DEVELOPMENT OF STRUCTURAL INFORMATICS METHOD FOR BINDING PEPTIDES

by

John R. Patterson

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2023

Approved by:

____

Dr. Donald Jacobs

____

Dr. Anthony Fodor

____

Dr. Jun-tao Guo

____

Dr. Xiuxia Du

____

Dr. Michael Matthews

ABSTRACT

JOHN R. PATTERSON. Development of Structural Informatics Method for Binding Peptides. (Under the direction of DR. DONALD JACOBS)

In the continuous pursuit of advanced therapeutics, the field of bioinformatics has innovated tools that allow unprecedented control over the proteome, profoundly shaping our understanding and manipulation of biological domains. Computational approaches to protein design grapple with the intricacies of protein behavior, encompassing everything from interaction dynamics to stability challenges. Methods in structural bioinformatics for peptide design typically hinge on the datasets of structures that have statistics applied to ascertain the effectiveness of protein design and modulation. When dealing with proteins that are poorly resolved, disordered, or niche, this task usually falls to experts in structural biology and often requires significant laboratory resources.

This thesis discusses an automated pipeline, devised to integrate remote sequence homology, structural modeling, and binding simulations of peptides to disordered proteins. Significant design and testing underpin this pipeline, aiming to generate binding peptides to any sequence, sidestepping the absolute requirement for an expert or a tedious process to produce leads. The utility of this pipeline is assessed across a diverse set of protein systems to refine its methodology. With the recent rise of machine learning-driven predictive or generative models, we explore their potential when integrated with our pipeline in attempt to address challenges in the computation of peptide binder design.

ACKNOWLEDGEMENTS

I distinctly remember when I decided to apply for this degree. I had called my mom, discussed the program and how I thought it might get me into jobs that I actually wanted. Figured there wasn't anything to lose by applying, but I wasn't quite sure I had another handful of years I could push through school. The doubt distilling down to my uncertainty on being able to achieve. It was at this point that my mom hit me with a classic. "Can't never could do nothin'". The phrase is amalgamated with my fathers voice, after hearing it on every occasion that one could conceive. Most notably perhaps during an outing that involved attempted mass turkey homicide. My father had great sayings, could literally talk a day away with any other living soul. He passed away in 2021. I miss him greatly.

So, inevitably, I applied to the department, hoping my computer science background and chemistry foreground would make my case for potential. And I heard nothing. Not a peep. Figuring life was rolling along, I even went as far as accepting a position teaching at a college. Until one random night in October Dr. Gibas sent me an email asking if I was planning to respond to the inquiry email for admittance. Suffocatingly exhilarating is the only way I can describe the process of feeling like maybe you compromised with your path forward to suddenly feeling infinite potential. It was a wild week of interviews, reading, researching, and driving around Charlotte. I knew of Dr. Jacobs through the grape vine, being involved with computational chemistry he had actually even been recommended to me for a class. If I recall, I believe it was Dr. Gibas that actually recommended I speak with him. Once again, if I recall correctly, our meeting went on for so long that his afternoon scheduled group meeting was approaching; so he invited me. I have literally been working with his Biomolecular Physics Group (BMPG) since then. I knew his outlook on science, connections, and rigor of experiment was exactly what I craved in research. Over the years I have been constantly amazed at his ability to motivate, converse, and

generally be a great mentor.

Naturally, as I sit in retrospect, this path was the only one I would have chosen. I met my wife, who is lovely beyond calculation, and I cannot imagine any aspect of my current life without her. The research, skills, and people I have gotten to know and work with were all amazing. Inside the BMPG my good friends Chris, Lonnie, Jenny, and Tyler all kept the environment alive and helped out in both research and life. I will miss working with everyone in the BMPG. My time working with the New Zealand group was also great. I never expected to work literally internationally. Everyone at PFR, Dr. Rikkerink, Dr. Wood, Dr. David, and Dr. Sun were immensely helpful and educational to me. Inside the Bioinformatics department and UNCC there were many people I found invaluable. In our IT department Jon Halter has literally given me an entire education on the secrets of linux, for which I am grateful. I found utility and enjoyment with educators and people in UNCC like Dr. Guo, Dr. Fodor, Dr. Du, Lauren Slane, Dr. Gibas, Dr. Mays, Dr. Nesmalova, and Dr. Jacobs too. I'd also like to thank Dr. Fodor for tangentially introducing me to my wife. There are actually many educators, particularly, who have contributed to my growth in appreciable ways. I doubt I could remember all of them, but some really stick with you. Mr. Adair, Mrs. Brenda Royal, Mr. Bass, Dr. Lee, Prof. Laura Bond. My friends, pets, and family all supported me in immeasurable ways. Even when I droned on about the vast unbounded intrestingness of chemistry, computers, or wiggly booger things.

