediction score and any tested attribute. The algorithms exhibited lower confidence in predicting structures for sequences with low identity scores when tested against the UniProt and PDB databases. Other deployed algorithms (i.e., trRoseeta) were concluded to be more relevant to domain manipulation as well. Reliable structure prediction from these algorithms highly depends on the model's architecture and training data. Ultimately, it was concluded that none of these algorithms have any

value except they show how good they can be at reproducing a known or partially-known structure. The comparison of AF2 and ESMF resulted in R^2 of 0.69, vouching for their orthogonality. However, the R^2 value of physiochemical attributes was as low as 0.07. Lack of significant correlation of predictability scores with physicochemical and functional properties cannot vouch for in-vivo stability and molecular functionality of a protein. Furthermore, when novel randomized and mutated sequences are provided to these algorithms, they fail to predict structures with acceptable accuracy. This is majorly due to the unavailability of similar folds in the training dataset (i.e., UniProt and PDB) of these algorithms. Although it might seem that these algorithms go beyond regurgitating available data, it might not be the case. In this context, these algorithms are considered no different than GPT4, which also regurgitates available data. These algorithms do not play well in proving the Levinthal paradox as solved, yet it remains unsolved.

INTRODUCTION

Biological products (biologics) encompass a broad range of therapeutics, including vaccines, blood components, allergenics, somatic cells, gene therapy, tissues, and recombinant therapeutic proteins. Biologics can consist of sugars, proteins, nucleic acids, or their intricate combinations. This diverse class of compounds has revolutionized medicine, offering new and effective treatment options for a wide range of diseases and conditions. These remarkable biopharmaceuticals, derivatives of naturally occurring human proteins, may range from small peptides like insulin to larger proteins like monoclonal antibodies. Before the advent of recombinant technology, biologics were derived from diverse natural sources, including humans, animals, and microorganisms. The boom of recombinant DNA technology has drastically improved the selective binding of target-enabled therapeutic proteins to interrupt disease progression, enhance immune responses, modulate cell signaling, or replace deficient proteins, leading to improved clinical outcomes. These recombinant proteins exhibit specific molecular structures and functions, making them highly versatile and potent therapeutic agents. The therapeutic potential of proteins lies in their folds, functional regions, and the ability of these functional regions to interact with specific molecular targets involved in disease pathways. Moreover, these proteins' high specificity and affinity allow for precise targeting, reducing potential side effects and improving patient safety, which can be studied through various in-silicon and in-vivo techniques.

1.1 Therapeutic proteins

Therapeutic proteins are typically large, complex molecules designed to mimic or enhance the natural functions of endogenous proteins within the human body. They

are produced through recombinant DNA technology, involving the genetic engineering of cells such as bacteria, yeast, or mammalian cells to express and produce the desired protein. This advanced manufacturing process enables the production of therapeutic proteins on a large scale, ensuring consistent quality and purity.

Recombinant forms of naturally occurring proteins have significantly contributed to the treatment of numerous diseases, including cancer, diabetes, autoimmune disorders, infectious diseases, and genetic disorders. They have demonstrated remarkable efficacy, often outperforming traditional small-molecule drugs in terms of specificity and therapeutic impact. Additionally, the advent of personalized medicine has further expanded the applications of therapeutic proteins, as they can be tailored to target specific genetic mutations or disease sub-types.

Protein-based therapeutics are highly successful with great potential. They can be categorized into groups based on their molecular composition and pharmacological activity. The molecular categorization is based on the type of class the molecule belongs to, i.e., enzymes, growth factors, hormones, interferons, interleukins, thrombolytics, Fc-fusion proteins, anticoagulants, and blood factors. These compounds are further categorized according to their molecular modes of action into groups that attach non-covalently to the target, such as monoclonal antibodies (mAbs); affect covalent bonds, such as enzymes; and exert activity without specific contacts, such as serum albumin. Each class of therapeutic protein exhibits unique characteristics that make them suitable for treating specific diseases i.e., treatment of infections, cancers, immunological disorders, and other diseases. Based on their pharmacological activity they can be divided into five sub-classes i.e., replacing an absent or abnormal protein; enhancing an already-existing pathway; offering a novel function or activity; interfering with a molecule or organism; and delivering other compounds or proteins, such as a radionuclide, cytotoxic drug, or effector proteins (1).

Each class of compounds has played a significant role in the treatment of various disease conditions. For instance, monoclonal antibodies have revolutionized the treat-

ment of various cancers and autoimmune diseases by selectively targeting specific molecules on the surface of cells. Trastuzumab (Herceptin) is an Anti-HER2 antibody that is used in combination with standard adjuvant chemotherapy for breast cancer treatment and has shown to significantly prolonged survival in high-risk patients (2). Similarly, Obinutuzumab is used in the treatment of patients with previously untreated chronic lymphocytic leukemia, while Idarucizumab is used when the reversal of an anticoagulant is needed during urgent surgical procedures in order to control uncontrolled bleeding. A recombinant enzyme, Sebelipase alfa is used in the treatment of patients with lysosomal acid lipase deficiency (3). At the same time, growth factors are crucial in stimulating tissue repair and regeneration. One of the best-therapeutically-characterized growth factors, the Heparin-binding-Epidermal Growth factor, plays a vital role in wound healing. It binds to the EGFR sub-types HER1 and HER472, promoting 're-epithelialization'. Besides wound healing, it has a critical role in angiogenesis, cardiac valve tissue development, and the maintenance of normal heart function (4).

1.2 Biosimilars

Biosimilars, also known as follow-on biologics, are copies of therapeutic proteins approved as having "no clinically meaningful difference" when compared to their reference-approved therapeutic proteins for which a full regulatory filing has already been approved. Extensive analytical, toxicological, clinical pharmacy, and clinical efficacy comparisons are conducted with their reference products to ensure that the innate variability of protein structure that determines efficacy and toxicity from batch-to-batch is highly similar to the reference product. By establishing the similarity and interchangeability of biosimilars with their reference products, regulatory agencies ensure that patients can confidently rely on these alternative treatment options. Since biosimilars are developed to have similar properties and effects as their reference products, they can be utilized for the same therapeutic purposes. This offers

Table 1.1. Biosimilars of Epoetin (EPO)

Country	Company	Product
Apotex	Canada	Apo-EPO
Biocon	India	Erypro Safe
Biosidus	Argentina	Zyrop
Hospira (Pfizer)	USA	Retacrit (epoetin zeta)
Hexal	Australia	Epoetin alfa Hexal

opportunities for expanding patient access to essential treatments while potentially reducing healthcare costs.

Patients with chronic renal failure develop anemia which is caused by delayed production of erythropoietin (EPO) by the peritubular cells of the kidney. This condition is mainly treated by EPO containing medicinal products like Epoetin-alpha and its biosimilars as shown in Table ???. A biosimilar of epoetin-alpha, called epoetin-zeta, was granted marketing authorization by the European Medicines Agency in 2007. Structurally, epoetin-zeta has higher levels of N-glycans with lactose-amine extensions and lower levels of N-glycan sialylation relative to its reference product which indicates that their structures are slightly different making it a biosimilar (5). Economically, a global epoetin-zeta cost analysis report showed a total cost saving of nearly 45% in comparison to epoetin alfa, for Spain and it showed that the availability of epoetin-zeta decreased the cost by £7.9 per week, for patients in UK (6).

1.3 Unraveling the complexity of Protein Folding

Therapeutic proteins and their biosimilars have emerged as indispensable tools in modern medicine with their remarkable ability to modulate biological processes and target specific disease-related molecules. However, the efficacy and safety of these proteins heavily depend on their three-dimensional structure, which governs their functional properties. Understanding and predicting protein structure is of paramount

importance as it allows researchers and scientists to gain insights into their behavior, interactions, and potential therapeutic applications. Accurate structure prediction is a cornerstone in drug design, rational engineering of proteins, and developing novel therapeutics. By unraveling the complexity of protein folding, a treasure trove of knowledge with immense potential for improving human health and addressing some of the most challenging medical conditions can be unlocked.

Protein folding process can be conceptually divided into four levels: primary, secondary, tertiary, and quaternary structure Figure 1.1. The primary structure refers to the linear sequence of amino acids that make up the polypeptide chain using 19 different amino acids in combinations. This sequence determines the protein's folding pattern and ultimately its function. The secondary structure involves the folding of the peptide chain into regular patterns such as alpha helices or beta sheets, stabilized by hydrogen bonds. Tertiary structure describes the overall three-dimensional arrangement of the protein, including the spatial orientation of secondary structure elements. Finally, multiple polypeptide chains interact together to form the quaternary structure. Furthermore, multiple combinations of amino acids and polypeptide chains fold to form different conformations of proteins due to many degrees of freedoms.

The remarkable complexity of proteins was recognized by Cyrus Levinthal in 1969. For instance, considering three states for each bond and examining a 101-residue sequence with 100 covalent bonds (peptide linkages) and 199 distinguishable phi & psi bond angles, theoretically, a protein has $3^{100} = 5 \times 10^{47}$ potential conformations. To explore all these possibilities at a protein sampling rate of 3×10^{20} per year, it would take approximately 10^{27} years, emphasizing the immense challenge of exhaustively testing every conformation (7), (8), (9). The innate programming of proteins coded in their genetic material enables them to fold into their native states with minimum energy and maximum stability with a minute window of error, in a fraction of a second. This significant flexibility and precise folding of therapeutic proteins is essential for their stability, activity, and interaction with disease targets.

Certain Pre-translational modifications and post-translational modifications play crucial roles in shaping the structure of proteins as well. Pre-translational modifications occur during protein synthesis and involve modifications to the nascent polypeptide chain before it is fully synthesized. These modifications include signal peptide cleavage, where a signal sequence is removed to guide the protein to its correct cellular location, and the addition of certain amino acids or protein tags that facilitate protein folding and stability. Some examples of such pre-translational modifications are splicing, capping, addition of a poly-A tail, histone modification, DNA methylation, and transcript modification.

Post-translational modifications (PTMs) occur after the protein has been fully synthesized and can profoundly impact protein structure and function. Common PTMs include phosphorylation, glycosylation, acetylation, methylation, ubiquitination, and proteolytic cleavage. These modifications can alter protein conformation, stability, enzymatic activity, subcellular localization, and interaction with other molecules. In glycosylation, for instance, the attached sugar molecules can influence protein folding and stability, as well as mediate protein-protein interactions and cell recognition processes. Besides this, methylation and acetylation can modify the charges and hydrophobicity of amino acids, leading to changes in protein structure and function. Additionally, PTMs can create binding sites for other proteins or signaling molecules, allowing for intricate regulatory networks within cells. The interplay between different PTMs can further regulate protein structure and function, adding an additional layer of complexity.

Overall, both pre-translational and post-translational modifications are essential for proteins' 3D fold formation and their therapeutic function. They expand the functional repertoire of a protein, regulate its activity, and enable it to participate in various cellular processes. By elucidating the folding pathways and factors influencing protein folding, therapeutic proteins with enhanced efficacy, reduced side effects, and improved pharmacokinetics (PK), pharmacodynamics (PD), bioavailability can be

studied, ultimately advancing the development of novel and more effective treatments for various diseases. Hence understanding and predicting these different levels of protein folding and the effects of pre-and post-translational modifications is crucial.

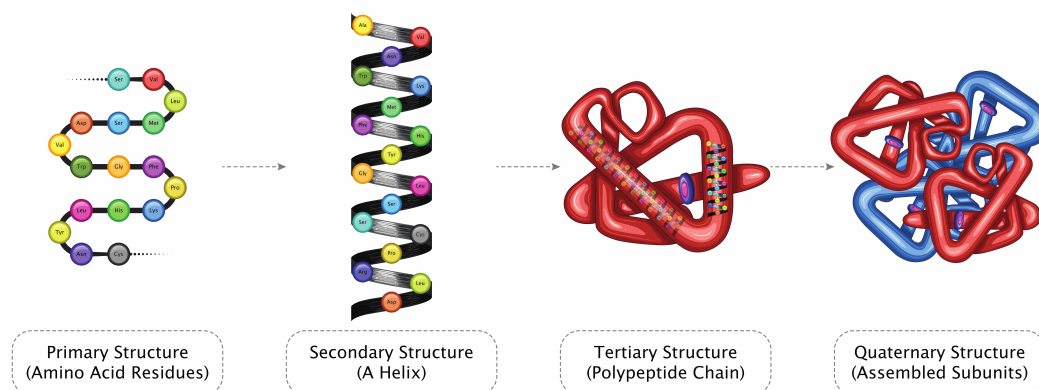


Figure 1.1. The four levels of protein folding: primary, secondary, tertiary, and quaternary (10)

1.4 Protein Structure Prediction

The primary amino acid chain is the only determinant of the 3D structure of a protein-based therapeutic (for both biosimilar and its reference product), thus the amino acid side chains are critical; charged amino acid sides can form ionic bonds, and polar amino acids can form hydrogen bonds. Weak Van der Waals interactions mediate interactions between hydrophobic side chains. These side chains primarily form non-covalent bonds. Cysteines are the only amino acids that have the ability to form covalent bonds, and they do so by utilizing their side chains. The arrangement of the amino acids of a given protein depends on side-chain interactions. Thousands of non-covalent bonds between amino acids stabilize folded proteins. A faithful translation of the genetic code depends on several sequential molecular recognition events, each with an inherent error rate. The overall error rate of protein synthesis has been estimated at one misincorporated amino acid per 10⁴ codons. It reflects accumulated mistakes from all steps involved in translation. These error rates are dependent on

the thermodynamic stability of the amino acid chain that can be projected through experimental methods and computational methods of protein structure identification.

1.4.1 Shifting the Paradigm: From Experimental Structure Prediction to In-silico Computer-Based Approaches

The experimental structure prediction of proteins has long been an empirical approach in providing invaluable insights into protein structure-function relationships, paving the way for understanding molecular mechanisms and guiding drug discovery efforts. Through meticulous laboratory techniques and advanced instrumentation, scientists have strived to decipher the three-dimensional arrangement of proteins, unveiling their intricate folding patterns and functional architectures. The techniques employed include X-ray crystallography, Nuclear Magnetic Resonance (NMR), Cryo-electron microscopy (Cryo-EM), and Circular dichroism spectroscopy, among others. The inherent variability of these methods makes them reliable depending upon samples' quality, equipment's accuracy, and results reproducibility. These experimentally identified structures can be easily retrieved from UniProt and RCSB Protein Databank (PDB) databases. The PDB database has over 174,825 experimentally generated structures available as of 2023 (11). Certain novel proteins and therapeutics have unique structures that can be predicted using computational methods like I-TESSER, SWISS-MODEL, MODELLER, Rosetta, Phyre2, etc. which are template-based homology modeling, protein threading, and ab initio approaches. While prediction methods for protein structure exhibit substantial variations in their specific procedures, there are fundamental steps that remain consistent across different approaches. These steps typically involve the selection of templates, the reconstruction of the structure, the refinement of the predictions, and subsequent analysis. A range of structure prediction tools can be found in the provided in Table [1.2](#)

Table 1.2. Structure prediction methods are the respective tools that can be employed for the prediction of structure

Methods and Models	Programs and Tools
Homology Modeling/Comparative Modeling: Create a 3D model of the target protein using a homologous protein's empirically confirmed structure as a guide.	MODELLER, SWISS-MODEL, Phyre2, RaptorX, I-TASSER
Ab Initio Modeling: Build a 3D model of the target protein by sampling the protein's conformational space without using any experimental data.	Rosetta, QUARK, AlphaFold, ESMFold, PCONS5
Threading: Build a 3D model of the target protein by aligning the protein sequence with the sequences of proteins of known structure.	MUSTER, 3D-PSSM, LOMETS, HHpred
Hybrid Modeling: Combine two or more modeling approaches to improve the accuracy of the predicted structure.	CABS-flex, PrimeX, GalaxyHomomer
Knowledge-based methods: Use existing knowledge about protein structure and function to predict the structure of the target protein.	ProSMoS, ProQ3D, I-TASSER-2GO
Template-free methods: Build a 3D model of the target protein without using templates or homologous proteins.	CONFOLD2, MetaPSICOV, trRosetta
Fragment-assembly methods: Build a 3D model of the target protein by assembling fragments of known protein structures.	PEP-FOLD3, Robetta, QUARK

1.4.2 Structure prediction in Machine Learning era

The accuracy of protein structure prediction algorithms has improved with significant advances in Machine Learning (ML) and Artificial Intelligence (AI). The template-free AI models utilized in this context are trained using sequence and 3D structural data extracted from publicly accessible databases i.e., UniProt (12), RCSB PDB, Uniclust (13), and MGNify (14) etc. Independent of templates, highly precise protein structure prediction tools include AlphaFold2 (15), trRosetta (16), Robetta (17), RoseTTA Fold (18), ESMFold (19), RaptorX (20) and OmegaFold (21). To predict protein structures from amino acid sequences, each one employs a distinct AI model and algorithm.

1- AlphaFold2 (2021):

AlphaFold2 is a CASP14 winner with 90% accuracy of structure prediction. In comparison with AlphaFold1, AF2 has replaced the convolutional neural network with an attention-based architecture, removing the rigid information flow from the local neighbors of the convolutional networks with a flow dynamically controlled by the network. This has increased its accuracy many folds. Its predictions have been validated experimentally and it was determined that the results have ended up being similar to predicted structures despite their challenging nature and having very few related sequences.

Since AF2 depends upon Multiple Sequence Alignment, it is limited by the availability of current knowledge, data, and experimentally derived structures present in the databases, i.e., PDB. Another limitation is that since AF2 was trained on PDB, which may not have structures of proteins in their natural fold states, i.e., some of the PDB structures were documented in the presence of other proteins during the solving of the fold. This limitation is most clearly observable for proteins with multiple native structures. Furthermore, the implementation of AF2 requires extensive data resources to download the databases before actual structure prediction.

2- ESMFold (2022):

ESMFold is a protein structure prediction model which uses transformer models to encode protein sequences 60 times faster than AF2, eliminating MSA while maintaining high-quality predictions using as many as 15 billion parameters. Its biggest advantage is its ability to predict structures many times faster than any other tools available making it excellent for identifying remote homology and conservation in a large collection of novel sequences. ESMFold generates structure predictions using only one sequence as input by leveraging the internal representations of the language model, ESMFold producing more accurate atomic-level predictions than AlphaFold2 or RoseTTAFold. However, similar to AlphaFold2, it is limited by the training data that required significant computational resources to run it, which can limit its accessibility.

3- RoseTTAFold (2021):

RoseTTAFold is a "three-track" neural network, meaning it simultaneously takes into account potential three-dimensional protein structure, interactions between the amino acids in a protein, and patterns in protein sequences. This architecture enables the network to collectively reason about the relationship between a protein's chemical components and its folded structure by exchanging one-, two-, and three-dimensional information. AlphaFold2-like precision and constraints apply to it.