I will not miss being in school, but I will miss the feeling of achievement and comradery that I was gifted during my time in PhD. Regardless, there will be more research, more to learn, and more to do. I can only hope it will be as fruitful as the last few years have been.

-JP

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

MD	Molecular Dynamics
PDB	Protein Data Bank
ML	Machine Learning
SAR	Structural Activity Relationship
QSAR	Quantitative Stability Activity Relationship
IDP	Intrinsically Disordered Protein
MoRF	Molecular Recognition Feature
MSA	Multiple Sequence Alignment
BLAST	Basic Local Alignment Search Tool
NLP	Natural Language Processing
NN	Neural Networks
LLM	Large Language Models
DR	Dimensionality Reduction
PSP	Protein Structure Prediction
PCA	Principle Component Analysis
SBDD	Structure-Base Drug Discovery
ESM	Evolutionary Scale Model
CNN	Convolutional Neural Network
p53	Tumor Protein p53
MDM2	Mouse Double Minute 2
SIRT	Sirtuin

CHAPTER 1: Structural Bioinformatics Applied to Proteins and Peptides

1.1 Background

Proteins are fundamental to the function of living organisms. They play an essential role in various biological processes, such as catalysis, signaling, transport, and provide structure. Even before the tetravalence of carbon, developed by Kekulé, in the mid 1700s albumin, fibrin, and gelatin were recognized protein substances due to their characteristic reactivity towards heat and acid [1]. Wide adoption and understanding of proteins being essential for life had propagated throughout the scientific community being foundational to modern dietetics [2]. Formal eponymous discovery of proteins in the 19th century when the chemists Mulder and Berzelius first identified that living organisms were composed of a variety of substances, including water, lipids, and nitrogen-containing compounds all sharing the empirical formula $C_{400}H_{620}N_{100}O_{120}P_1S_1$ [3]. However, the exact chemical nature of these compounds remained unclear for several decades until the early 20th century when scientists began to unravel the chemical nature of polypeptides.

As biochemist Archibald Garrod unraveled the mysteries of alkaptonuria in 1902, he introduced the pioneering concept of "inborn errors of metabolism", implicating the inability to metabolize aromatic amino acids like tyrosine and phenylalanine. This failure led to the accumulation of byproducts such as homogentisic acid. While Garrod's work elucidated the metabolic roles of these amino acids, it also alluded to the potential enzyme malfunctions as their root cause [4]. In a stroke of historical synchronicity, the same year witnessed the unveiling of the peptide bond by both Franz Hofmeister and Emil Fischer [1,5]. Fischer's ensuing endeavors encompassed the conceptualization of the lock-and-key model for enzyme functionality and a collaborative synthesis of the first optically active peptides with Otto Warburg [5,6].

While these revelations illuminated the rudiments of protein composition and function, the precise architecture of proteins remained unknown. The limelight shifted to Frederick Sanger when he decoded the sequence of insulin, heralding sophisticated sequencing techniques for the future. Further pushing the frontiers, J.B. Sumner achieved a monumental feat by crystallizing the enzyme jackbean urease [7,8]. Concurrently, Pauling's theoretical musings on the secondary structures of proteins, such as the α-helix and the β-sheet, set the stage for the empirical determination of protein structures by 1958 [9]. In a major contribution, Max Perutz and John Kendrew unraveled the structure of myoglobin, highlighting unforeseen complexities in protein structures [10].

Exploiting the power of X-ray crystallography and ancillary techniques, the scientific community resolved the three-dimensional structures of proteins. It was discerned that tertiary structures derived their stability from hydrophobic and hydrogen bond interactions. With their evolved surfaces and functional pockets proteins are adept at executing diverse roles in native environments [11,12]. Christian Anfinsen's research established protein structure was encoded within its amino acid sequence [13].