4- trRosetta (2021):

A web-based platform for quick and precise protein structure prediction, trRosetta (transform-restrained Rosetta) is powered by deep learning and Rosetta. A deep neural network is initially used to predict the inter-residue geometries, including distance and orientations, using the input of a protein's amino acid sequence. In the context of Rosetta, the predicted geometries are subsequently turned into restrictions to steer the structure prediction based on direct energy minimization. It has good prediction accuracy but is only applicable to monomer models.

5- OmegaFold (2022):

In order to learn single- and pairwise-residue representations as effective features

that model the distribution of sequences, OmegaFold uses a deep transformer-based protein language model that was trained on a sizable collection of unaligned and unlabeled protein sequences. The 3D coordinates of each heavy atom are lastly predicted by a structural module. The OmegaFold protein manufacturing model is 'super-fast' since it is independent of evolutionary data. It performed better on single-sequence inputs than AlphaFold2 and RoseTTAFold. Additionally, OmegaFold outperformed AlphaFold 2 in terms of statistical prediction accuracy, most likely as a result of the benefits of its single-sequence-based prediction strategy for both antibody loops and orphan proteins. On the CAMEO dataset, OmegaFold structures were as accurate as RoseTTAFold structures (0.75 mean LDDT score) and RoseTTAFold structures with a mean local-distance difference test (LDDT) score of 0.82.

6- RaptorX (2016):

A web-based program called RaptorX makes predictions about structures and characteristics based on protein sequences. This service makes use of a powerful internal deep learning model named DeepCNF (Deep Convolutional Neural Fields) to forecast secondary structure (SS), solvent accessibility (ACC), and disorder regions (DISO). The complicated hierarchical structure shows the dependency between nearby property labels in addition to the complex interaction between sequence and structure. The experimental findings demonstrated that this server can achieve 84 percent Q3 accuracy for 3-state SS, 72 percent Q8 accuracy for 8-state SS, 66 percent Q3 accuracy for 3-state solvent accessibility, and 0.89 area under the ROC curve (AUC) for disorder prediction when evaluated on CASP10, CASP11, and the other benchmarks. However, the structure is affected by the availability of sparse evolution data (22).

Conclusion:

Although ML-based tools predict structures with high accuracy, they do not predict several important aspects of these protein structures i.e., metal ions, cofactors, and other ligands. Post-transnational modifications, such as glycosylation or phosphorylation, and complexes conjugated with DNA, RNA, and their complexes, are also not

accounted for during these predictions. In addition, amino acid side chains are not always accurately placed. Each of these listed features may be crucial for protein function, and many of these are necessary for the integrity of the folds within the proteins.

1.5 Orthogonality: AlphaFold2 and ESMFold

The Levinthal paradox was better understood when it was shown that a slight shift in the free energy in the thermodynamic profile of the amino acid chain could explain the repeatability of the 3D structure. However, the structure variability upon protein translation remains due to pre-translation modifications triggered primarily by the thermodynamic instability of the amino acid chain. Predicting the 3D structure and these instabilities using AI-based models has been challenging until AF2 presented its ability to provide higher than 90% confidence upon repeated prediction, simulating repeated protein translation. AF2 employs Evoformer, a neural network architecture that combines elements of both evolutionary and transformer-based models. Evoformer architecture incorporates an attention mechanism inspired by transformers, enabling the networking order to record distant interactions between amino acids in a protein sequence. It uses an attention matrix to weigh the importance of different pairwise interactions, allowing for a more accurate prediction of protein structures. AlphaFold's training process involves two key components: multiple sequence alignments (MSA) and the prediction of distance maps. MSA is derived from diverse sets of evolutionarily related protein sequences, providing valuable information about co-evolving residues. This MSA data is used to train the model and generate sequence profiles that capture the evolutionary conservation of amino acids and come directly from a publicly accessible database i.e., UniProt, PDB, etc. By training the network to predict the distances between pairs of residues in a protein sequence, AF2 infers the 3D spatial arrangement of the protein which it achieves by incorporating a combination of convolutional layers, residual connections, and attention mechanisms within

the Evoformer architecture. Further optimization is guided by a scoring function that considers various physical and geometric properties of proteins. This refinement stage helps improve the accuracy of the final predicted structures.

Similar to AF2, ESMF is a cutting-edge protein structure prediction method that harnesses evolutionary sequence co-variation analysis and large-scale language models. By examining the co-evolutionary patterns among amino acid residues in a protein family in order to capture valuable information about residue interactions and structural constraints, ESMF makes accurate predictions. The strength of ESMF lies in its ability to integrate diverse sources of information and incorporation of predicted secondary structure information and contact maps. These additional inputs provide valuable insights into local structural elements and the spatial proximity of amino acids, further refining the accuracy of the predicted protein structures. Both AF2 and ESMF structural module generates model confidence predictions which are displayed as predicted local distance difference test (pLDDT) scores that range from 0 to 100 and predicted Template Modeling (pTM) scores that range from 0 to 1. A cross-comparison serving as an orthogonal confirmation shows that the AF2 and ESMF algorithm predicts the structure of protein domains with an accuracy matching that of experimental methods.

1.5.1 Predicted Local Distance Difference Test

The predicted Local Distance Difference Test (pLDDT) is a measure of confidence or reliability assigned to each residue in the predicted protein structure. It represents the predicted accuracy of the local distance difference, which is the difference between the predicted distances and the true distances in the experimentally determined protein structure. The pLDDT score ranges from 0 to 100, with higher scores indicating higher confidence in the predicted local structure. Regions with high pLDDT scores (e.g., > 80) are considered to have accurate predictions, while lower scores (e.g., <50) indicate regions where the predictions may be less reliable. Lower

pLDDT scores may also indicate that the fold is in intrinsically disordered protein regions (IDPRs).

1.5.2 Predicted Template Modeling

The predicted Template Modeling, or TM-scores (pTM), on the other hand, is the global metric of structure assessment and evaluates the overall quality of the predicted protein structure by comparing it to experimentally determined structures of similar proteins available in the Protein Data Bank (PDB). The pTM score assesses how well the predicted structure aligns with the known structure of a related protein template. It ranges from 0 to 1 and a higher pTM score signifies a better alignment and a higher likelihood of the predicted structure being accurate.

1.5.3 Predicted Aligned Error

Predicted Aligned Error (PAE) is a metric used in the evaluation of protein domain structure predictions generated by AF2 and ESMF. PAE measures the average deviation between the predicted and experimental positions of aligned residues in a protein structure. The predicted structure is first aligned with the experimental or reference structure followed by the computation of deviation or error for each aligned residue by measuring the distance between the predicted and experimental positions. PAE is usually shown as a heatmap (Figure 1.2) with residue numbers running along the axis and color at each pixel indicating the PAE value for the corresponding pair of residues. If the relative position of two domains is confidently predicted then the PAE values will be low (less than 5Å) for pairs of residues with one residue in each domain.

High pLDDT demonstrates strong confidence in the residue structure. Residues that have low pLDDT scores The global superposition TM-score is an additional beneficial statistic to qualify or quantify the structure prediction. AF2 calculates the predicted template modeling, or TM-scores (pTM), based on a pairwise error pre-

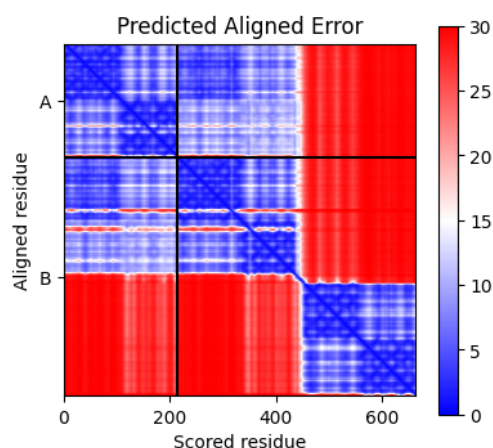


Figure 1.2. PAE plot of Trastuzumab (Herceptin) generated by AlphaFold2.

diction, the predicted aligned error (PAE), which calculates the error of each amino acid's location, is then calculated. It has been demonstrated that the protein sequence's MSA depth has a significant impact on prediction accuracy. It was also determined that quick and accurate protein structure predictions can be made using just MSA. Since interactions, functions, and the effects of missense variation depend on the 3D structure of proteins, the excellent predictability of the 3D structure created by the AF2 and ESMF suggests that testing biosimilars to determine molecular biosimilarity will not be as necessary, which will lower the cost of development. Regarding two different scores, the pLDDT and pTM, they allow the user to judge the reliability of the predicted structure. The authors of AF2 showed that the scores for different proteins have a high correlation with actual prediction accuracy (structural similarity to the native structure).

LITERATURE REVIEW

The review of the literature is divided into three major sections. The first section provides an overview of the ability of AlphaFold2 to perform near-to-experimentally driven structure prediction. The second section provides a review of the impact of mutations on the structure of proteins and their ability to bind. Moreover, the last section overviews the possible interpretations of prediction scores from these AI-driven prediction tools to prove their usefulness. The goal is to summarize the findings and conclusions of key research that has been undertaken in relation to the sections stated above.

2.1 Structure Prediction - AlphaFold2

AlphaFold2 predicts protein structures with an accuracy competitive with experimental structures in the majority of cases using a novel deep network-based model (23). In a study published in *Communications Biology*, the quality and usability of AF2 was demonstrated by testing its ability to predict the structure of protein complexes that were identified biochemically, but for which no experimental structural information was available. The study found that AF2 predicted the structures of the protein complexes with high accuracy (15). Recently, a study integrated AF2 with Cryo-Electron Microscopy to build an almost complete structure of the Nuclear Pore Complexes and revealed the first-ever structure of *Plasmodium falciparum* surface protein (Pfs48/45) (24), (25). AF2 was also used to identify a new distinct fold in rotavirus group B revealing its functionality and the predicted structure of stress-inducible phosphoprotein 1 (STIP1) revealing its role as a neuroprotective factor against Parkinson's disease. AF2 was also used to decipher which DNA mutations

are involved in genetic traits (26), (27), (28).

Even though AF2 does not take specific properties of Transmembrane (TM) proteins into account, the reliability of the generated TM structures, when quantitatively investigated for specific membrane proteins (e.g., dimer modeling and stability in molecular dynamics simulations) using a template-free model showed that AF2 performs well in the case of TM proteins and its neural network is not over-fitted. Thus, it was concluded that applications of AF2 structural models can advance TM protein-associated studies to a higher level (29). Since AF2 was not trained to handle phase boundaries that triggered an assessment of structural correctness, it was tested in a new database (30), (TmAlphaFold database: TMDDET) and AF2 was found not equipped to deal with structures for which it has barely any template, such as stand-alone TM segments.

A template-free, ab-initio protein model, used by AF2, showed that it was of sufficient quality to phase the native ORF8 dataset by Molecular Replacement (MR). The study claims that this approach can prove useful for future structural determination campaigns where a homologous structure is not available but could aid in the determination of pre-existing “unsolvable” datasets (31). Evidence that AF2 model has probably learned the energy function to rank the quality of predicted protein structures with reliable accuracy, without using any coevolution data and MSA proposes that it is a good starting point for structural optimization; significant for proteins with no structural homologs and MSA available, hence leading to a potential improvement of protein design methods (32). Providing either a template or a MSA for a receptor allows AF2 to identify the correct structure of the receptor and it might also identify the binding sites and conformations, and where a peptide must compete for two or more binding receptors, MSA is helpful (33). AF2 has proven to be useful to predict the structure of a whole host adhesion device from the *Lactobacillus casei* bacteriophage J-1. As exemplified by the human gut phageome, these AF2-based structure predictions can be used to revisit phage genome annotations and efficiently characterize newly discovered phages (34).

Furthermore, AF2's high prediction confidences for fold switchers indicate that it uses sophisticated pattern recognition to search for one most probable conformer rather than protein biophysics to model a protein's structural ensemble, hence it fails predictions for proteins whose characteristics cannot be completely inferred from their solved structures. These results highlight the importance of viewing protein structure as a whole and imply that fold-switching sequences may reveal a tendency for many stable secondary and tertiary structures through careful investigation (35).

2.2 Impact of Mutations on proteins structure

protein structure can now be better understood, including its stability and function and extending its applications, such as predicting the structural context of mutations associated with a disease or an escape from an immune response. Multiple studies were conducted on the applicability and interpretation of AF2 scores resulting in contradicting conclusions.

In order to test the impact of mutations on the fold prediction and hence the binding ability of proteins, multiple studies were conducted regardless of the disclaimer "has not been validated for predicting the effect of mutations" provided by the developers of AF2 (36). It was concluded that the AF2 models do not use template structures and do not improve binding free energy prediction ($\Delta\Delta G$); hence the prediction of the impact of a mutation on protein stability remains unresolved through AF2 ((37)).

Since it is crucial to create stable proteins efficiently and logically to be used in industry and health and help understand protein function where stability effects play a major role (38), (39) a comparison of several models to predict both the unfolded wild-type structure and the structures of the folded and unfolded mutant concluded that structure-based approaches only slightly outperform their sequence-based counterparts (40), (41),(42). The potential of AF2 predictions in the designability of new therapeutics and structural stability testing through mutagenesis analysis failed when the confidence scores of AF2 prediction showed no meaningful impact on the stability

of predictions and, therefore, no direct way to use AF2 for the prediction of $\Delta\Delta G$ upon mutation in the sequence has been identified yet (43). It was also shown that the best templates are homology models for the prediction of protein stability change upon mutation if the protein 3D structure is not available (44), (41).

2.3 AlphaFold2 scores interpretation

Through multiple studies, it has been confirmed that there is no correlation between the pLDDT and pTM scores and protein stability ($\Delta\Delta G$) as impacted by mutations on protein stability and function (45). The first structural analysis of hereditary cancer genes listed the thermodynamic stability predicted from AF2 structures as moderate but suggest that the confidence score of AF2 is a strong descriptor for variant pathogenicity, and the confidence score for a given variant in the AF2 structure could alone predict pathogenicity more robustly than even the stability predictors with an Area Under the Receiver Operating Characteristic Curve (AUROC) score. Moreover, the study concluded that the scores provided by AF2 according to the binding affinity values seem to work especially well in a comparative analysis study of strong binders competing against weak binders during docking (46).

In another study, factors that contribute to the inaccuracies of AF2 were tested 98-fold-switching proteins, which adopt at least two different yet stable secondary and tertiary structures, are the focus of this study. Five predicted, and two experimentally determined structures of each fold-switching protein were compared in terms of topological similarity. Generally speaking, 94% of AF2 predictions correctly predicted one of the experimentally determined conformations but not the other. Despite these biased findings, AF2's estimated confidences were moderate-to-high for 74% of fold-switching residues. This is in contrast to intrinsically disordered proteins' generally low confidence, despite their structural heterogeneity. Keeping this in view, since AF2 performs well at discriminating disordered regions, the pLDDT scores can be used for the characterization of the local dynamics of intrinsically disordered regions. A

thorough analysis of the pLDDT score could provide insight into the structural transientness, as well as the local function and dynamics (i.e., disorder–order transition) of IDP motifs, further enhancing the applicability of AF2 (47).

Although predicting protein structures is challenging due to their ensemble nature, using a curated collection of apo-holo conformations improvements in the holo form prediction in 70% of the cases were observed, even though it failed to capture observed conformational diversity as effectively as estimating a single conformation. The flexibility of the protein's main chain, specifically in the context of apo-holo pairs of conformers, showed a negative correlation with pLDDT scores – this relation can be used to infer local conformational changes linked to ligand binding transitions in a single 3D model (35).

AF2 is not likely to accelerate the experimental determination of 3D structures by improving the models for molecular replacement and it does not help resolve other problems or assist in protein folding applications. The lack of applicability and interpretation of the prediction scores has left a huge gap in understanding how this deluge of algorithms with claims to predict structure with high accuracy can be used. Even though the creators of these algorithms have said that no importance should be given to the predictability scores, it is difficult for the scientists to not dig into these scores, as each protein has its own score that is reproducible.

Is it because of the length and nature of the amino acid chain, is it because of the difference in thermodynamic energy, is it because of the 3D structure formed or is it entirely dependent on what the algorithm has been taught how to regurgitate the structures available for comparison in the databases like the PDB, UniProt or others. Another question that needs answering is why are these algorithms getting better in their predictability scores? Is it because they are getting better at reading the structures available for simulation, or is it because they are becoming capable of understanding the Levinthal paradox? These questions will be answered in this thesis

2.4 Study Rationale

Previous studies have highlighted the potential of protein structure prediction algorithms like AlphaFold2 and ESMFold in the prediction of 'close-to-native-state-like' protein structures and the effects of mutations on this predicted structure. However, there is a need to critically evaluate the practical utility of these tools and their ability to efficiently predict novel structures. This study aims to assess the correlation between prediction scores and physicochemical properties of FDA-approved therapeutic proteins. It also examines the impact of mutations on prediction scores and any effect this has on the functionality of these proteins. By investigating the limitations and applications of these algorithms, this research seeks to provide insights into their reliability and potential in protein structure prediction.

2.5 Objectives

The following are the objectives:

- To evaluate the reliability and accuracy of protein structure prediction tools for therapeutic proteins.
- To identify and compare the most variable regions and classes of therapeutic proteins, considering physiological, chemical, and functional properties.
- To establish the relationship between protein structure similarity and receptor binding mechanisms of action, utilizing 3D domain analysis.
- To investigate the impact of mutations on the structural and functional attributes of therapeutic proteins.
- To test the extent of dependence of AI-based structure prediction tools on the training data.

- To employ AI-based structure prediction tools to evaluate and rank therapeutic proteins (potential biosimilars) for the risk of structural variability based on prediction scores.

MATERIALS AND METHODS

This research was based on the observation and the premise that if a given amino acid sequence ends up with the same 3D structure, then the amino acid sequence similarity should suffice to demonstrate the structural similarity of a biosimilar candidate with its reference product. The second incentive came from the observation that each sequence when put through structure-prediction algorithms, gives a single score of confidence, so the score must somehow correlate with the nature of the sequence since it is not a random score, and reproducible if the algorithm is run multiple times. Connecting the confidence of predictability with protein structure has not been reported, and if a correlation can be established, this will allow rank ordering therapeutic proteins for the risk of **structure variability** when multiple batches are produced; thus, claiming that the proteins with high predictability confidence score will more likely be proven biosimilar. A presumption is made that a low score means uncertainty that might end up as variability during the pre-translation stages.

The anticipation was to identify proteins that are least likely to show variability in their structure and thus functional properties, an observation that might allow us to reduce the testing, which may help lower the current cost burden of \$100-300 Million for the development of each biosimilar.

3.1 Data Collection

204 FDA-approved therapeutic protein's amino acid sequences ranging from 5 to 1000, were obtained from their regulatory filing (48), listings in the FDA Purple Book (49), Orange Book (50), the patents reporting amino acid sequences, the Inight Drug (51), Kegg Pathway (52), and DrugBank databases (53). The sequences in the

UniProt database were found to have residual differences compared to the amino acid sequences in patents; hence, the resources mentioned above were used and cross-checked through these references for similarity. Conjugated, modified, pegylated, or combination protein products were excluded. The selected therapeutic proteins were all commercialized products with proven safety and efficacy.

The list of 204 products included 188 protein molecules and 16 molecules with amino acid sequence lengths less than 40, classified as polypeptides and not treated as biological drugs by the FDA (54). The data provided in Appendix A, separate these two classes, 'peptides' for 16 polypeptides and 'proteins' for the remaining 188 products. Each category's file contains information of the therapeutics name, its brand name, accession number from Inxight, KeggDrug or DrugBank databases, and Biologics License Application (BLA) number or New Drug Application (NDA) number acquired from FDA approval documentation. Furthermore, details of sequence and sequence length along with molecular weight computed by Cusabio tool (55), and the type of therapeutic molecule (i.e., enzyme, monoclonal antibody, blood factor, cytokine, growth factor, hormone, inhibitors, fusion protein, recombinant human protein, etc.) are identified in the files provided in Appendix A.

3.2 Structure prediction tools and scores

The amino acid sequences were put into 3D structure prediction with AF2 and ESMF through ColabFold (56) using the UCSF ChimeraX tool (57) for AF2 and an independent Google Colaboratory notebook for ESMF (<https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/ESMFold.ipynb#scrollTo=CcyNpAvhTX6q>) (58). The two confidence scores, pLDDT and pTM, for all proteins are reported in the Appendix A as 'AlphaFold pLDDT Score', 'AlphaFold pTM Score', 'ESMFold pLDDT Score' and 'ESMFold pTM Score'.

The correlations between the amino acid chain length, pLDDT, and pTM scores, orthogonal comparison between the two algorithms, structure-function relation between

the predicted molecules, structure-sequence-similarity/identity relations, and molecular ranking were compared to in anticipation of finding any biosimilarity marker.

3.3 Physicochemical properties

The physicochemical parameters, including hydrophobicity, isoelectric point, extinction coefficients, and instability index, for all 204 molecules were computed through a Python script (Appendix C) employing the Expasy ProParam package (<https://web.expasy.org/protparam/>) (59). The hydrophobicity was calculated using the GRAVY (grand average of hydropathy) index to measure the aggregation of hydropathy of amino acid residues. The isoelectric point (pI) was used to account for the pH of the protein at a net neutral charge. In addition, the theoretical molecular extinction coefficients for both reduced and non-reduced cysteine residue structures were calculated to determine the protein concentration by measuring its absorbance at 280 nm wavelength. Finally, the proteins' instability index was calculated based on the composition of its amino acids, with higher values indicating greater instability and more variations in protein degradation. These scores were further compared to gain insights into therapeutic proteins' physiological and functional properties.

(A) Cysteine Extinction Coefficient:

The cysteine (cys) extinction coefficient represents the theoretical molar absorption coefficient of the protein at a wavelength of 280 nm, assuming that all cysteine residues are involved in disulfide bonds. It calculates the absorption of light at 280 nm due to the presence of disulfide bonds in the protein.

(B) Reduced Cystein Extinction Coefficient:

On the other hand, the reduced cys extinction coefficient provides the theoretical molar absorption coefficient of the protein at 280 nm, considering that all cysteine residues are in the reduced state and hence there are no disulfide bonds. It

calculates the absorption of light at 280 nm when all cysteine residues are reduced and not involved in disulfide bonds.

3.4 Protein-target interaction

Docking is an established technique to acquire insight in the binding mode, affinity, and propensity of a protein-based-therapeutic to its target where the structures of the subunits have been determined either experimentally or computationally. In this study, LZerD (<https://lzerd.kiharalab.org/about/howtouse/>), a web server for pairwise and multiple protein-protein docking, which uses a soft protein surface representation with 3D Zernike descriptors and explores the binding pose space using geometric hashing, was employed for molecular docking (60). The LZerD suite of methods has been ranked near the top of all server groups in recent rounds of CAPRI (Critical Assessment of PRedicted Interactions) - a community-wide blind experiment for testing computational algorithms in blind predictions of experimentally determined 3D structures of protein complexes (61), (62), (63), (64).

The LZerD server was used to dock cytokines, hormones, and fusion proteins with their respective targets identified through the DrugBank database and retrieved from the PDB database. The complexes included one therapeutic protein with a high pLDDT score and one with a low pLDDT score as ranked by AF2 and ESMF.

PDB structures often have non-standard chain names and residue numbering that can cause compatibility issues with the docking tools. For standardization, chains were renamed, and residues were renumbered using UCSF Chimera software (65). Docked complexes with the highest rank-sum from GOAP, DFIRE, and ITScore scores, from the LZerD server were given to PRODIGY server and their Gibbs free energy/binding affinity (ΔG), dissociation constant (K_d), Interfacial Contacts (ICs) and Non-Interacting Surfaces (NIS) values were computed.

The combination of GOAP, DFIRE, ITScore, and PRODIGY represents a robust approach for protein-protein docking, accounting for various factors affecting the bind-

ing affinity values. GOAP is an orientation and distance-dependent all-atom statistical potential computation using a distance-scaled finite ideal-gas reference state for the distance-dependent components; the parameters of DFIRE (66). DFIRE employs distance-dependent structure-derived potentials (DDPs), that are used to predict the energy of a protein-protein complex, accounting for the reference state, which serves as a baseline for comparing the interactions in the complex. ITScore, on the other hand, is an energy evaluation method for the electrostatic and van der Waals interactions between the protein and target on the atomic level using statistical mechanics principles (67). Finally, PRODIGY calculates the binding energy values based on the 3D structure of the docked complex in the form of Gibbs free energy (ΔG) and dissociation constant (Kd) (68). It also enumerates the number of Interatomic Contacts (ICs) made at the interface of a protein-protein complex within a 5.5 Å distance threshold and classifies them according to the polar, apolar, and charged character of the interacting amino acids. Using a combination of these methods provides a comprehensive assessment of protein-protein interactions and improves the accuracy of the docking predictions acquired from LZerD server. Collectively these physiological, chemical, and functional parameters were employed to analyze their relationship with prediction scores as discussed in the next chapter.

3.5 Protein domains & pTM score

The pTM score provides a more accurate estimation of the precision of the predicted protein structure about its native structure. The pTM score indicates the quality of the predicted model based on the similarity between the predicted structure and the experimentally determined structure of a related protein, which serves as a template for the prediction. CASP experiments have found the TM-score a useful metric for evaluating the accuracy of predicted protein structures, indicating that the best models for most targets were equal to or better than the best template available in the PDB. Overall, the results suggest that the TM-score is a valuable tool for assessing

3.6 Investigating relationship between Predictions, Structural and Sequential Data

protein structure and domains. 13 out of 188 protein molecules within the dataset had significantly low pTM scores (<0.5) (69),(70). Since structural flexibility is highly dependent upon the environment and interactions of proteins, the unavailability of such experimental data for specific proteins may have led to lower pTM scores. Sequence data is the core of the training set for both prediction models; hence lower sequence identity between the domains of the target and template could also lead to lower scores since accurate template-based modeling is highly dependent upon the quality of the template structure available. The presence of multiple domains, where each has a unique structural feature, also influences the overall prediction score.

Additionally, flexible linkers between the domains may greatly impact the prediction of such regions. These significantly lower pTM scores could have resulted from the presence of intrinsically disordered regions, multiple domains, complex folds, lower sequence conservation, predicted secondary structure, lower prediction power of the model, or simply due to the unavailability of specific combinations of residues in the training set. A trRosetta domain-based analysis using the AF2 and ESMF pTM scores was carried out to compare the impact of multiple domains and the prediction power of the modeling tool. NCBI Conserved Domain Database (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>) (71) was used to identify domains in the proteins and categorize therapeutics into single-domain or multi-domains as shown in Table 4.8 (section 4.7).

3.6 Investigating relationship between Predictions, Structural and Sequential Data

With increased prediction scores with the better and more complex model architecture of prediction algorithms, they should be able to predict novel protein folds regardless of the unavailability of similar sequential/structural data in PDB, UniProt, and other databases. To correlate the dependence of predictions by

AF2 and ESMF on the availability of structural and sequential data present in PDB, the PDB Blast Query coverage and Percentage Identity scores were listed down. The highest similarity sequences against respective therapeutic proteins from UniProt were also retrieved after alignment. The sequence alignment identity and similarity scores were recorded in Table 4.10. Every UniProt protein has a respective structure stored in AlphaFold Database (<https://alphafold.ebi.ac.uk/>) (72). These proteins, and pLDDT scores extracted from their respective mmCIF files are also listed for further analysis in order to find a correlation.

3.7 Randomization & Mutation

The domains sequences identified from NCBI-CDD database were ‘shuffled’ using Molbiotool’s Random Sequence Generator to ensure the highest mutation rate (<https://molbiotools.com/randomsequencegenerator.php>) (73). First, a single domain was randomized through shuffling, and then all the domains of Trastuzumab, Etanercept, Coagulation Factor VIIa, and Darbepoetin alfa were randomized in order to produce novel molecules. These mutated sequences were ran through BLAST PDB, and their query coverage, along with their percentage identity scores, were retrieved. This data was generated and collected to identify any similarities between the randomized sequence combinations and folds present in the UniProt and PDB database (as shown in Table 4.11) (section 4.9). If the data was available, the AF2 and ESMF models must have learned them during the training phase. However, if the coverage and identity scores were extremely low or zero, the AF2 and ESMF models would rely solely on their training to predict the structure. Since ESMF does not use MSA, it can be anticipated that this model would perform better than MSA -dependent model AF2.

3.8 Affinity Maturation: Trastuzumab

3.8.1 Trastuzumab & Tyrosine-protein kinase erbB-2 (HER2)

Trastuzumab is a Monoclonal Antibody commonly used to treat certain types of breast cancer. It specifically targets the receptor tyrosine-protein kinase erbB-2, which is overexpressed in these cancer cells. By binding to erbB-2, Trastuzumab inhibits signaling pathways involved in cell growth and survival. Receptor tyrosine-protein kinase erbB-2, also known as HER2, is a protein that plays a crucial role in cell growth and differentiation. Overexpression of erbB-2 is associated with aggressive forms of breast cancer and hence serves as a target for therapies like Trastuzumab.

3.8.2 Alanine Scanning

Alanine scanning is an in-silico technique used to study protein-protein interactions and identify key amino acid residues that contribute to the binding affinity between two proteins. It involves systematically substituting specific amino acids with alanine and measuring the resulting change in binding affinity values. The increase in affinity value vouches for the increased therapeutic potential of a protein when tested experimentally, also known as affinity maturation.

3.8.3 mCSM-PPI2

The mCSM-PPI2 tool (https://biosig.lab.uq.edu.au/mcsm_ppi2/) is a computational tool designed for predicting the effects of mutations on protein-protein interactions (74). It utilizes a machine learning-based approach to estimate changes in binding affinity upon mutation through alanine scanning. By analyzing the interactions between residues in the complex, mCSM-PPI2 can provide insights into how specific mutations impact binding strength.

3.8.4 Trastuzumab & HER2 affinity maturation

An experimentally docked complex of Trastuzumab, Pertuzumab, and the receptor tyrosine-protein kinase erbB-2 was acquired from RCSB PDB (ID: 6OGE). The Pertuzumab FAB Heavy and Light chains were removed from the complex through PyMol tool, resulting in a modified molecule. The remaining structure consisted of Trastuzumab FAB Light and Heavy chains, complexed with the Receptor tyrosine-protein kinase erbB-2 represented by chains D, E, and A respectively in the .pdb file. The truncated complex was subjected to alanine scanning using the mCSM-PPI2 tool. The identified positions can be systematically mutated from the alanine scanning analysis using 19 possible amino acid substitutions. The resulting mutant variants can then be assessed for their binding affinity values using the PRODIGY server. This analysis aims to evaluate the impact of different amino acid changes on the binding affinity and facilitate the identification of potential modifications that could lead to the development of a therapeutically enhanced molecule. By leveraging computational tools like PRODIGY, a comprehensive understanding of the binding properties can be obtained, aiding in the design and optimization of novel therapeutic candidates.

RESULTS

An in-depth analysis of physicochemical and functional parameters resulted in multiple significant conclusions and each has been discussed in detail below. Briefly, the very first finding was that both algorithms (AF2 and ESMF) yield comparable and reproducible scores, confirming the orthogonality of structure prediction uncertainty. However, no correlation was found between any physicochemical or functional properties of therapeutic proteins with the predictability scores. The binding site predictions also do not correlate binding properties with protein structure prediction scores. The predictability confidence scores vary, even if smaller changes or mutations are introduced in the amino acid chain sequence. This observation came from comparing the scores of proteins whose amino acid chain sequences reported in the PDB or UniProt was different from the commercial product sequences.

4.1 Sequence Length vs Molecular Weight

The first comparison and correlation made was between the sequence length and molecular weight with R^2 0.99 despite the differences in the molecular weight of amino acids. This establishes the normal distribution of amino acids in the set of therapeutic proteins studied.

4.2 Proteins & Peptides Correlation

There was no linear or nonlinear relationship between the sequence length and the pLDDT or pTM scores (Figures 4.1, 4.2, 4.3). Only peptides showed a weak positive correlation coefficient (R^2 0.40) in pLDDT from AF2, but not in the ESMF

score. The two algorithms correlated with the correlation (R^2) 0.60 for pTM scores (Figure 4.4), between the two algorithms.

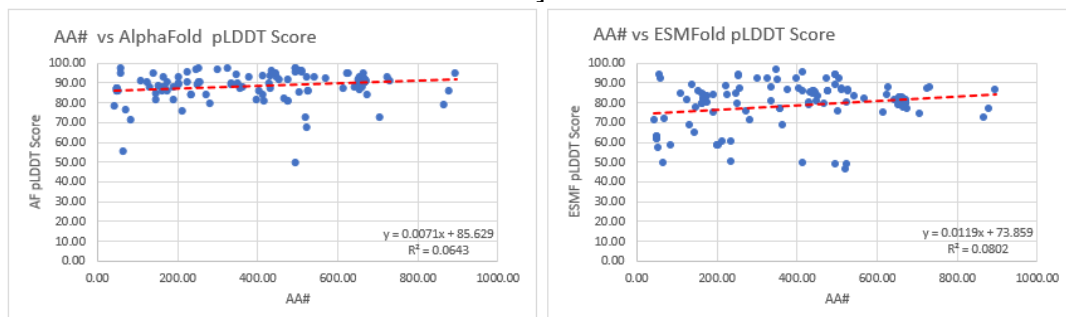


Figure 4.1. Correlation between amino acid chain length and pLDDT scores from AlphaFold2 and ESMFold of therapeutic proteins

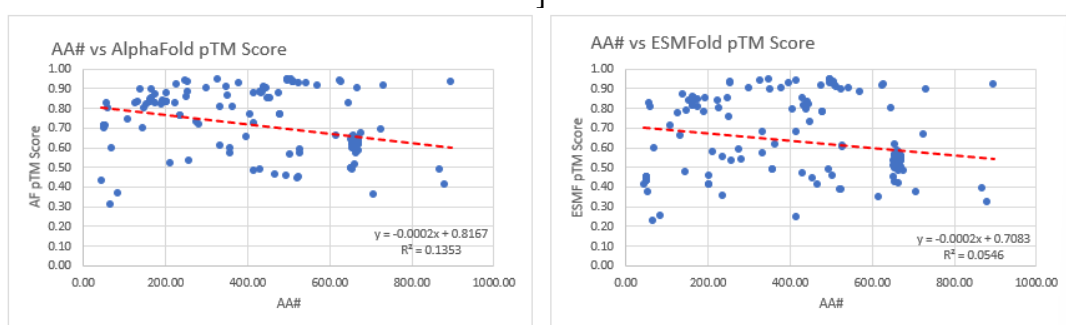


Figure 4.2. Correlation between amino acid chain length and pTM scores from AlphaFold2 and ESMFold of therapeutic proteins

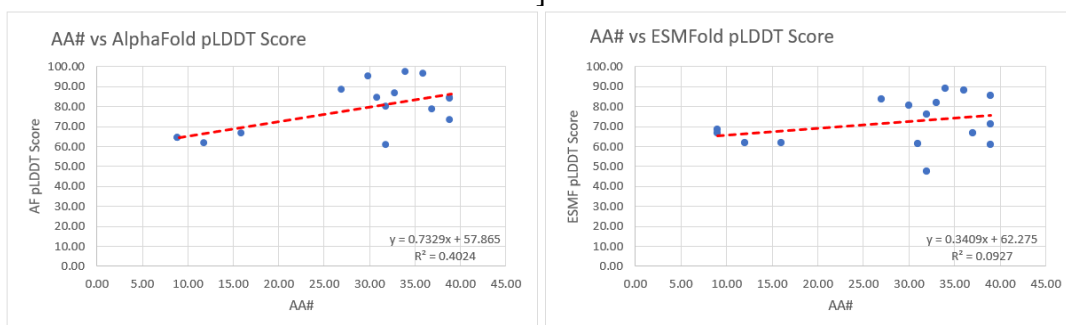


Figure 4.3. Correlation between amino acid chain length and pLDDT scores from AlphaFold2 and ESMFold of therapeutic peptides

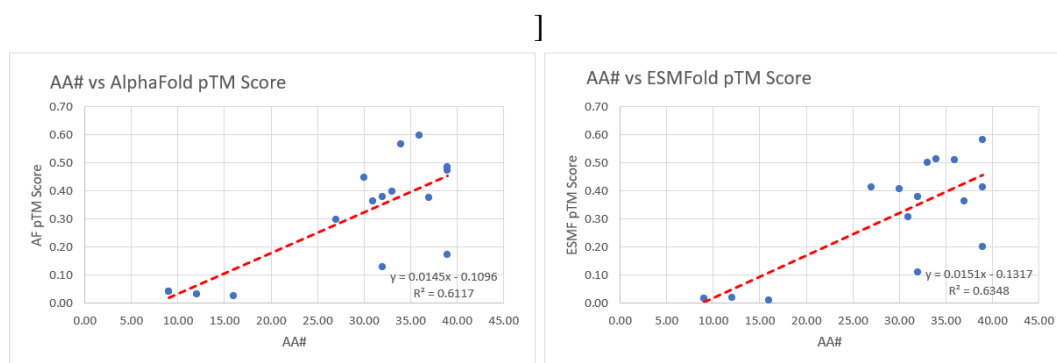


Figure 4.4. Correlation between amino acid chain length and pTM scores from AlphaFold2 and ESMFold of therapeutic peptides

4.3 Orthogonal comparison between AF2 vs. ESMF

The Pearson correlation of the pLDDT scores between the AF2 and ESMF on a random subset of around 4000 metagenomic sequences was reported to be 0.79 ((?)), however, it was found the Pearson correlation to be 0.72 from the data of 204 molecules (peptides and proteins). The Pearson correlation (corr.) of the pTM scores between the AF2 and ESMF was 0.88. For comparison, two cut-offs were used: **AA#<40** and **40<AA#<1000**. The first was to understand the predictability of polypeptides below 40 amino acids from AF2 and ESMF, which resulted in a correlation of 0.83 using the pLDDT score and 0.95 using the pTM score. The second cut-off was of all the proteins above 40 amino acids and below 1000, which resulted in a correlation of 0.69 using the pLDDT score and 0.84 using the pTM score. The Pearson correlation (corr.) and correlation coefficient (R^2) for pTM scores from both algorithms were in better agreement than the pLDDT scores (Figure 4.5). Furthermore, the correlation for proteins was relatively lower than for peptides, hence implicating that the predictability of these tools decreases with an increase in structural complexity.