Better understanding about protein conformations and dynamics took root [14,15] as experimental techniques became more sophisticated. The need to disseminate this vast knowledge beyond traditional academic journals became paramount. Thus emerged repositories, consolidating structural data for open access and analysis [16]. As this field evolved, discernible patterns in protein folds, their functionalities, and structural encapsulations began to emerge. The expansive structural databases coupled with versatile experimental tools unveiled the immense structural diversity of proteins [17–19].

Protein structural diversity can be exploited to guide research in engineering novel functions via rational design. Pioneering studies demonstrated that, in some cases, only the functional sites of proteins were essential for their activity. This led to the simplification of full-length proteins down to their bare essentials for functionality [20, 21]. However, this approach had its limitations, as evidenced by the enduring challenges in enzyme design today [22]. Advancements in the field eventually enabled de novo design, where the designed sequences are not identifiable in any homological sequence searches, not necessarily implying wholly unique structures. An early triumph in this realm was the creation, expression, and characterization of 3-helix bundles, marking the onset of an applied field dedicated to functional peptidic polymers [23, 24].

Structural bioinformatics has emerged as an indispensable tool for the creation of innovative therapeutics and materials. Protein 3D structural information is captured through experimental methods such as: X-ray crystallography (XRC), nuclear magnetic resonance (NMR), cryo-electron microscopy (cryo-EM), each with its advantages and disadvantages [25, 26]. Other experimental methods to characterize protein structures include SAXS, CD, FRET, and Raman spectroscopy [27]. These methods facilitate high-resolution mapping of large macromolecular structures, ushering in fresh possibilities for structural exploration that is leveraged in medicinal applications.

1.2 Proteins

While proteins exhibit a rich diversity of structures, they are governed by the underlying physics and chemistry that determine enthalpic and entropic contributions of their folded conformations [13]. In living organisms, a vast array of molecular interactions orchestrate chemical reactions, guided by the principles of thermodynamics and kinetics. Proteins have evolved to adopt conformational ensembles that align with both thermodynamic and kinetic stabilities that ensure functionality, particularly in their native chemical environments [28]. The specificity inherent to protein function is partially encoded in its primary structure, as the three-step progression of sequence, structure, and function in proteins.