4.4 pLDDT & pTM Rank ordering Biosimilars

The prediction scores represent the reliability of predicted structures and implicitly account for the structural variability that may occur in-vivo systems due to

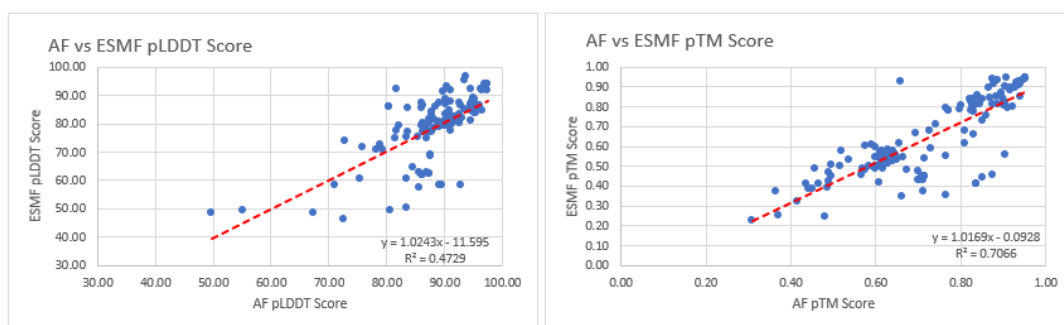


Figure 4.5. Correlation between the pLDDT and pTM Scores of both AlphaFold2 and ESMFold for proteins

pre-translation modifications during multiple batch productions. These scores can be used to rank order, and thus make a basis for reduced testing of biosimilar candidates with higher scores. However, since the correlation scores do not conform to any known physicochemical or function properly, these can be best labeled as random scores with little justification that recognizable folds with higher structural stability can be of any predictive value. The Table B.1 (Appendix B) reports a rank order of therapeutic proteins following the pLDDT scores, and B.2 (Appendix B) uses pTM scores using both algorithms.

It was noteworthy that all FDA and EMA-approved biosimilars demonstrated pLDDT scores greater than 80 using the AF2 predictions (Table 4.1). If there were any correlation between the confidence score and the free energy, it would have been a reasonable assumption that these proteins are more stable, leading to the same 3D structure, and thus, extensive comparative testing may not be required (9).

Table 4.1. Confidence scores of approved biosimilars using AF2 and ESMF

Product	AA#	pLDDT (AF2)	pTM (AF2)	pLDDT (ESMF)	pTM (ESMF)
Teriparatide	34	96.80	0.57	89.09	0.51
Etanercept	467	82.10	0.47	79.23	0.41
Ranibizumab	445	94.50	0.90	84.93	0.82
Adalimumab	665	91.50	0.65	81.90	0.53
Rituximab	664	91.30	0.64	81.43	0.54
Trastuzumab	664	91.00	0.61	82.01	0.58
Bevacizumab	667	90.69	0.90	81.01	0.56
Filgrastim	175	90.20	0.87	83.34	0.84
Follitropin alfa	203	89.60	0.84	58.57	0.42
Interferon beta	165	87.70	0.84	84.39	0.86
Erythropoietin-alpha	168	87.70	0.84	83.95	0.85
Interferon alpha	165	87.00	0.84	79.58	0.81
Insulin aspart	51	86.30	0.71	62.55	0.45
Insulin glargine	53	85.60	0.71	57.46	0.38
Somatropin	191	81.50	0.82	74.81	0.78
Infliximab	440	95.00	0.91	84.49	0.79

4.5 Physicochemical Attributes

The 3D structures reflect the underlying intra-molecular interaction that governs protein folding and stability, such as hydrogen bonding, electrostatic interactions, hydrophobicity, and van der Waals forces. The protein's structure, stability, binding pockets, binding affinity, and desolvation are all heavily influenced by the intramolecular interactions, which also have an impact on the protein's physicochemical qualities.

AF2 and ESMF algorithms implicitly capture these attributes, being trained on a large dataset of experimentally known protein structures. However, there was a very weak correlation in proteins (Table 4.2) and in the peptide (Table 4.3), resulting in a conclusion that certain protein chunks govern its therapeutic potential rather than the entire structure for which AF2/ESMF provides prediction scores.

Both AF2 and ESMF consider known structures of homologous proteins, physicochemical properties of amino acids, and conformational constraints to generate a 3D structure that is energetically favorable and perform molecular dynamic simulation relaxation step, i.e., AMBER force field in AF2, to find the closest-to-native-state structure. Although they might not typically predict pre-translation variability, which refers to the variability in the amino acid sequence of a protein that arises from alternative splicing, epigenetic modifications, or genetic variations, due to the presence of these modifications in their training data, they might, again, implicitly account for such changes. Furthermore, they may indirectly point to pre-translation variability by generating multiple alternative models that account for different conformations, i.e., the generation of five relaxed and unrelaxed conformations in the case of AF2. As a result, these algorithms were projected to correlate with pre-translation variability on the structure and function of the protein, however, the predicted structures were found to be uncertain and could not be used for any purpose unless validated experimentally, in which case, the predictability becomes a moot point.

Table 4.2. Correlations between structural and physicochemical parameters of 188 therapeutic proteins with strong correlations (>0.5)

	AA	MW	AF pLDDT	AF pTM	ESMF pLDDT	ESMF pTM	Hydrophobicity	Isoelectric point	Extinction coefficients (red-cys)	Extinction coefficients (cys)	Instability Index
AA	1.00										
MW	1.00	1.00									
AF pLDDT	0.25	0.25	1.00								
AF pTM	-0.37	-0.36	0.51	1.00							
ESMF pLDDT	0.28	0.29	0.69	0.42	1.00						
ESMF pTM	-0.23	-0.22	0.42	0.84	0.67	1.00					
Hydrophobicity	-0.31	-0.32	0.07	0.35	-0.14	0.21	1.00				
Isoelectric point	0.20	0.19	0.23	-0.16	0.11	-0.15	-0.24	1.00			
Extinction coefficients (red-cys)	0.92	0.92	0.32	-0.26	0.35	-0.13	-0.31	0.20	1.00		
Extinction coefficients (cys)	0.91	0.92	0.32	-0.26	0.35	-0.12	-0.31	0.20	1.00	1.00	
Instability Index	0.19	0.18	-0.05	-0.19	-0.05	-0.16	-0.42	0.31	0.15	0.15	1.00

Table 4.3. Correlations between structural and physicochemical parameters of 16 therapeutic peptides with strong correlations (>0.5)

	AA	MW	AF pLDDT	AF pTM	ESMF pLDDT	ESMF pTM	Hydro- phobicity	Isoelectric point	Extinction coefficients (red-cys)	Extinction coefficients (cys)	Instability Index
AA	1.00										
MW	1.00	1.00									
AF pLDDT	0.63	0.69	1.00								
AF pTM	0.78	0.81	0.93	1.00							
ESMF pLDDT	0.3	0.37	0.83	0.73	1.00						
ESMF pTM	0.80	0.81	0.91	0.95	0.78	1.00					
Hydrophobicity	-0.37	-0.42	-0.27	-0.33	-0.26	-0.36	1.00				
Isoelectric point	-0.01	0.00	-0.14	-0.16	-0.23	-0.15	-0.34	1.00			
Extinction coefficients (red-cys)	0.51	0.58	0.55	0.61	0.38	0.49	-0.40	-0.50	1.00		
Extinction coefficients (cys)	0.51	0.59	0.55	0.61	0.39	0.50	-0.40	-0.50	1.00	1.00	
Instability Index	0.13	0.16	-0.02	-0.02	-0.05	-0.03	0.04	0.03	0.06	0.06	1.00

Proteins must be folded into their native stable states to perform their function, which typically involves binding to their respective targets. They have the inherent ability of stable fold formation and strong binding interactions, acquired through adaptation and conservation, even when these changes do not directly increase the organism's fitness (75). The distribution of polar and apolar residues on the surface mediate protein-target interactions, influencing their specificity and affinity. Multiple studies have concluded that the charged residues interact with targets through the exposed surfaces rather than the interface to affect the binding ability of the interacting proteins. Enhanced intra-molecular electrostatic interactions lower the desolvation penalty. In contrast, the inter-molecular interactions with charged residues on the target molecule enable better complementarity and electrostatic steering, resulting in increased solubility and bioavailability of these proteins in living systems (76). Physicochemical parameters like hydrophobicity and isoelectric point also play a crucial role in these interactions, contributing to the stability of 3D folds formed. The computed list of physicochemical parameters was analyzed to gain insights into therapeutic proteins' physiological and functional properties.

The bioavailability, pharmacokinetics, and pharmacodynamics of a therapeutic drug are greatly influenced by its structural elements, as well as the concentration and dosage of the drug. The extinction coefficient is often used in protein purification, quantification, and structural studies where accurate protein concentration determination is required for therapeutics (77), (78). The extinction coefficient computed through ProtParam, validated through the findings of Gill and von Hippel (79), is based on Beer-Lambert Law, which states that the absorbance of a solution (water) is directly proportional to the concentration of the solute (protein) and the path length of light through the solution. Since the path length is fixed (1cm), the absorbance is proportional to the protein concentration. With increased amino acid length, the protein extinction coefficients, reduced cystine, and oxidized cystines both increase, exhibiting a Pearson correlation of 0.92 and 0.91 respectively. Larger proteins, like

monoclonal antibodies, tend to have more chromophores (molecules that absorb light), such as aromatic amino acids like tryptophan and tyrosine, resulting in a strong positive correlation. These chromophores contribute to the overall light absorption of the protein, leading to a higher extinction coefficient. The average extinction coefficient is $105 \text{ M}^{-1}\text{cm}^{-1}$ for monoclonal antibodies with an average molecular weight of 69.25 kD and $13.54 \text{ M}^{-1}\text{cm}^{-1}$ for hormones with an average molecular weight of 15.29 kD (Table 4.4).

Since the extinction coefficient is influenced by the electronic transitions that occur within a molecule, it is in turn influenced by the structure. Confidently predicted structures can provide insight into the likely folding of a protein, which can in turn inform predictions of its extinction coefficient. This information can be useful in the development of methods for measuring the concentration of a protein in a solution, which is critical for ensuring proper dosing and efficacy of a therapeutic, leading to a reduction in resource expense during clinical testing. In conjugation with the presence of chromophores, protein stability is also affected by the amino acids' innate half-life and disulfide linkages. However, a lack of significance of the predictability scores renders such applications fruitless.

Table 4.4. Average extinction coefficients (reduced cystine) for all types of molecules with the average amino acid number and molecular weight

Type of Molecule	AA#	MW	AF pLDDT	ESMF pLDDT	Extinction coefficients (rd cys)	Instability Index
Hormone	137.50	15.29	84.54	65.68	13.54	40.05
Cytokine	200.56	23.03	85.76	74.16	23.88	50.87
Growth factor	241.67	26.89	88.73	80.72	32.58	50.78
Inhibitors	293.33	33.35	81.92	80.85	26.35	41.30
Enzyme	456.61	51.11	91.40	87.38	89.34	38.04
Blood factor	557.25	62.50	86.10	84.57	80.33	41.60
Fusion protein	571.20	64.01	84.77	78.07	77.21	43.40
Monoclonal Antibody	633.33	69.25	91.15	81.39	105.51	47.18

Proteins with a higher abundance of residues with a lower half-life tend to have a relatively higher instability index. Therefore, they may have a shorter lifespan in vivo being more prone to degradation. However, even when the pLDDT scores are high, few proteins have higher instability index (more susceptible to degradation), i.e., Choriogonadotropin alfa has AF2 pLDDT score 83.40 and the chances of its degradation in vivo are high (instability index 67.46). Similarly, Sargramostim has a confidence-predicted structure from AF2 with a pLDDT score of 90.10. Still, the instability index is 63.87, indicating that a reliable structure prediction cannot vouch for the structure's stability in vivo. This eliminates the possibility of correlating pLDDT scores with instability indexes and hence a confidently predicted structure cannot vouch for the stability of a protein in in-vivo systems.

If the pLDDT score of a structure is high, it vouches for its structure reliability. Still,

the thermodynamic stability of the structure, as represented by the scores, comes from the sequences that are not necessarily part of the functional domains or terminal chains responsible for the physicochemical and functional properties. Proteins with low predictability scores are more likely to show pre-translation modifications between batches and may exhibit greater variability in their physicochemical properties. Although all the computed physicochemical properties depend on the sequence's residual composition, they remain unaffected by the predicted structures and scores, resulting in no correlation.

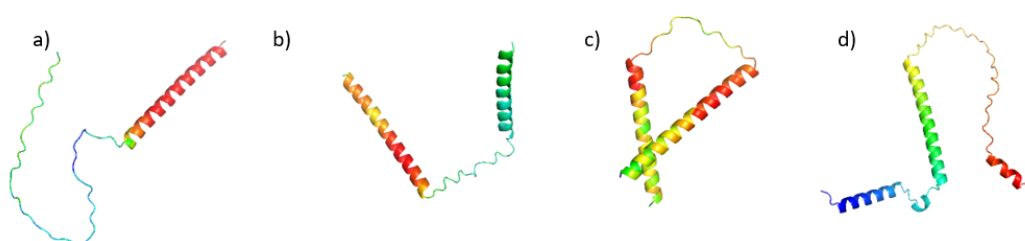


Figure 4.6. AF2 has a slight tendency to produce overfitted results, but in the case of parathyroid, it ‘under-predicted’ or misfolded the alpha-helix. Expaty ProtScale was used to verify chemically the presence of two alpha helices in parathyroid, out of which, only one was predicted by AF2, unlike ESMF and trRosetta. a) Misfold predicted structure by AlphaFold2 with missing alpha-helix on left. b) Confidently predicted structure by ESMFold with both helices present c) Confidently predicted structure by trRosetta d) Misfold structure present in AlphaFold Database with a partially missing alpha-helices

Despite using similar sequences in structure prediction from AF2/ ESMF, the variability in their conformational folds would result in different experimentally driven physicochemical parameters, including light absorbance and stability seen through extinction coefficient and instability index. The folds produced by AF2 and ESMF vary for similar sequences, i.e., parathyroid (Figure 4.6), affecting the accessibility of residues to the solution. For instance, if a protein region predicted to be unstable in one conformation becomes buried or stabilized in a different conformation, the extinction coefficient and instability index may vary drastically.

4.6 Proteins Interactions: Effects of Structural folds

The atomic pLDDT by AF2 and ESMF measures atomic-level prediction accuracy based on the degree of agreement between the predicted model and the experimental structure. In principle, certain portions of a protein hold therapeutic potential with residues responsible for binding. Parathyroid (PTH) structures predicted from AF2 and ESMF when docked to the PTHR1 receptor (PDB: Q03431) through LZerD server and evaluated through the PRODIGY server produced a ΔG value of -11.1 and -10.3 respectively. Despite the ability of AF2 to misfold structures (Figure 6), as observed in other studies, (80),(81) the predicted structure has a higher pLDDT/pTM score (pLDDT: 71.00, pTM: 0.37) and binding affinity in comparison to ESMF predicted structure (pLDDT: 58.50, pTM: 0.25), indicating that if the domains are strongly predicted and the rest of the structure does not produce hindrance, strong binding can be obtained. Therefore, it is in fact possible for a protein residue to have low atomic pLDDT scores and contribute towards a strong binding affinity with its target and vice versa.

The literature evidence (82) and residue-residue pair file (.ic) produced by the PRODIGY server indicated that the residues 1-37 of PTH contributed to the binding with PTHR1. Few of the residues involved in binding - Ser1, Ser3, Glu4, Ile5, Leu7, Met8, Leu11, His14, Leu15, Ser17, Met18, Glu19, Arg20, Phe34, of PTH structure predicted from ESMF had low pLDDT, but when predicted from AF2, had high pLDDT, as represented in Table 4.5. The lower confidence residues lead to lose interactions hence lowering the binding affinity for ESMF-PTH. The ICs and NIS being the measures of the number and type of interactions between the PTH and PTHR1, vary. These scores are based on the number of salt bridges, hydrogen bonds, and electrostatic interactions formed between the charged, polar, and apolar residues of the protein and its target. Therefore, they depend highly on the 3D structure (Table 4.7). One explanation for this variation in the IC/NIS scores is that in the AF2-PTH structure, charged residues

Table 4.5. Interacting PTH-PTHr1 residues pLDDT from AF2 and ESMF, although they have different pLDDT scores but produced similar binding interactions and scores

PTH residue type	PTH residue number	AF2 residual pLDDT	ESMF residual pLDDT (averaged)
Ser	1	85.82	48.86
Ser	3	94.47	66.06
Glu	4	95.84	63.25
Ile	5	96.16	65.74
Leu	7	96.63	65.94
Met	8	97.32	68.10
Leu	11	97.24	61.35
His	14	96.93	64.82
Leu	15	97.02	69.84
Ser	17	96.32	66.47
Met	18	96.94	66.27
Glu	19	96.60	48.86
Arg	20	96.58	66.06
Phe	34	97.32	63.25

are positioned in such a way that they form multiple interactions with apolar residues of the PTHr1, resulting in a higher charged-apolar IC score in comparison to ESMF-PTH. Therefore, the IC and NIS scores between the two predicted structures differ due to variations in their predicted 3D conformations, which can affect the strength and number of protein-target interactions.

Table 4.6. AF and ESM predicted cytokines, hormones and fusion proteins binding affinity values calculated from PRODIGY server

	Target	AF	ESMF			
		PRODIGY G(kcal mol ⁻¹)	pLDDT	PRODIGY G(kcal mol ⁻¹)	pLDDT	
Cytokines	<i>Aldesleukin</i>	ILR2 (PDB: 2B5I)	-12.20	87.60	-14.50	68.90
	<i>Denileukin diftitox</i>	CD25 (PDB: 1Z92)	-11.30	72.50	-13.70	46.47
Hormones	<i>Aprotinin</i>	Mesotrypsin (PDB: 5TP0)	-10.30	97.00	-10.90	94.25
	<i>Parathyroid</i>	PTH1R (PDB: Q03431)	-11.1	71.00	-10.30	58.50
Fusion proteins	<i>Alefacept</i>	CD2 (PDB: 1CDB)	-9.40	72.80	-12.80	74.24
	<i>Belatacept</i>	CD86 (PDB: 1NCN)	-10.90	87.80	-10.40	77.00

In some cases, proteins, by nature, can retain their functional properties regardless of conformational variation, given that the domains were predicted confidently and the remaining structure does not produce hindrance in binding. For Aldesleukin, Denileukin diftitox, and Aprotinin, the binding affinity (ΔG) values with lower pLDDT predicted complexes from ESMF are relatively better, in comparison to higher pLDDT scores from AF2 (Table 4.6). Aldesleukin showed a binding affinity of -12.20 (AF pLDDT: 87.60) and -14.5 (ESMF pLDDT: 68.90). Similar observations were seen for Denileukin diftitox and Aprotinin, while Parathyroid, Alefacept, and Belatacept binding to their respective targets produced better energy values for higher pLDDT scored structures from AF2 in comparison to ESMF predicted structures. Besides binding affinity, variation in the Interfacial Contacts (ICs) and Non-Interacting Surfaces (NIS) between charged (e.g., glutamic acid, aspartic acid), polar (e.g., serine, cysteine), and apolar (e.g., alanine, valine) were observed (Table 4.7). Several structural differences in proteins can influence binding affinity (83), (84). The ICs and NIS of proteins, and residue-pairs with charged and aromatic side chains are important for binding. These residues influence the formation of cationic, electrostatic, and aromatic interactions between the protein and target molecule helping explain the drastic variance in the binding affinity. From the given data in Table 4.7, the ΔG is more favorable for the Aldesleukin structure predicted from ESMF, in comparison to the AF2 predicted structure from the same sequence. The ΔG value for the ESMF predicted structure is more negative (-14.5) than the ΔG value for the AF2 predicted structure (-12.2), which indicates a more energetically favorable interaction in the Aldesleukin-ESMF case. Additionally, the K_d value for the ESMF predicted structure is lower (2.2×10^{-11}) than that of the AF2 predicted structure (1.1×10^{-9}), indicating a stronger binding affinity of Aldesleukin-ESMF to its target Interleukin Receptor 2 (ILR2). Overall, this data suggests that the ESMF-predicted structure has a stronger interaction and higher binding affinity with its target, regardless of a lower pLDDT (68.90) value.

Table 4.7. PRODIGY results of AF2 and ESMF predicted structures docked with their respective targets

Protein-protein complex	AF PTH-PTH1	ESMF PTH-PTH1	AF Aldesleukin-ILR2	ESMF Aldesleukin-ILR2	AF Trastuzumab-HER2	ESMF Trastuzumab-HER2
G	-11.1	-10.3	-12.2	-14.5	-12.4	-12.3
Kd	7.4×10^{-9}	2.7×10^{-8}	1.1×10^{-9}	2.2×10^{-11}	7.6×10^{-10}	9.7×10^{-10}
ICs charged-charged	1	2	3	7	4	10
ICs charged-polar	7	4	14	12	16	8
ICs charged-apolar	30	34	22	21	20	22
ICs polar-polar	3	1	5	5	6	4
ICs polar-apolar	21	14	18	27	24	18
ICs apolar-apolar	47	37	11	16	19	31
NIS charged	22.06	22.13	26.04	25.64	21.95	22
NIS apolar	48.63	48.95	30.57	30.92	38.15	37.79

Extending the argument, comparing the ICs and NIS data for the predicted structures of Aldesleukin, it can be seen that Aldesleukin-ESMF has a higher number of ICs than AF2. This suggests that the Aldesleukin-ESMF structure has more extensive interactions with its target molecule than the Aldesleukin-AF2 structure. Specifically, the Aldesleukin-ESMF structure has more polar-apolar contacts (Table 4.7), which are interactions between polar and non-polar amino acids. This indicates that the Aldesleukin-ESMF structure is more polar in nature and may have more hydrogen bonding interactions with the target molecule, hence resulting in better affinity values. On the other hand, Aldesleukin-AF2 has more charged-apolar contacts, which suggests that its structure is more hydrophobic in nature and has relatively more Van der Waals interactions with the target molecule. The values of NIS charged and NIS apolar for Aldesleukin-AF2 and Aldesleukin-ESMF are very similar, with only a small difference between the two. This indicates that both structures have similar areas that do not interact with the target molecule and that the two methods have produced comparable results in terms of predicting these non-interacting surface areas. However, it must be noted that the number of NIS does not account for the nature or location of the interacting regions.

As deduced, larger molecules like Monoclonal Antibody (mAb), i.e., Trastuzumab, predicted by AF2 and ESMF have reliable pLDDT (AF: 91.00, ESMF: 0.61) and pTM (AF: 82.01, ESMF: 0.58) scores indicating higher confidence in predicted structures. Slight variance in the ΔG , Kd, ICs, and NIS is observed for Trastuzumab-AF2 and Trastuzumab-ESMF when docked to their respective target HER2 (Table 4.7). This high variance in pLDDT scores and ΔG of protein complexes indicate that there is no definitive correlation between the structural and functional parameters hence leading to a conclusion that these parameters are independent of each other.

Even when the sequence remains the same, predicted structures through AI-based tools alone cannot be wholly relied upon to interpret protein's functional properties and hence requires further analysis or experimental validation. Sequence complex-

Table 4.8. Single and multi-domain molecules with low pTM (<0.5) predicted from AF2 and ESMF

Name	AF2 pTM	ESMF pTM	Domain
Lepirudin	0.31	0.23	Single
Alefacept	0.36	0.37	Multiple
Parathyroid/Preotact	0.37	0.25	Single
Rilonacept	0.42	0.33	Multiple
Lixisenatide	0.43	0.41	Single
Denileukin diftitox	0.44	0.38	Multiple
Tagraxofusp	0.45	0.39	Multiple
Elosulfase Alfa	0.46	0.49	Multiple
Etanercept	0.47	0.41	Multiple
Menotropins	0.48	0.25	Multiple
Eftrenonacog Alfa	0.49	0.39	Multiple
Aflibercept	0.49	0.47	Multiple
Tositumomab	0.49	0.43	Multiple

ity, structural flexibility, intrinsically disordered regions, or pre-translational modifications might lead to structural variability or poorly predicted structures. While pre-translational modifications may affect protein folding and stability, they may not necessarily directly impact the specificity of the protein for its target or its binding affinity. Therefore, it can be inferred that the functional elements responsible for physicochemical properties and binding are surface elements that are not inevitably engaged in the process of folding to the extent that the structure prediction becomes correlated. This indicates how inherent structural variability results in 3D folds that might, and in some cases, not determine the functional aspects hence not affecting the pharmacology or toxicology of the protein.

4.7 TrRosetta: Domains-based analysis

Nearly all the multidomain molecules (mAbs, fusion proteins, etc.), had higher prediction scores directing that both AF2 and ESMF perform well on multi-domain proteins, making single and multiple-domain molecules equally likely to have a lower

pTM score. Multi-domain molecules with longer sequence lengths tend to have larger radii of gyration, resulting in increased complexity (85). The radius of gyration is the measure of the compactness of a protein, defined as the root-mean-square distance of the constituent atoms of a molecule from its center of mass over a trajectory (86). A larger radius of gyration indicates a more extended or less compact structure, it might add up to the challenge of structure prediction for the prediction tools to model a structure accurately, resulting in lower scores. However, the results negate this hypothesis.

Proteins with a larger radius of gyration and multiple domains resulted in relatively better prediction scores (i.e., Afibercept and Tositumomab), while the proteins with a smaller radius of gyration and single domains had relatively lower pTM scores (i.e., Lepirudin, Parathyroid, Lixisenatide). The larger molecules might have features that make them easier to model accurately, such as distinctive folds, recognizable structural motifs, or simply better MSA, resulting in relatively better scores. Lastly, the prediction power of these tools also plays a vital role in determining the quality of the predicted structure.

Extending the repeated-measures analysis, monomer proteins (9 out of 14 listed in Table 4.9) with lower pTM scores were predicted through Yang Servers trRosetta (16), which was announced as a winner of a biannually held competition, CASP-15 (2022) (87). Momentous improvements in the pTM scores were seen (Table 4.9), hence backing up the inference that the accuracy of predicted protein structures increases with the prediction power of an AI-based tool and the algorithm and data used during its training. Anomalous low pTM scores for Parathyroid and Rilonecept could indicate that these proteins have more atypical or complex structures, which are not well represented by the known structures used as templates in the modeling process. These low pTM scores could indicate the presence of structurally divergent domains from already known structural data or that the domain's boundaries are difficult to determine from the available experimental data. The drastic increase in pTM scores

Table 4.9. Single and multi-domain molecules with low pTM scores predicted from AF2, ESMF, and trRosetta

Name	AF2 pTM	ESMF pTM	trRosetta pTM	Domain
Lepirudin	0.31	0.23	0.72	single
Parathyroid/Preotact	0.37	0.25	0.31	single
Rilonacept	0.42	0.33	0.45	multiple
Lixisenatide	0.43	0.41	0.54	single
Denileukin diftitox	0.44	0.38	0.76	multiple
Tagraxofusp	0.45	0.39	0.70	multiple
Elosulfase Alfa	0.46	0.49	0.52	multiple
Etanercept	0.47	0.41	0.67	multiple
Aflibercept	0.49	0.47	0.66	multiple

with prediction power and better training data indicates that it might be possible to predict complex protein structures with accuracy closer to the experimentally driven structures.

4.8 Analyzing the Learning of AF2 & ESMF based on available data

The listed PDB Blast Query coverage and Percentage Identity scores analyzed the dependence of the predictions made by AF2 and ESMF on the availability of structural and sequential data found in the PDB database. The pLDDT scores for Laronidase and Velaglucerase alfa for both predicated and AF2 Database are relatively high (pLDDT: 90+), given that the structural folds and sequences were already present in PDB as well as in UniProt databases, as supported by the percentage identity and sequence alignment identity scores. As for Alefacept and Eftrenonacog Alfa, the scores are slightly low (around 70 to 80), which is also well supported by the availability of training data being relatively less (lower query coverage, percentage identity, alignment identity, and similarity) in comparison to Laronidase and Velaglucerase alfa. For Elosulfase Alfa, the predicated pLDDT is relatively lower (49.7) in com-

Chapter 4 4.8 Analyzing the Learning of AF2 & ESMF based on available data

parison to the AF2 DB score (96.06) for the sequence acquired against Elosulfase from UniProt with 45% sequence identity. The high variation in the score is due to the differences in the rest of the 55% sequence. Another reason for a low predicted score for Elosulfase could be due to the unavailability of such combinations of amino acids in the training data of AF2. As seen before, in the case of Velaglycerase alfa, the AF2 model performs reliable predictions with higher structure and sequence identity scores. While AF2 has been trained on a vast amount of protein structure and sequence data, it is still possible for specific combinations of mutations to fall outside the ones encountered during training, as in the case of Lepirudin.

The pLDDT score from the AF2 DB for Hirudin variant-1 (UniProt ID: P01050) is 89.93. The first two residues in this variant of Hirudin are mutated to form an FDA-approved drug Lepirudin. With only two residual differences, it might seem that the predicted AF2 score has dropped to 55.10, however, when predicted from AF2 Colab Fold, Hirudin predictions result in a pLDDT score of 57.8. Compared to AF2 DB scores and structures predicted from AF2 full model, the AF2 Colab notebook uses no templates (homologous structures) and a selected portion of the BFD database. Although predictions from AF2 Colab Fold have been validated on several thousand PDB structures and the accuracy of predicted scores attained was tested to be near-identical to the full AF system on multiples targets, yet a small fraction is anticipated to cause a large drop in accuracy due to the smaller MSA and lack of templates available. This is why smaller proteins predicted from AF2 do not have reliable scores. Only two residual differences in Lepirudin caused a significant drop in the pLDDT score when predicted from Colab Fold. Keeping this in view, if AF2 and ESMF models are highly dependent on the availability of folds and sequence patterns in the training dataset, how would they perform on novel folds?

Table 4.10. BLAST PDB and UniProt compared with AF2 pLDDT scores

Name	PDB-BLAST Query Coverage (%)	PDB-BLAST Percentage Identity (%)	Sequence Alignment Identity (%)	Sequence Alignment Similarity (%)	Predicted AF2 pLDDT	UniProt- AF2 DB pLDDT
Elosulfase Alfa	99.00	62.29	45.00	45.00	49.70	96.06
Lepirudin	100.00	96.92	96.90	98.50	55.10	89.93
Alefacept	56.00	63.25	35.40	35.40	72.80	83.66
Eftrenonacog Alfa	51.00	85.89	45.50	45.50	78.80	80.05
Laronidase	100.00	99.84	95.70	95.70	95.10	95.13
Velaglucerase Alfa	100.00	100.00	92.70	92.70	97.30	93.55

4.9 Randomization & Mutation: Novel molecules predictions

After the randomization and mutation of single domains, followed by all the domains of Trastuzumab, Etanercept, Coagulation Factor VIIa, and Darbepoetin alfa, an analytical comparison was carried out. The original trastuzumab molecule had an AF2 pLDDT score of 91 and an ESMF pLDDT score of 82.01. When the first domain in the heavy chain of trastuzumab was mutated, the percentage identity dropped to 75.43% from 93.73%. When all the domains within this molecule were mutated, the AF2 pLDDT dropped to 25.20 with a percentage identity of 100% for only 3% query coverage. Clearly backing up the decrease in prediction score due to the unavailability of known fold and sequence combinations. The most interesting instances were Etanercept and Darbepoetin alfa with all the domains mutated. There were absolutely no hits for both sequences, which means that AF2 would not have any data available for MSA and probably would have never seen fold for such sequences, hence relying on its own learning. AF2, when left on its own learning, performs relatively better for smaller molecules (Darbepoetin alfa pLDDT: 40), in contrast to larger molecules (Etanercept pLDDT: 32.20). Overall, this data supports the hypothesis that AF2 and ESMF perform well with known folds and sequence combinations as compared to

Table 4.11. AF2 and ESMF prediction scores comparison for mutated single and multiple domains

Trastuzumab	Query Coverage (%)	Percentage Identity (%)	AF pLDDT	AF pTM	ESMF pLDDT	ESMF pTM
original	99.00	93.73	91.00	0.61	82.01	0.58
one-domain-mutated	99.00	75.43	79.50	0.53	71.90	0.46
all-domains-mutated	3.00	100.00	25.20	0.15	19.19	0.13
Etanercept:	Query Coverage (%)	Percentage Identity (%)	AF pLDDT	AF pTM	ESMF pLDDT	ESMF pTM
original	49.00	100.00	82.10	0.47	79.23	0.41
one-domain-mutated	37.00	100.00	68.50	0.38	68.34	0.39
all-domains-mutated	0.00	0.00	32.20	0.17	24.84	0.13
Coagulation Factor-VIIa:	Query Coverage (%)	Percentage Identity (%)	AF pLDDT	AF pTM	ESMF pLDDT	ESMF pTM
original	62.00	100.00	86.10	0.77	87.42	0.79
one-domain-mutated	37.00	100.00	48.80	0.25	43.46	0.24
all-domains-mutated	25.00	40.87	28.10	0.18	25.19	0.14
Darbepoetin alfa:	Query Coverage (%)	Percentage Identity (%)	AF pLDDT	AF pTM	ESMF pLDDT	ESMF pTM
original	86.00	95.18	87.70	0.84	83.95	0.85
domain-mutated	0.00	0.00	40.00	0.29	41.64	0.19

novel sequences.

4.10 Trastuzumab: Alanine scanning results

The results obtained from the mCSM-PPI2 analysis revealed that when the THR residue at position 129 of the Light chain in Trastuzumab (chain D) was mutated to ALA (THR129ALA), the binding affinity between Trastuzumab and chain A (the receptor tyrosine-protein kinase erbB-2) increased significantly. This indicates that the presence of THR at position 129 may weaken the interaction between Trastuzumab and the receptor. The substitution to alanine likely eliminates a potential steric hindrance or alters the hydrogen bonding pattern, resulting in improved binding affinity. These findings suggest that the specific amino acid residue at po-

sition 129 of Trastuzumab plays a critical role in its interaction with the receptor tyrosine-protein kinase erbB-2 (HER2). The observed increase in binding affinity upon mutation provides valuable insights for understanding the molecular basis of the Trastuzumab-receptor interaction and potentially guiding future therapeutic interventions. Other than this, all the point mutations resulted in an increase in binding affinity (decrease in the score) as shown in Table A.3 and Figure 4.7. A significant decrease is observed for the mutation of PHE on position 173 of the Heavy chain of trastuzumab (chain E) to ALA. Detailed scoring and description of each point mutation of Chain D and E are shown in the link provided in Appendix A.3.