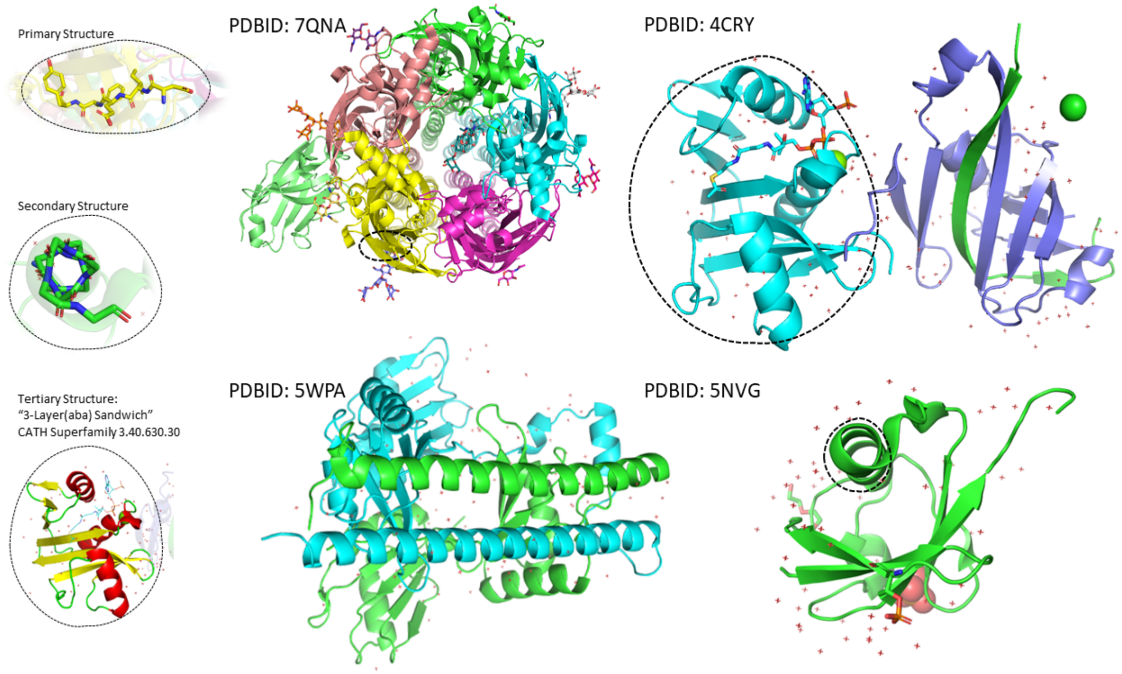


Figure 1.1: The fundamental units of protein structures are depicted on the left. These structures progress from primary to secondary, and then to tertiary levels, creating the protein polymer's folds. Folded proteins can then bind to themselves (5WPA), another protein (4CRY), and can bind together with many other proteins to form very large nanoscale structures (7QNA). Proteins are modifiable even after being folded or bound to others via post translation modifications, such as the phosphate on a serine of 5NVG.

The sequence-structure relationship is a hierarchical relationship that is categorized into primary, secondary, and tertiary structural levels. The primary structure is the linear amino acid sequence in a polypeptide chain. Secondary structure pertains to the local spatial conformation along the backbone, introducing topological complexity. The tertiary structure describes the three-dimensional folded arrangement of the secondary structures, which causes buried and exposed residues [15]. Quaternary level structures are assemblies of folded domains from separate polymer chains.

The dynamics of protein compositions play a direct role in influencing their functions [29–31]. As illustrated in Figure 1.1, the fundamental components of protein structures span from secondary to tertiary and quaternary levels. Other factors that modify the dynamics of proteins are post-translational chemical modification and interactions with solutes including higher order protein-protein interactions. The conformational states of proteins can be altered by various factors, including molecular cofactor interactions and environmental conditions such as temperature and pH. However, the true essence of a functional protein lies in its stabilizing elements that determine its native conformational state. Factors like hydrophobic interactions, hydrogen bonding, disulfide bonds, and salt bridges are crucial for maintaining protein stability. Typically, proteins demonstrate functionality in their native folded states, even if such functions aren't immediately evident [32, 33].

Analyzing the folded states of proteins has spurred the growth of databases like the RCSB Protein Data Bank (PDB). Techniques such as X-ray diffraction (XRD), electron microscopy (EM), and nuclear magnetic resonance (NMR) spectroscopy facilitated this [34]. Modern advancements include neutron scattering data. While some methods reveal fixed protein conformations, like XRD or neutron diffraction, others like NMR or circular dichroism methods showcase an ensemble of structures in a solution state.

Protein structures are commonly categorized into fold families and sheds light on evolutionary protein relationships, even when hard to discern from sequences. Recognized databases in this domain include SCOPe, CATH, and ECOD, among others [35, 36]. Recent studies have employed unsupervised machine learning to expedite the identification of obscure structural relationships in proteomes [37]. Notably, structural conservation, more than sequence, is observed in proteins performing identical functions across different species [38, 39].

Protein structures are complicated and derive specific activity from evolution. This structure-activity relationship (SAR) is a century-old principle in structural biology, pivotal in developing major pharmaceuticals [40]. However, SAR alone sometimes only offers a superficial understanding, failing to encapsulate the dynamics defining biological systems [41]. By integrating known dynamics and mathematical correlations, quantitative structure–activity relationships (QSAR) can formulate hypotheses for biological functionalities [42, 43].