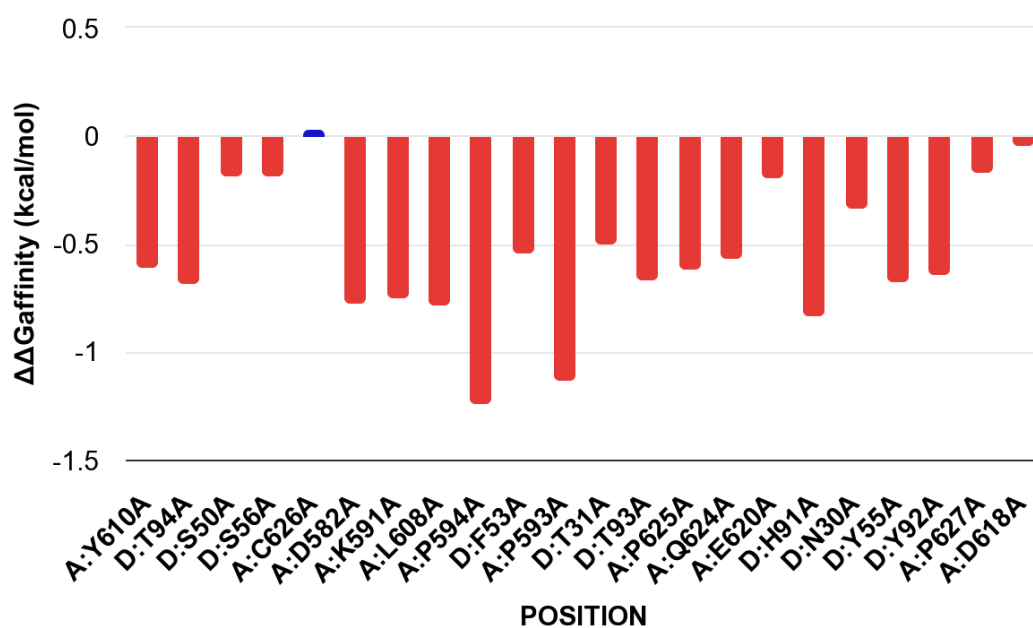


Figure 4.7. Trastuzumab-HER2 alanine Scanning Affinity Chart

Table 4.12. Alanine Scanning results for Trastuzumab-HER2

Chain	Wild-type	Residue #	Mutant	mCSM-PPI2-prediction	Affinity
E	PHE	173	ALA	-2.544	Decreasing
D	PHE	116	ALA	-2.236	Decreasing
E	ARG	50	ALA	-1.972	Decreasing
E	TRP	110	ALA	-1.942	Decreasing

DISCUSSION

The data presented in this thesis, especially for biosimilars indicates that the products that proved the safety and efficacy have high pLDDT and pTM scores. Extending these findings, Protein-based therapeutics (88) are highly successful with great potential (89). They can be categorized into five groups based on their pharmacological activity: (a) replacing an absent or abnormal protein; (b) enhancing an already-existing pathway; (c) offering a novel function or activity; (d) interfering with a molecule or organism; and (e) delivering other compounds or proteins, such as a radionuclide, cytotoxic drug, or effector protein. Additionally, they are categorized according to the molecular types they belong to, such as enzymes, growth factors, hormones, interferons, interleukins, thrombolytics, Fc fusion proteins, anticoagulants, blood factors, bone morphogenetic proteins, and engineered protein scaffolds. These compounds are further categorized according to their molecular modes of action into groups that (a) attach non-covalently to the target, such as mAb; (b) affect covalent bonds, such as enzymes; and (c) exert activity without specific contacts, such as serum albumin. Most therapeutic proteins available on the market today are made of recombinant proteins, and hundreds more are undergoing clinical trials for the treatment of infections, cancers, immunological disorders, and other diseases.

The recombinant products have grown fast with over 200 FDA-approved therapeutic protein products (90), (91). However, developing these products takes years and billions of dollars, so they are allowed a 12-year exclusivity in the market, regardless of their intellectual property status. When these exclusivities expire, biosimilars are introduced, the copies of the first-licensed recombinant therapeutic proteins to reduce patient costs (92). While there are many differences in the regulatory approval requirements globally, the primary metrics include establishing molecular biosimilarity

(often labelled as product-related attributes) and process-related attributes (9). Unfortunately, these testing costs range between USD 100 to 300 million, a significant barrier to adopting biosimilars. The primary amino acid chain is the only determinant of the 3D structure of a protein-based therapeutic (for both biosimilar and its reference product), thus the amino acid side chains are critical; charged amino acid sides can form ionic bonds, and polar amino acids can form hydrogen bonds. Weak Van der Waals interactions mediate interactions between hydrophobic side chains. These side chains primarily form non-covalent bonds. Cysteines are the only amino acids that have the ability to form covalent bonds, and they do so by utilizing their side chains. The arrangement of the amino acids of a given protein depends on side-chain interactions. Thousands of noncovalent bonds between amino acids stabilize folded proteins (93).

A faithful translation of the genetic code depends on several sequential molecular recognition events, each with an inherent error rate. The overall error rate of protein synthesis has been estimated at one misincorporated amino acid per 10⁴ codons. It reflects accumulated mistakes from all steps involved in translation (94). These error rates are dependent on the thermodynamic stability of the amino acid chain that should be projected using the AF2 and ESMFold algorithms, and other tools. How these error rates are related to the protein's functionality remains to be established.

Predicting the 3D structure using AI-based models has been challenging until AlphaFold2 (AF2) presented its ability to provide higher than 90% confidence upon repeated prediction, simulating repeated protein translation. The AF2 algorithm is based on a network-based pair-wise residue distance model. The ESMFold (ESMF) works by leveraging a large-scale language model and a 3-Dimensional equivariant of a transformer model, trRosetta, is an attention-based neural network; a cross-comparison serving as an orthogonal confirmation shows that the AF2 algorithm predicts the structure of protein domains with an accuracy matching that of experimental methods.

The question about the pathway of protein folding was posed by Levinthal in 1968 (95). An estimate of 10300 was made if a polypeptide of 100 residues will have 99 peptide bonds, and therefore 198 different phi and psi bond angles. If each of these bond angles can be in one of three stable conformations, the protein may misfold into a maximum of 3198 different conformations (including any possible folding redundancy (96)). Despite many efforts this paradox remains unsolved, though discussions of energy landscape framework following pathways that minimize the free energy along the folding funnel, where proteins navigate a rugged landscape of free energy to reach their native structure has been suggested (97), (98). Kinetic studies demonstrate that a number of small domains fold by a nucleation-condensation mechanism, in which the final or major transition state resembles an expanded version of the native structure, with numerous long interactions partially forming to stabilise an extended folding nucleus, and stronger interactions consolidating (99). A more traditional framework technique that uses the docking of prefabricated parts of regular secondary structures to create additional tiny domains (100).

In a wholly human-based model, Keil and colleagues (101) simulated the structure of trypsin based on the structure of alpha-Chymotrypsin discovered by Blow and colleagues, about the same time that Levinthal developed his paradox. According to Christian Anfinsen's research, proteins' three-dimensional structures can only be predicted using information from their main structures. As a result, the first hypotheses about protein structure were based on using a homologous protein's known, previously established structure as a template (102), (103).

With the help of 50 years of rigorous experimental work and the CASP competition, 175,000 three-dimensional structures in the Protein Data Bank are now available. With significant advances in structure prediction AlphaFold2 has proven the power of machine learning by identifying patterns in primary sequences that accurately predict three-dimensional folds (104). The core of AlphaFold2 is a neural network that has been trained on the vast majority of protein structure data in the Protein Data Bank

(105) to predict distributions of distances between pairs of residues' C β atoms and to build an artificial force field to control folding without using a specific template but rather patterns derived from many proteins (106). Sequence databases and multiple sequence alignments have also been actively utilised. Using the sequence as the experimental data, AlphaFold is arguing that if a method relies on data from multiple sequence alignments, whether directly or indirectly, it is an experimental method for determining structure, similar to X-ray crystallography, NMR spectroscopy, or electron microscopy.

This work concludes that the high confidence in the structure prediction of therapeutic proteins does not reflect stability of protein domains, or their properties to bind with receptors, leading to conclusion that these scores have little significance. Fact that over time these algorithms have improved their scores is a result of better functionality of algorithms in regurgitating a known structure. The dream of finding a thermodynamic clue to the structuring of a sequence into a 3D structure could have allowed simulation of domains that might be docked with receptors to create new therapies; and, also support arguments to reduce the testing of biosimilars if the amino acid chains are identical or highly similar.

The 3D protein structure computation and predictions heavily claimed as major breakthroughs have failed to shake the classic Levinthal paradox. These algorithms simply reproduce what is already known, adding random errors that are anticipated to be minimized in the future, yet, giving any significance to these scores will not likely bring any drug discovery, or safety evaluation of therapeutic proteins.

CONCLUSION AND FUTURE PERSPECTIVES

The newer structure prediction methods include improving pairwise and higher-order residue distance constraints from multiple sequence alignments (107), (108), and understanding how this information is eventually encoded into a predicted 3D structure (107), (106), (109), (110). These developments have been reviewed recently (111), showing how the increasing use of neural network models forms the backbone of predicting protein structures from their primary sequence. This is supported by the rise of protein sequence and structure databases (112), (113), which serve as critical resources for input and training sophisticated prediction methods.

The structural complexity of proteins depends on the number of amino acids, resulting in lesser confidence in the structure prediction, as predicted by Levinthal. This work shows that for the category of therapeutic proteins above 40 amino acids, there is a weak or no correlation between the number of amino acids and their pLDDT or pTM scores. In the case of polypeptides, it was shown a reverse observation that a smaller number provides more complexity in prediction and lesser confidence in structure predictability, as applied to polypeptides.

Both tools have a significantly positive correlation, representing their orthogonality. The finding of this paper suggests that the predictions based on sequence alone can not be used to describe the folding of a structure and its accuracy. The pLDDT and pTM scores do not correlate with any structural or functional parameters and hence they cannot be used to determine proteins stability in in-vivo systems or propose concentration and dosing for better efficacy, neither can they be correlated with the binding properties of proteins even though they are all dependent upon the amino

acid sequence. While pre-translational modifications may affect protein folding and stability, they may not necessarily directly impact the specificity of the protein for its target or its binding affinity. Hence, it can be concluded that the surface elements responsible for the physicochemical properties and binding are not necessarily involved in the folding process to a degree that correlates with structure prediction therefore not affecting the pharmacology or toxicology of the protein.

Even though AF2 has a slight tendency to misfold structures, it performs reliable predictions on multidomain molecules. However, these predictions can be significantly improved with better prediction power tools like trRosetta, hence concluding that it might be possible to predict complex protein structures with an accuracy closer to the experimental structures in the near future with more robust prediction tools.

Few FDA and EMA-approved biosimilars demonstrated pLDDT scores greater than 80 using the AF2 predictions, thus it can be concurred that there is less variability of the 3D structure, and these molecules may not require extensive testing to establish molecular biosimilarity. Extending this argument, 188 proteins were rank-ordered in the context of structural variability using the AF2 pLDDT score. Using these ranks as evidence of higher structural stability, reduced testing of biosimilars that were prepared from similar amino acid sequences to their reference product can be proposed. While the future algorithms may bring better predictability, given the innate nature of translation, the variability caused by the thermodynamic instability is unlikely to raise these scores very high or change the relative ranking of therapeutic proteins. For now, it can be concluded that high confidence in the structure predictability assures a 3D structure will be similar between a biosimilar candidate and its reference product. Future studies may include structure prediction of post-translational modifications (114), further reducing the testing to establish the biosimilarity of biological drugs (115). The training of AF2 and ESMF models is highly dependent on structural and sequential data from multiple databases i.e., PDB and UniProt. If these models have encountered a certain fold in their training set, they are more likely to provide an

accurate prediction. In order to analyse the effect of availability of structural and sequential data, BLAST against PDB and UniProt databases for FDA approved therapeutic proteins from the dataset was employed.

In the future, the randomized and mutated sequences can be predicted from trRosetta to include it in the orthogonal comparison and conclude if the prediction power of the model affects the folds prediction. Additionally, further analysis is required to find which amino acid contributes the most to destabilizing the overall protein structure and fold formation. Besides this, testing whether the lack of availability of training data and/or the prediction power of AI-based tools affects the modeling of protein regions referred to as loops. Loop modeling still remains unresolved and hence requires in-depth in-silico analysis.

REFERENCES

- [1] D. S. Dimitrov, “Therapeutic proteins,” *Therapeutic Proteins: Methods and Protocols*, pp. 1–26, 2012.
- [2] G. P. Adams and L. M. Weiner, “Monoclonal antibody therapy of cancer,” *Nature biotechnology*, vol. 23, no. 9, pp. 1147–1157, 2005.
- [3] A. L. Erwin, “The role of sebelipase alfa in the treatment of lysosomal acid lipase deficiency,” *Therapeutic advances in gastroenterology*, vol. 10, no. 7, pp. 553–562, 2017.
- [4] S. Barrientos, O. Stojadinovic, M. S. Golinko, H. Brem, and M. Tomic-Canic, “Growth factors and cytokines in wound healing,” *Wound repair and regeneration*, vol. 16, no. 5, pp. 585–601, 2008.
- [5] G. Grampp and S. Ramanan, “The diversity of biosimilar design and development: implications for policies and stakeholders,” *BioDrugs*, vol. 29, pp. 365–372, 2015.
- [6] M. L. Davis-Ajami, J. Wu, K. Downton, E. Ludeman, and V. Noxon, “Epoetin zeta in the management of anemia associated with chronic kidney disease, differential pharmacology and clinical utility,” *Biologics: Targets and Therapy*, pp. 155–167, 2014.
- [7] R. Zwanzig, J. Kiefer, and G. Weiss, “Proc. natl. acad. sci.(us),” 1966.
- [8] T. Schmidt, A. Bergner, and T. Schwede, “Modelling three-dimensional protein structures for applications in drug design,” *Drug discovery today*, vol. 19, no. 7, pp. 890–897, 2014.
- [9] S. K. Niazi, “Molecular biosimilarity—an ai-driven paradigm shift,” *International Journal of Molecular Sciences*, vol. 23, no. 18, p. 10690, 2022.
- [10] S. K. Niazi and Z. Mariam, “Reinventing therapeutic proteins: Mining a treasure of new therapies,” *Biologics*, vol. 3, no. 2, pp. 72–94, 2023.
- [11] A. Kouranov, L. Xie, J. de la Cruz, L. Chen, J. Westbrook, P. E. Bourne, and H. M. Berman, “The rcsb pdb information portal for structural genomics,” *Nucleic acids research*, vol. 34, no. suppl_1, pp. D302–D305, 2006.
- [12] U. Consortium, “Uniprot: a worldwide hub of protein knowledge,” *Nucleic acids research*, vol. 47, no. D1, pp. D506–D515, 2019.
- [13] M. Mirdita, L. Von Den Driesch, C. Galiez, M. J. Martin, J. Söding, and M. Steinegger, “Uniclust databases of clustered and deeply annotated protein sequences and alignments,” *Nucleic acids research*, vol. 45, no. D1, pp. D170–D176, 2017.

-
- [14] A. L. Mitchell, A. Almeida, M. Beracochea, M. Boland, J. Burgin, G. Cochrane, M. R. Crusoe, V. Kale, S. C. Potter, L. J. Richardson, *et al.*, “Mgnify: the microbiome analysis resource in 2020,” *Nucleic acids research*, vol. 48, no. D1, pp. D570–D578, 2020.
- [15] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, “Highly accurate protein structure prediction with alphafold,” *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.
- [16] Z. Du, H. Su, W. Wang, L. Ye, H. Wei, Z. Peng, I. Anishchenko, D. Baker, and J. Yang, “The trRosetta server for fast and accurate protein structure prediction,” *Nature protocols*, vol. 16, no. 12, pp. 5634–5651, 2021.
- [17] D. E. Kim, D. Chivian, and D. Baker, “Protein structure prediction and analysis using the rosetta server,” *Nucleic acids research*, vol. 32, no. suppl_2, pp. W526–W531, 2004.
- [18] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, *et al.*, “Accurate prediction of protein structures and interactions using a three-track neural network,” *Science*, vol. 373, no. 6557, pp. 871–876, 2021.
- [19] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, *et al.*, “Language models of protein sequences at the scale of evolution enable accurate structure prediction,” *BioRxiv*, 2022.
- [20] J. Peng and J. Xu, “Raptorx: exploiting structure information for protein alignment by statistical inference,” *Proteins: Structure, Function, and Bioinformatics*, vol. 79, no. S10, pp. 161–171, 2011.
- [21] R. Wu, F. Ding, R. Wang, R. Shen, X. Zhang, S. Luo, C. Su, Z. Wu, Q. Xie, B. Berger, *et al.*, “High-resolution de novo structure prediction from primary sequence,” *BioRxiv*, pp. 2022–07, 2022.
- [22] S. Wang, W. Li, S. Liu, and J. Xu, “Raptorx-property: a web server for protein structure property prediction,” *Nucleic acids research*, vol. 44, no. W1, pp. W430–W435, 2016.
- [23] M. van Breugel, I. Rosa e Silva, and A. Andreeva, “Structural validation and assessment of alphafold2 predictions for centrosomal and centriolar proteins and their complexes,” *Communications Biology*, vol. 5, no. 1, p. 312, 2022.
- [24] K.-T. Ko, F. Lennartz, D. Mekhail, B. Guloglu, A. Marini, D. J. Deuker, C. A. Long, M. M. Jore, K. Miura, S. Biswas, *et al.*, “Structure of the malaria vaccine candidate pfs48/45 and its recognition by transmission blocking antibodies,” *Nature Communications*, vol. 13, no. 1, p. 5603, 2022.