1.2.1 Protein Stability and Dynamics

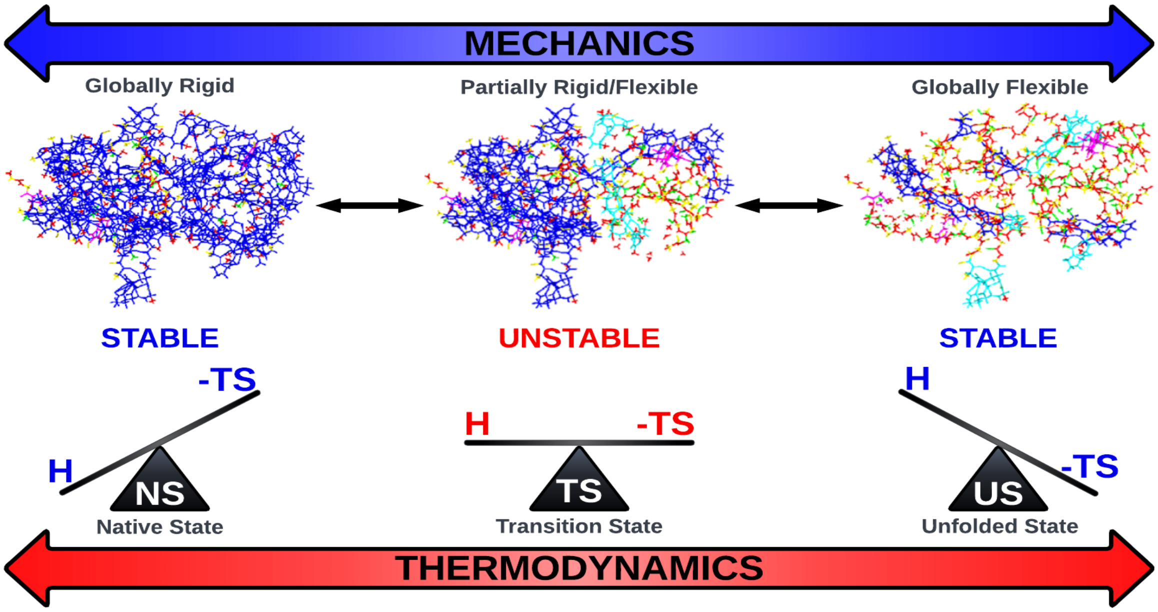


Figure 1.2: Protein stability depicted as states of mechanical properties. Native states of proteins undergo a free energy transition to unfold states. Demonstrated in network flexibility, shown here in yellow and red, as determined by covalent and hydrogen bond network.

The challenge of correctly folding tertiary protein structures, especially larger ones, is well-recognized [44]. Most long protein chains, exceeding 100 residues, do not quickly attain their native folded state. Outside of the cellular environment where they are synthesized, proteins can be notoriously difficult to fold. Once removed far from their native environments, these molecules easily denature. Stability is a contextual category for proteins where function, structure, and robustness can each have different definitions. Proteins are often denoted as stable if their tertiary structure can be isolated, structurally characterized, and can be folded readily outside native biological expression. Proteins can be stable over a range of different conditions, with robust proteins spanning a wide range, especially once folded.

The stability of proteins is directly related to the conformation, this can be viewed as a constraint network defined by covalent bonds or other strong interactions, like hydrogen or ionic bonds. When framed as a mechanical network problem flexible and rigid clusters can be identified in protein structures, as depicted in Figure 1.2. These networks consist of stable sub-components that interact to stabilize the protein in its environment. Factors such as temperature, pH, salt concentration, and the presence of certain small molecules or other proteins can interact with these constraint networks; perturbing the conformation in response. [45–47].

Exemplified in Figure 1.3, the resolved conformation of a protein is usually a static point in a large space of conformations. However, on time scales relevant to their function or the instruments used for measurements, proteins in solution explore an ensemble of conformations. Early protein computational methods primarily emphasized the role of side chain rotamers, to aid in predicting protein stability [48]. This provided a limited accuracy as protein stability involves not just the configurations of the side chains but also the protein backbone structure.

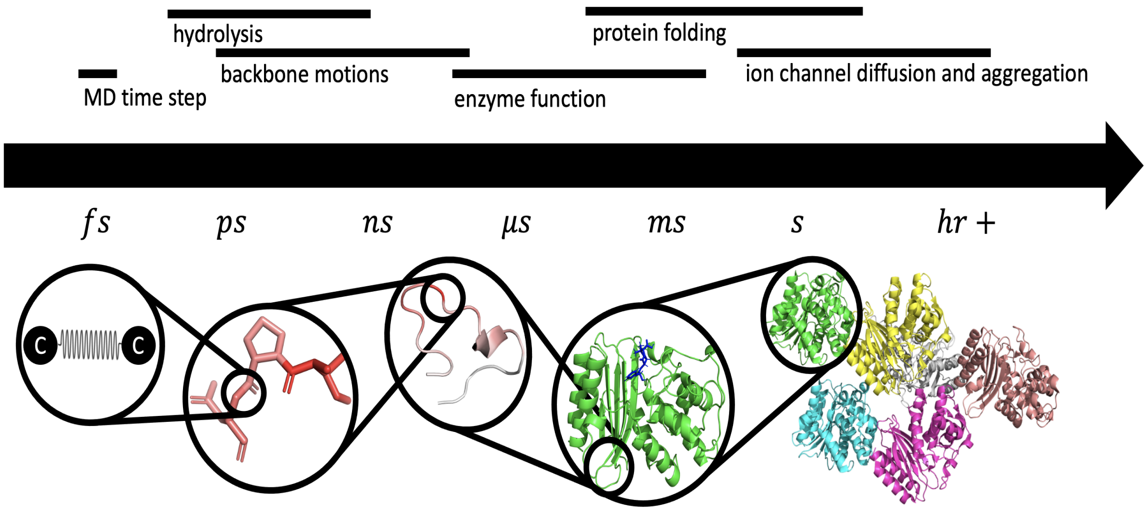


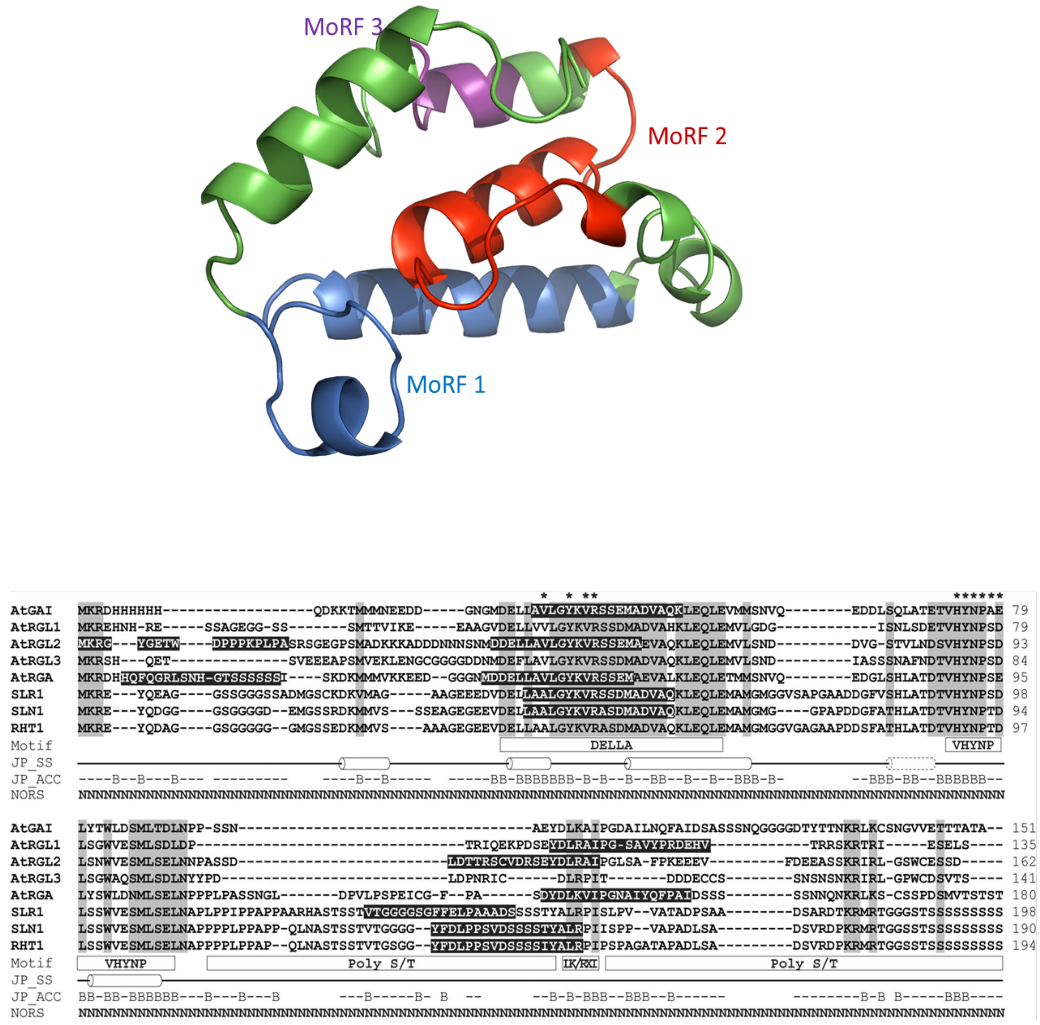
Figure 1.3: Protein function and dynamics occurs on time scales that span many orders of magnitude depending on the process of interest. Left to right: bond vibrations, dihedral torisons, backbone dynamics, protein dynamics, and a complex of proteins interacting dynamically.