-
- [25] H.-B. Guo, A. Perminov, S. Bekele, G. Kedziora, S. Farajollahi, V. Varaljay, K. Hinkle, V. Molinero, K. Meister, C. Hung, *et al.*, “Alphafold2 models indicate that protein sequence determines both structure and dynamics,” *Scientific Reports*, vol. 12, no. 1, p. 10696, 2022.
- [26] L. Hu, W. Salmen, B. Sankaran, Y. Lasanajak, D. F. Smith, S. E. Crawford, M. K. Estes, and B. V. Prasad, “Novel fold of rotavirus glycan-binding domain predicted by alphafold2 and determined by x-ray crystallography,” *Communications biology*, vol. 5, no. 1, p. 419, 2022.
- [27] P. Fontana, Y. Dong, X. Pi, A. B. Tong, C. W. Hecksel, L. Wang, T.-M. Fu, C. Bustamante, and H. Wu, “Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-em and alphafold,” *Science*, vol. 376, no. 6598, p. eabm9326, 2022.
- [28] J. S. Y. Tan, B. Lee, J. Lim, D. R. Ma, J. X. Goh, S. Y. Goh, M. Y. Gulam, S. M. Koh, W. W. Lee, L. Feng, *et al.*, “Parkinson’s disease-specific autoantibodies against the neuroprotective co-chaperone stip1,” *Cells*, vol. 11, no. 10, p. 1649, 2022.
- [29] T. Hegedűs, M. Geisler, G. L. Lukács, and B. Farkas, “Ins and outs of alphafold2 transmembrane protein structure predictions,” *Cellular and Molecular Life Sciences*, vol. 79, no. 1, p. 73, 2022.
- [30] G. Tusnányi, Z. Dosztányi, and I. Simon, “Tmdet: web server for detecting transmembrane domains by using 3d structure of proteins,” *Bioinformatics*, vol. 21, no. 7, p. 1, 2005.
- [31] T. G. Flower and J. H. Hurley, “Crystallographic molecular replacement using an in silico-generated search model of sars-cov-2 orf8,” *Protein Science*, vol. 30, no. 4, pp. 728–734, 2021.
- [32] J. P. Roney and S. Ovchinnikov, “State-of-the-art estimation of protein model accuracy using alphafold,” *Physical Review Letters*, vol. 129, no. 23, p. 238101, 2022.
- [33] L. Chang and A. Perez, “Alphafold encodes the principles to identify high affinity peptide binders,” *BioRxiv*, pp. 2022–03, 2022.
- [34] A. Goulet and C. Cambillau, “Present impact of alphafold2 revolution on structural biology, and an illustration with the structure prediction of the bacteriophage j-1 host adhesion device,” *Frontiers in Molecular Biosciences*, vol. 9, 2022.
- [35] D. Chakravarty and L. L. Porter, “Alphafold2 fails to predict protein fold switching,” *Protein Science*, vol. 31, no. 6, p. e4353, 2022.
- [36] L. M. Bertoline, A. N. Lima, J. E. Krieger, and S. K. Teixeira, “Before and after alphafold2: An overview of protein structure prediction,” *Frontiers in Bioinformatics*, vol. 3, 2023.
-

-
- [37] M. A. Pak and D. N. Ivankov, “Best templates outperform homology models in predicting the impact of mutations on protein stability,” *Bioinformatics*, vol. 38, no. 18, pp. 4312–4320, 2022.
- [38] A. G. Street and S. L. Mayo, “Computational protein design,” *Structure*, vol. 7, no. 5, pp. R105–R109, 1999.
- [39] “Rational design of a structural and functional nitric oxide reductase - PubMed — pubmed.ncbi.nlm.nih.gov.” <https://pubmed.ncbi.nlm.nih.gov/19940850/>. [Accessed 05-Jun-2023].
- [40] S. Khan and M. Vihinen, “Performance of protein stability predictors,” *Human mutation*, vol. 31, no. 6, pp. 675–684, 2010.
- [41] O. Calderaru, T. L. Blundell, and K. P. Kepp, “A base measure of precision for protein stability predictors: structural sensitivity,” *BMC bioinformatics*, vol. 22, pp. 1–14, 2021.
- [42] Y. Zhang, P. Li, F. Pan, H. Liu, P. Hong, X. Liu, and J. Zhang, “Applications of alphafold beyond protein structure prediction,” *bioRxiv*, pp. 2021–11, 2021.
- [43] M. A. Pak, K. A. Markhieva, M. S. Novikova, D. S. Petrov, I. S. Vorobyev, E. S. Maksimova, F. A. Kondrashov, and D. N. Ivankov, “Using alphafold to predict the impact of single mutations on protein stability and function,” *Plos one*, vol. 18, no. 3, p. e0282689, 2023.
- [44] “AUROC x2014; PyTorch-Metrics 0.11.4 documentation — https,” [Accessed 05-Jun-2023].
- [45] P.-S. Huang, S. E. Boyken, and D. Baker, “The coming of age of de novo protein design,” *Nature*, vol. 537, no. 7620, pp. 320–327, 2016.
- [46] P. R. Patil, A. M. Burroughs, M. Misra, F. Cerullo, I. Dikic, L. Aravind, and C. A. Joazeiro, “Mechanism and evolutionary origins of alanine-tail c-degron recognition by e3 ligases pirh2 and crl2-klhdc10,” *bioRxiv*, pp. 2023–05, 2023.
- [47] C. J. Wilson, W.-Y. Choy, and M. Karttunen, “Alphafold2: a role for disordered protein/region prediction?,” *International Journal of Molecular Sciences*, vol. 23, no. 9, p. 4591, 2022.
- [48] “THPdb: A Database of FDA approved Therapeutic Peptides and Proteins — webs.iiitd.edu.in.” <https://webs.iiitd.edu.in/raghava/thpdb/length.php>. [Accessed 06-March-2023].
- [49] R. D. Kelly and Y. Onoe, “The fda purple book: Black and blue all over,” *Biotechnology Law Report*, vol. 34, no. 3, pp. 75–80, 2015.
- [50] U. Food and D. Administration, “Orange book: approved drug products with therapeutic equivalence evaluations,” 2013.
-

-
- [51] V. B. Siramshetty, I. Grishagin, Đ.-T. Nguyn, T. Peryea, Y. Skovpen, O. Stroganov, D. Katzel, T. Sheils, A. Jadhav, E. A. Mathé, *et al.*, “Ncats inxight drugs: a comprehensive and curated portal for translational research,” *Nucleic Acids Research*, vol. 50, no. D1, pp. D1307–D1316, 2022.
- [52] M. Kanehisa, “The kegg database,” in *‘In Silico’ Simulation of Biological Processes: Novartis Foundation Symposium 247*, vol. 247, pp. 91–103, Wiley Online Library, 2002.
- [53] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, “Drugbank: a knowledgebase for drugs, drug actions and drug targets,” *Nucleic acids research*, vol. 36, no. suppl_1, pp. D901–D906, 2008.
- [54] “U.S. Food and Drug Administration.” Accessed 06-Jun-2023.
- [55] “Molecular Weight Calculator — cusabio.com.” <https://www.cusabio.com/m-299.html#:~:text=Note%3A%20The%20molecular%20weight%20of,of%20a%205%27phosphate%20residue.&text=Note%3A%20The%20molecular%20weight%20of%20A%2C%20U%2C%20C%2C,and%20345.2%20g%20Fmol%20respectively.> [Accessed 06-Jun-2023].
- [56] E. Bisong and E. Bisong, “Google colaboratory,” *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*, pp. 59–64, 2019.
- [57] E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, and T. E. Ferrin, “Ucsf chimeraX: Structure visualization for researchers, educators, and developers,” *Protein Science*, vol. 30, no. 1, pp. 70–82, 2021.
- [58] R. R. B. H. Z. Z. W. L. A. d. S. C. M. F.-Z. T. S. S. C. A. R. Zeming Lin, Halil Akin, “ESMFold Notebook.” <https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/ESMFold.ipynb#scrollTo=CcyNpAvhTX6q>, 2022. Accessed: 06-Jun-2023.
- [59] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, and A. Bairoch, *Protein identification and analysis tools on the ExPASy server*. Springer, 2005.
- [60] C. Christoffer, V. Bharadwaj, R. Luu, and D. Kihara, “Lzrd protein-protein docking webserver enhanced with de novo structure prediction,” *Frontiers in molecular biosciences*, vol. 8, p. 724947, 2021.
- [61] J. Janin, K. Henrick, J. Moult, L. T. Eyck, M. J. Sternberg, S. Vajda, I. Vakser, and S. J. Wodak, “Capri: a critical assessment of predicted interactions,” *Proteins: Structure, Function, and Bioinformatics*, vol. 52, no. 1, pp. 2–9, 2003.
- [62] M. F. Lensink, S. Velankar, M. Baek, L. Heo, C. Seok, and S. J. Wodak, “The challenge of modeling protein assemblies: the casp12-capri experiment,” *Proteins: Structure, Function, and Bioinformatics*, vol. 86, pp. 257–273, 2018.

-
- [63] M. F. Lensink, G. Brysbaert, N. Nadzirin, S. Velankar, R. A. Chaleil, T. Gerguri, P. A. Bates, E. Laine, A. Carbone, S. Grudinin, *et al.*, “Blind prediction of homo- and hetero-protein complexes: The casp13-capri experiment,” *Proteins: Structure, Function, and Bioinformatics*, vol. 87, no. 12, pp. 1200–1221, 2019.
- [64] M. F. Lensink, N. Nadzirin, S. Velankar, and S. J. Wodak, “Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: Capri 7th edition,” *Proteins: Structure, Function, and Bioinformatics*, vol. 88, no. 8, pp. 916–938, 2020.
- [65] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, “Ucsf chimera—a visualization system for exploratory research and analysis,” *Journal of computational chemistry*, vol. 25, no. 13, pp. 1605–1612, 2004.
- [66] H. Zhou and J. Skolnick, “Goap: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction,” *Biophysical journal*, vol. 101, no. 8, pp. 2043–2052, 2011.
- [67] S.-Y. Huang and X. Zou, “Itscorepro: an efficient scoring program for evaluating the energy scores of protein structures for structure prediction,” *Protein Structure Prediction*, pp. 71–81, 2014.
- [68] L. C. Xue, J. P. Rodrigues, P. L. Kastiris, A. M. Bonvin, and A. Vangone, “Prodigy: a web server for predicting the binding affinity of protein–protein complexes,” *Bioinformatics*, vol. 32, no. 23, pp. 3676–3678, 2016.
- [69] Y. J. Huang, B. Mao, J. M. Aramini, and G. T. Montelione, “Assessment of template-based protein structure predictions in casp10,” *Proteins: Structure, Function, and Bioinformatics*, vol. 82, pp. 43–56, 2014.
- [70] V. Modi, Q. Xu, S. Adhikari, and R. L. Dunbrack Jr, “Assessment of template-based modeling of protein structure in casp11,” *Proteins: Structure, Function, and Bioinformatics*, vol. 84, pp. 200–220, 2016.
- [71] “NCBI Conserved Domain Search — https.” <https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>. [Accessed 10-Jun-2023].
- [72] A. P. S. Database, “AlphaFold Protein Structure Database — alphafold.ebi.ac.uk.” <https://alphafold.ebi.ac.uk/>. [Accessed 10-Jun-2023].
- [73] “MOLBIOTOOLS - Molecular Biology Online Apps — molbiotools.com.” <https://molbiotools.com>. [Accessed 10-Jun-2023].
- [74] C. H. Rodrigues, Y. Myung, D. E. Pires, and D. B. Ascher, “mcsm-ppi2: predicting the effects of mutations on protein–protein interactions,” *Nucleic acids research*, vol. 47, no. W1, pp. W338–W344, 2019.
- [75] M. Manhart and A. V. Morozov, “Protein folding and binding can emerge as evolutionary spandrels through structural coupling,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 6, pp. 1797–1802, 2015.
-

-
- [76] A. Patil and H. Nakamura, “The role of charged surface residues in the binding ability of small hubs in protein-protein interaction networks,” *Biophysics*, vol. 3, pp. 27–35, 2007.
- [77] E. C. Hilario, A. Stern, C. H. Wang, Y. W. Vargas, C. J. Morgan, T. E. Swartz, and T. W. Patapoff, “An improved method of predicting extinction coefficients for the determination of protein concentration,” *PDA journal of pharmaceutical science and technology*, vol. 71, no. 2, pp. 127–135, 2017.
- [78] H. Yuan, Z. Li, X. Wang, and R. Qi, “Photodynamic antimicrobial therapy based on conjugated polymers,” *Polymers*, vol. 14, no. 17, p. 3657, 2022.
- [79] S. Gill, “C. & von hippel, ph (1989),” *Anal. Biochem*, vol. 182, pp. 319–326.
- [80] G. Rigi, G. Kardar, A. Hajizade, J. Zamani, and G. Ahmadian, “The effects of a truncated form of staphylococcus aureus protein a (spa) on the expression of cytokines of autoimmune patients and healthy individuals,” 2022.
- [81] A. O. Stevens and Y. He, “Benchmarking the accuracy of alphafold 2 in loop structure prediction,” *Biomolecules*, vol. 12, no. 7, p. 985, 2022.
- [82] R. W. Cheloha, S. H. Gellman, J.-P. Vilardaga, and T. J. Gardella, “Pth receptor-1 signalling—mechanistic insights and therapeutic prospects,” *Nature Reviews Endocrinology*, vol. 11, no. 12, pp. 712–724, 2015.
- [83] P. L. Kastritis, J. P. Rodrigues, G. E. Folkers, R. Boelens, and A. M. Bonvin, “Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface,” *Journal of molecular biology*, vol. 426, no. 14, pp. 2632–2652, 2014.
- [84] M. M. Gromiha, K. Yokota, and K. Fukui, “Energy based approach for understanding the recognition mechanism in protein–protein complexes,” *Molecular BioSystems*, vol. 5, no. 12, pp. 1779–1786, 2009.
- [85] Q.-Y. Tang, W. Ren, J. Wang, and K. Kaneko, “The statistical trends of protein evolution: a lesson from alphafold database,” *Molecular Biology and Evolution*, vol. 39, no. 10, p. msac197, 2022.
- [86] L. Mi, N. Bogatyreva, and O. Galzitskaia, “Radius of gyration is indicator of compactness of protein structure,” *Molekuliarnaia biologii*, vol. 42, no. 4, pp. 701–706, 2008.
- [87] “Groups Analysis: zscores - CASP15 — https.” https://predictioncenter.org/casp15/zscores_final.cgi. [Accessed 10-Jun-2023].
- [88] “Therapeutic Biologics Applications (BLA) — fda.gov.” <https://www.fda.gov/drugs/types-applications/therapeutic-biologics-applications-bla>. [Accessed 10-Jun-2023].
- [89] D. Ds, “Therapeutic proteins,” *Therapeutic Proteins. Methods in Molecular Biology (Methods and Protocols)*, 2012.
-

-
- [90] S. S. Usmani, G. Bedi, J. S. Samuel, S. Singh, S. Kalra, P. Kumar, A. A. Ahuja, M. Sharma, A. Gautam, and G. P. Raghava, “Thpdb: Database of fda-approved peptide and protein therapeutics,” *PloS one*, vol. 12, no. 7, p. e0181748, 2017.
- [91] P. Book, “Lists of licensed biological products with reference product exclusivity and biosimilarity or interchangeability evaluations,” *Food and Drug Administration*. Available from: <https://www.fda.gov/drugs/therapeutic-biologics-applications-bla/purple-book-lists-licensed-biological-products-reference-product-exclusivity-and-biosimilarity-or>, 2014.
- [92] S. K. Niazi, “The coming of age of biosimilars: a personal perspective,” *Biologics*, vol. 2, no. 2, pp. 107–127, 2022.
- [93] “Three imperatives for RD in biosimilars — https.” <https://www.mckinsey.com/industries/life-sciences/our-insights/three-imperatives-for-r-and-d-in-biosimilars>. [Accessed 10-Jun-2023].
- [94] N. M. Reynolds, B. A. Lazazzera, and M. Ibba, “Cellular mechanisms that control mistranslation,” *Nature Reviews Microbiology*, vol. 8, no. 12, pp. 849–856, 2010.
- [95] C. Levinthal, “Are there pathways for protein folding?,” *Journal de chimie physique*, vol. 65, pp. 44–45, 1968.
- [96] C. Levinthal, “How to fold graciously,” *Mossbauer spectroscopy in biological systems*, vol. 67, pp. 22–24, 1969.
- [97] K. A. Dill and J. L. MacCallum, “The protein-folding problem, 50 years on,” *science*, vol. 338, no. 6110, pp. 1042–1046, 2012.
- [98] O. J. Nelson and G. Wolynes Peter, “Theory of protein folding,” *Curr Opin Struct Biol*, vol. 14, no. 1, pp. 70–5, 2004.
- [99] L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, “The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding,” *Journal of molecular biology*, vol. 254, no. 2, pp. 260–288, 1995.
- [100] S. Gianni, N. R. Guydosh, F. Khan, T. D. Caldas, U. Mayor, G. W. White, M. L. DeMarco, V. Daggett, and A. R. Fersht, “Unifying features in protein-folding mechanisms,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 23, pp. 13286–13291, 2003.
- [101] B. Keil, V. Dlouha, V. Holeyšovský, and F. Šorm, “Hypothesis of three-dimensional arrangement of polypeptide chain in trypsin,” *Collection of Czechoslovak Chemical Communications*, vol. 33, no. 7, pp. 2307–2315, 1968.
- [102] B. W. Matthews, P. Sigler, R. Henderson, and D. Blow, “Three-dimensional structure of tosyl- α -chymotrypsin,” *Nature*, vol. 214, no. 5089, pp. 652–656, 1967.
-

-
- [103] C. B. Anfinsen, “Principles that govern the folding of protein chains,” *Science*, vol. 181, no. 4096, pp. 223–230, 1973.
- [104] “Home - Prediction Center — predictioncenter.org.” <https://predictioncenter.org/>. [Accessed 10-Jun-2023].
- [105] R. P. D. Bank, “RCSB PDB: Homepage — https.” <https://www.rcsb.org/>. [Accessed 10-Jun-2023].
- [106] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland, *et al.*, “Improved protein structure prediction using potentials from deep learning,” *Nature*, vol. 577, no. 7792, pp. 706–710, 2020.
- [107] J. Thomas, N. Ramakrishnan, and C. Bailey-Kellogg, “Graphical models of residue coupling in protein families,” in *Proceedings of the 5th international workshop on Bioinformatics*, pp. 12–20, 2005.
- [108] S. D. Dunn, L. M. Wahl, and G. B. Gloor, “Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction,” *Bioinformatics*, vol. 24, no. 3, pp. 333–340, 2008.
- [109] J. Xu, “Distance-based protein folding powered by deep learning,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 34, pp. 16856–16865, 2019.
- [110] M. AlQuraishi, “End-to-end differentiable learning of protein structure,” *Cell systems*, vol. 8, no. 4, pp. 292–301, 2019.
- [111] M. AlQuraishi, “Machine learning in protein structure prediction,” *Current opinion in chemical biology*, vol. 65, pp. 1–8, 2021.
- [112] A. Godzik, “Metagenomics and the protein universe,” *Current opinion in structural biology*, vol. 21, no. 3, pp. 398–403, 2011.
- [113] “Protein data bank: the single global archive for 3d macromolecular structure data,” *Nucleic acids research*, vol. 47, no. D1, pp. D520–D528, 2019.
- [114] D. V. Laurents, “Alphafold 2 and nmr spectroscopy: Partners to understand protein structure, dynamics and function,” *Frontiers in molecular biosciences*, 2022.
- [115] D. P. Ismi, R. Pulungan, *et al.*, “Deep learning for protein secondary structure prediction: Pre and post-alphafold,” *Computational and Structural Biotechnology Journal*, 2022.

Proteins and Peptides Complete Data - Github URL

The field of therapeutics encompasses a wide range of molecules that are used for medicinal purposes. The U.S. Food and Drug Administration (FDA) plays a crucial role in ensuring the safety and efficacy of these therapeutic molecules. To facilitate their regulatory processes, the FDA classifies therapeutics into distinct categories. One such classification is based on the molecule's composition and structure, explicitly dividing them into two main classes: proteins and peptides. In this categorization, molecules with an amino acid sequence length below 40 are considered peptides rather than therapeutic proteins. This classification helps distinguish and analyze these molecules based on their unique characteristics and potential therapeutic applications. For a comprehensive analysis, data for proteins and peptides have been compiled into separate data sets, allowing for a detailed examination of their properties and attributes.

A.1 Proteins Data

Protein therapeutics represent a significant portion of the FDA-regulated molecules used in medical treatments. The data set dedicated to protein therapeutics consists of 188 unique molecules. Each entry in the data set provides valuable information, including the name, accession number, BLA/NDA status (Biologics License Application/New Drug Application), brand name, and the type of molecule. Additionally, essential characteristics such as the amino acid count, molecular weight, and amino acid sequence are included. To further aid in understanding these proteins, scores from advanced structural prediction algorithms, AlphaFold2 and ESMFold are provided. These scores, such as AlphaFold pLDDT Score and AlphaFold pTM Score, assess the structural reliability and protein domain predictions, respectively. Other properties such as hydrophobicity (GRAVY), isoelectric point, extinction coefficients, and instability index contribute to a comprehensive analysis of protein therapeutics.

Link: Please refer to the [proteins data](#) for more information.

A.2 Peptides Data

Peptide therapeutics constitute a distinct subset within the realm of therapeutic molecules, characterized by their relatively shorter amino acid sequences. The data set dedicated to peptides comprises 16 molecules that fall under this category. These peptides, with their unique characteristics and potential therapeutic applications, are of great interest to researchers and clinicians. Similar to the protein data set, each entry provides relevant details, including the name, accession number, BLA/NDA status, brand name, and molecule type. Additionally, important information such as

amino acid count, molecular weight, and amino acid sequence is available for analysis. While peptides generally exhibit shorter sequences, they can still possess diverse properties crucial for their therapeutic potential. By examining properties such as AlphaFold2 and ESMFold scores, hydrophobicity (GRAVY), isoelectric point, extinction coefficients, and instability index, a comprehensive understanding of peptide therapeutics can be obtained.

Link: Please refer to the [peptides data](#) for more information.

A.3 Alanine Scanning Results

The results of the alanine scanning analysis conducted using the mCSM-PPI2 tool are present in the link below. The scanning focused on the Trastuzumab and receptor tyrosine-protein kinase erbB-2 complex, specifically examining the effects of alanine mutations on the binding affinity. Detailed information on the mutations and their corresponding changes in binding strength are provided. These results contribute to the understanding of the critical residues involved in the Trastuzumab-receptor interaction and shed light on the key determinants of binding affinity.

Link: Please refer to the [alanine scanning data](#) for more information.

These datasets for proteins and peptides as well as the data of alanine scanning provide a valuable resource for further analysis and investigation into their roles in therapeutic applications and molecule design. By examining their unique characteristics and properties, researchers and clinicians can gain insights into their potential efficacy, safety, and mechanisms of action.

Appendix B

Biosimilars - Rank order

Table B.1.
Rank order from pLDDT score AF2

Rank#	Name	pLDDT (AF)			
1	Asparaginase erwinia chrysanthemi	97.40	38	Panitumumab	91.90
2	Alglucerase	97.30	39	Emicizumab	91.90
3	Velaglucerase alfa	97.30	40	Pembrolizumab	91.90
4	Dornase alfa	97.20	41	Olipudase alfa	91.90
5	Aprotinin	97.00	42	Tocilizumab	91.70
6	Rasburicase	96.40	43	Fremanezumab	91.60
7	Ocriplasmin	96.40	44	Eculizumab	91.60
8	Taliglucerase alfa	96.20	45	Mepolizumab	91.60
9	Pancrelipase amylase	96.10	46	Ixekizumab	91.60
10	Digoxin Immune Fab (Ovine)	95.70	47	Gemtuzumab ozogamicin	91.60
11	Palivizumab	95.50	48	Alirocumab	91.60
12	Sacrosidase	95.40	49	Ipilimumab	91.50
13	Galsulfase	95.30	50	Pertuzumab	91.50
14	Laronidase	95.10	51	Elotuzumab	91.50
15	Infliximab	95.00	52	Nivolumab	91.50
16	Palifermin	94.90	53	Adalimumab	91.50
17	Idarucizumab	94.90	54	Ramucirumab	91.40
18	Vestronidase Alfa	94.80	55	Tildrakizumab	91.40
19	Bevacizumab	94.60	56	Reslizumab	91.40
20	Avalglucosidase alfa	94.60	57	Isatuximab	91.40
21	Ecallantide	94.50	58	Burosumab	91.40
22	Ustekinumab	94.50	59	Aducanumab	91.40
23	Ranibizumab	94.50	60	Mogamulizumab	91.40
24	L-asparaginase	93.60	61	Risankizumab	91.30
25	Glucarpidase	93.50	62	Ravulizumab	91.30
26	Ofatumumab	93.50	63	Obiltoxaximab	91.30
27	Interferon beta-1a	93.00	64	Bezlotoxumab	91.30
28	Thyrotropin Alfa	92.90	65	Eptinezumab	91.30
29	Sebelipase alfa	92.70	66	Odesivimab	91.30
30	Belimumab	92.60	67	Rituximab	91.30
31	Cerliponase alfa	92.60	68	Evinacumab	91.30
32	Idursulfase	92.60	69	Ansuvimab	91.30
33	Asfotase alfa	92.50	70	Cemiplimab	91.20
34	Brodalumab	92.40	71	Canakinumab	91.20
35	Romosozumab	92.10	72	Polatuzumab vedotin	91.20
36	Satralizumab	92.10	73	Sacituzumab govitecan	91.10
37	Crizanlizumab	91.90	74	Dostarlimab	91.10
			75	Anifrolumab	91.10

Rank#	Name	pLDDT (AF)
76	Efgartigimod alfa	91.10
77	Hyaluronidase (Ovine)	91.10
78	Margetuximab	91.00
79	Siltuximab	91.00
80	Trastuzumab	91.00
81	Ocrelizumab	91.00
82	Atezolizumab	91.00
83	Belantamab mafodotin	91.00
84	Obinutuzumab	91.00
85	Natalizumab	91.00
86	Mirvetuximab Soravtansine	91.00
87	Inebilizumab	91.00
88	Catridecacog	91.00
89	Maftivimab	90.90
90	Erenumab	90.90
91	Cetuximab	90.90
92	Benralizumab	90.90
93	Daratumumab	90.90
94	Atoltivimab	90.90
95	Loncastuximab tersirine	90.90
96	Becaplermin	90.90
97	Secukinumab	90.80
98	Muromonab	90.80
99	Tralokinumab	90.70
100	Sarilumab	90.70
101	Necitumumab	90.70
102	Tafasitamab	90.70
103	Teprotumumab	90.70
104	Golimumab	90.70
105	Galcaneuzumab	90.67
106	Naxitamab	90.60
107	Tezepelumab	90.50
108	Basiliximab	90.50
109	Denosumab	90.50
110	Vedolizumab	90.50
111	Dupilumab	90.50
112	Durvalumab	90.40
113	Urokinase	90.40
114	Olaratumab	90.30
115	Guselkumab	90.20
116	Dinutuximab	90.20

117	Alemtuzumab	90.20
118	Alpha-1-proteinase inhibitor	90.20
119	Filgrastim	90.20
120	Caplacizumab	90.10
121	Evolocumab	90.10
122	Ibalizumab	90.10
123	Sargramostim	90.10
124	Emapalumab	89.90
125	Lanadelumab	89.90
126	Chymopapain	89.70
127	Avelumab	89.60
128	Follitropin	89.60
129	Aflibercept	89.10
130	Reteplase	89.10
131	Iuspatercept-aamt	89.00
132	Urofollitropin	89.00
133	Brolucizumab	88.90
134	Anakinra	88.40
135	Daclizumab	88.20
136	Ibritumomab tiuxetan	88.20
137	Belatacept	87.80
138	Interferon beta-1b	87.70
139	Darbepoetin alfa	87.70
140	Erythropoietin	87.70
141	Omalizumab	87.60
142	Albiglutide	87.60
143	Aldesleukin	87.60
144	Drotrecogin alfa	87.50
145	Abatacept	87.40
146	Insulin degludec	87.40
147	Insulin detemir	87.40
148	Antithrombin Alfa	87.40
149	Interferon Alfacon 1	87.30
150	Eflapegrastim	87.00
151	Insulin Regular	87.00
152	Interferon alfa-2a	87.00
153	Tositumomab	86.60
154	Insulin aspart	86.30
155	Rilonacept	86.20
156	Tenecteplase	86.20
157	Oprelvekin	86.20
158	Coagulation Factor VIIa	86.10

Rank#	Name	pLDDT (AF)
159	Alteplase	86.00
160	Insulin lispro	86.00
161	Interferon alfa-2b	86.00
162	Insulin glargine	85.60
163	Insulin glulisine	85.50
164	Blinatumomab	85.40
165	Interferon gamma-1b	84.50
166	Protein S human	83.70
167	Coagulation factor IX	83.60
168	Dulaglutide	83.40
169	Choriogonadotropin alfa	83.40
170	Chorionic Gonadotropin (Human)	83.40
171	Etanercept	82.10
172	Agalsidase Beta	81.60
173	Metreleptin	81.60
174	Somatotropin Recombinant	81.50
175	Menotropins	80.60
176	Conestat alfa	80.30
177	Human C1-esterase inhibitor	80.30
178	Romiplostim	79.30
179	Eftrenonacog Alfa	78.80
180	Lixisenatide	78.20
181	Mecasermin	75.80
182	Lutropin alfa	75.30
183	Alefacept	72.80
184	Denileukin diftitox	72.50
185	Parathyroid/Preotact	71.00
186	Tagraxofusp	67.20
187	Lepirudin	55.10
188	Elosulfase Alfa	49.70

Table B.2.
Rank order from pTM score AF2

Rank#	Name	pTM (AF)			
1	Asparaginase erwinia chrysanthemi	0.95	45	Urofollitropin	0.84
2	Alglucerase	0.95	46	Interferon alfa-2b	0.84
3	Velaglucerase alfa	0.95	47	Follitropin	0.84
4	Taliglucerase alfa	0.95	48	Aldesleukin	0.83
5	Galsulfase	0.94	49	Sargramostim	0.83
6	Ocriplasmin	0.94	50	Albiglutide	0.83
7	Laronidase	0.94	51	Alpha-1-proteinase inhibitor	0.83
8	Dornase alfa	0.94	52	Oprelvekin	0.83
9	Pancrelipase amylase	0.94	53	Aprotinin	0.83
10	Sacrosidase	0.94	54	Somatotropin Recombinant	0.82
11	Avalglucosidase alfa	0.94	55	Anakinra	0.82
12	Vestronidase Alfa	0.93	56	Retepase	0.81
13	Sebelipase alfa	0.93	57	Drotrecogin alfa	0.81
14	Cerliponase alfa	0.93	58	Ecallantide	0.80
15	Idursulfase	0.93	59	Metreleptin	0.80
16	Palivizumab	0.92	60	Conestat alfa	0.77
17	Olipudase alfa	0.92	61	Human C1-esterase inhibitor	0.77
18	Catridecacog	0.92	62	Coagulation Factor VIIa	0.77
19	Infliximab	0.91	63	Choriogonadotropin alfa	0.77
20	L-asparaginase	0.91	64	Chorionic Gonadotropin (Human)	0.77
21	Rasburicase	0.90	65	Becaplermin	0.74
22	Ranibizumab	0.90	66	Dulaglutide	0.73
23	Bevacizumab	0.90	67	Coagulation factor IX	0.73
24	Idarucizumab	0.90	68	Romiplostim	0.72
25	Ustekinumab	0.90	69	Insulin degludec	0.71
26	Interferon beta-1a	0.90	70	Insulin detemir	0.71
27	Palifermin	0.90	71	Insulin Regular	0.71
28	Urokinase	0.89	72	Insulin glargine	0.71
29	Digoxin Immune Fab (Ovine)	0.89	73	Insulin aspart	0.71
30	Antithrombin Alfa	0.88	74	Insulin lispro	0.70
31	Hyaluronidase (Ovine)	0.88	75	Insulin glulisine	0.70
32	Glucarpidase	0.88	76	Interferon gamma-1b	0.70
33	Ofatumumab	0.88	77	Asfotase alfa	0.69
34	Thyrotropin Alfa	0.88	78	Protein S human	0.67
35	Filgrastim	0.87	79	Galcanezumab	0.66
36	Chymopapain	0.87	80	Eflapegrastim	0.66
37	Brolucizumab	0.86	81	Agalsidase Beta	0.66
38	Efgartigimod alfa	0.85	82	Satralizumab	0.66
39	Belimumab	0.85	83	Mogamulizumab	0.65
40	Interferon Alfacon 1	0.85	84	Inebilizumab	0.65
41	Interferon alfa-2a	0.84	85	Pembrolizumab	0.65
42	Interferon beta-1b	0.84	86	Adalimumab	0.65
43	Darbepoetin alfa	0.84	87	Alirocumab	0.64
44	Erythropoietin	0.84	88	Emicizumab	0.64
			89	Nivolumab	0.64

Rank#	Name	pTM (AF)
90	Ansuvimab	0.64
91	Golimumab	0.64
92	Panitumumab	0.64
93	Gemtuzumab ozogamicin	0.64
94	Tocilizumab	0.64
95	Evinacumab	0.64
96	Dupilumab	0.64
97	Alemtuzumab	0.64
98	Ixekizumab	0.64
99	Odesivimab	0.64
100	Rituximab	0.64
101	Aducanumab	0.64
102	Eptinezumab	0.64
103	Romosozumab	0.64
104	Elotuzumab	0.64
105	Bezlotoxumab	0.64
106	Mirvetuximab Soravtansine	0.64
107	Obiltoxaximab	0.64
108	Muromonab	0.64
109	Burosumab	0.63
110	Brodalumab	0.63
111	Natalizumab	0.63
112	Vedolizumab	0.63
113	Crizanlizumab	0.63
114	Mepolizumab	0.63
115	Polatuzumab vedotin	0.63
116	Anifrolumab	0.63
117	Loncastuximab tersirine	0.63
118	Teprotumumab	0.63
119	Denosumab	0.63
120	Ibalizumab	0.63
121	Pertuzumab	0.63
122	Isatuximab	0.63
123	Durvalumab	0.63
124	Canakinumab	0.63
125	Eculizumab	0.62
126	Belantamab mafodotin	0.62
127	Obinutuzumab	0.62
128	Atoltivimab	0.62
129	Fremanezumab	0.62
130	Ravulizumab	0.62
131	Daratumumab	0.62
132	Tafasitamab	0.62
133	Reslizumab	0.62
134	Benralizumab	0.62

135	Ipilimumab	0.62
136	Tildrakizumab	0.62
137	Olaratumab	0.62
138	Necitumumab	0.62
139	Atezolizumab	0.62
140	Basiliximab	0.62
141	Dostarlimab	0.62
142	Secukinumab	0.62
143	Cetuximab	0.62
144	Ocrelizumab	0.62
145	Erenumab	0.62
146	Naxitamab	0.62
147	Risankizumab	0.61
148	Siltuximab	0.61
149	Trastuzumab	0.61
150	Cemiplimab	0.61
151	Dinutuximab	0.61
152	Sarilumab	0.61
153	Ibspatercept-aamt	0.61
154	Sacituzumab govitecan	0.61
155	Maftivimab	0.61
156	Tralokinumab	0.61
157	Margetuximab	0.61
158	Lanadelumab	0.61
159	Tezepelumab	0.61
160	Abatacept	0.60
161	Guselkumab	0.60
162	Mecasernin	0.60
163	Ramucirumab	0.60
164	Evolocumab	0.60
165	Tenecteplase	0.59
166	Emapalumab	0.59
167	Avelumab	0.57
168	Alteplase	0.57
169	Belatacept	0.57
170	Blinatumomab	0.57
171	Caplacizumab	0.54
172	Lutropin alfa	0.52
173	Omalizumab	0.52
174	Daclizumab	0.50
175	Menotropins	80.60
176	Conestat alfa	80.30
177	Human C1-esterase inhibitor	80.30
178	Romiplostim	79.30
179	Eftrenonacog Alfa	78.80

Rank#	Name	pTM (AF)
175	Ibritumomab tiuxetan	0.50
176	Tositumomab	0.49
177	Aflibercept	0.49
178	Eftrenonacog Alfa	0.49
179	Menotropins	0.48
180	Etanercept	0.47
181	Elosulfase Alfa	0.46
182	Tagraxofusp	0.45
183	Denileukin diftitox	0.44
184	Lixisenatide	0.43
185	Rilonacept	0.42
186	Parathyroid/Preotact	0.37
187	Alefacept	0.36
188	Lepirudin	0.31

Physiochemical Attributes computation Source Code

```
1 pip install biopython
2
3 import pandas as pd
4 from Bio.SeqUtils import ProtParam
5
6 # Define the function to calculate GRAVY
7 def calc_gravy(seq):
8     # Calculate the GRAVY (Grand average of hydrophaticity) value for a
9     # given protein sequence.
10    kd = 'ACDEFGHIKLMNPQRSTVWY'
11    seq = ''.join(aa for aa in seq if aa in kd)
12    seq = seq.replace('\n', '') # Remove any newline characters
13    pp = ProtParam.ProteinAnalysis(seq)
14    return pp.gravy()
15
16 # Define the function to calculate isoelectric point
17 def isoelectric_point(seq):
18    pp = ProtParam.ProteinAnalysis(str(seq))
19    isoelectric_point = pp.isoelectric_point()
20    return isoelectric_point
21
22 # Define the function to calculate extinction coefficient
23 def extinction_coefficients(seq):
24    pp = ProtParam.ProteinAnalysis(str(seq))
25    extinction_coefficients = pp.molar_extinction_coefficient()
26    extinction_coefficients_no_cys = extinction_coefficients[0] # with
27    # reduced cysteines
28    extinction_coefficients_cys = extinction_coefficients[1] # with
29    # disulfid bridges
30    return extinction_coefficients_no_cys, extinction_coefficients_cys
31
32 # Define the function to calculate instability index (< 40 protein
33 # stable)
34 def instability_index(seq):
35    seq = seq.replace(' ', '')
36    seq = seq.replace('\n', '')
37    if 'X' in seq:
38        return 0
39    else:
40        pp = ProtParam.ProteinAnalysis(str(seq))
41        instability_index = pp.instability_index()
42        return instability_index
43
```



```

40 |
41 | # Read the input Excel file
42 | df = pd.read_excel('protparam_input_data.xlsx')
43 |
44 | # Apply the functions to the sequence column and store the results in
    |   new columns
45 | df['GRAVY'] = df['Sequence'].apply(calc_gravy)
46 | df['Isoelectric point'] = df['Sequence'].apply(isoelectric_point)
47 | df['Instability Index'] = df['Sequence'].apply(instability_index)
48 |
49 | % Create the columns for Extinction coefficients (reduced cys) and
    |   Extinction coefficients (cys)
50 | df['Extinction coefficients (reduced cys)'] = None
51 | df['Extinction coefficients (cys)'] = None
52 |
53 | % Loop through the rows of the dataframe and compute the extinction
    |   coefficients for each sequence
54 | for index, row in df.iterrows():
55 |     cys_extinction_coefficients, no_cys_extinction_coefficients =
        |     extinction_coefficients(row['Sequence'])
56 |     df.at[index, 'Extinction coefficients (reduced cys)'] =
        |     no_cys_extinction_coefficients
57 |     df.at[index, 'Extinction coefficients (cys)'] =
        |     cys_extinction_coefficients
58 |
59 | % Save the output to a new Excel sheet
60 | df[['Name', 'Sequence', 'GRAVY', 'Isoelectric point', 'Extinction
    |   coefficients (reduced cys)',
61 |     'Extinction coefficients (cys)', 'Instability Index']].to_excel('
    |   peptides - protparam_output_data2.xlsx', index=False)

```

Listing C.1 Python code