1.2.2 Motifs and Molecular Recognition Features

Protein motifs are defined by conserved regions present in both three-dimensional structures and amino acid sequences. Segments of conserved residues provide evolutionary insight into the functionality of a protein over time. Often these motifs are named for their function, like a Zinc Finger motif, or the conserved one-letter protein sequence might be used for the name. In many cases motifs have a limited set of residues that will allow the function or structure. For instance, the alpha helix, a quintessential secondary structure, derives its stability primarily adjacent residue backbone hydrogen bonds in addition to capping residues. Multiple alpha helices can assemble to produce stable structural arrangements in solution [23, 49].

Intrinsically disordered proteins (IDPs) are proteins, or specific regions within, that remain in a transient state lacking a defined structure that is characterizable. Nevertheless, their functional importance is underscored by their critical roles in signaling, transcription, and translation mechanisms [50–52]. A subset of IDPs contain molecular recognition features (MoRFs). Characterized by their secondary structure inside a disordered domain, these segments often initialize a transition from a disordered state to an ordered conformation upon binding to specific interacting entities [53]. Notably, a significant percentage of proteins, especially within eukaryotes and signaling pathways, exhibit extended disordered regions. Generically, MoRFs facilitate biological processes that often instigate additional signaling cascades [54–56].

A demonstrative example is provided by the conserved MoRFs, as illustrated in Figure 1.4. Comprehensive sequence alignments spanning diverse evolutionary trajectories delineate characteristic residues typically found within protein segments classified as disordered. These residues, emphasized by their grey shading, play a pivotal role in interaction mechanisms or chemical reactions, thereby necessitating their conservation to ensure functional integrity. A case in point is the N-terminal domains of DELLAs, intrinsically unstructured under physiological parameters [57]. Within these disordered domains a MoRF initializes the transition between disorder to order upon binding with partner proteins and sometime co-factors, such as GID1 in the presence of a Gibberellin ligand.



```
                                                                    *   * **                                                     ******
AtGAI  MKRDHHHHHH------------------QDKKTMMMNEEDD-----GNGMDELLAVLGYKVRSSEMADVAQKLEQLEVMMSNVQ--------EDDLSQLATETVHYNPAE 79
AtRGL1 MKREHNH-RE-----SSAGEGG-SS-------SMTTVIKE-----EAAGVDELLVVLGYKVRSSDMADVAHKLEQLEMVLGDG------------ISNLSDETVHYNPSD 79
AtRGL2 MKRG---YGETW----DPPPKPLPASRSGEGPSMADKKKADDDNNNSNMDDELLAVLGYKVRSSEMAEVAQKLEQLEMVLSND---------DVG-STVLNDSVHYNPSD 93
AtRGL3 MKRSH---QET--------------SVEEEAPSMVEKLENGCGGGDDNMDEFLAVLGYKVRSSDMADVAQKLEQLEMVLSND---------IASSSNAFNDTVHYNPSD 84
AtRGA  MKRDHHQFQGRLSNH-GTSSSSSI----SKDKMMMVKKEED--GGGNMDDELLAVLGYKVRSSEMAEVALKLEQLETMMSNVQ--------EDGLSHLATDTVHYNPSE 95
SLR1   MKRE---YQEAG----GSSGGGSSADMGSCKDKVMAG-----AAGEEEDVDELLAALGYKVRSSDMADVAQKLEQLEMAMGMGGVSAPGAADDGFVSHLATDTVHYNPSD 98
SLN1   MKRE---YQDGG----GSGGGGD--EMGSSRDKMMVS---SSEAGEGEEVDELLAALGYKVRASDMADVAQKLEQLEMAMGMG----GPAPDDGFATHLATDTVHYNPTD 94
RHT1   MKRE---YQDAG----GSGGGGG--GMGSSEDKMMVS----AAAGEGEEVDELLAALGYKVRASDMADVAQKLEQLEMAMGMGGVGAGAAPDDSFATHLATDTVHYNPTD 97
Motif                                                       DELLA                                                  VHYNP
JP_SS
JP_ACC ----B--B---B--- ---------    ----B--B-----  -----BB-BBBBBBB-B-B--B--BB--B--B--BBB-B-           --BBB-BB--BBBBBB--
NORS   NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

AtGAI  LYTWLDSMLTDLNPP-SSN-----------------------------------AEYDLKAIPGDAILNQFAIDSASSSNQGGGGDTYTTNKRLKCSNGVVETTTATA-- 151
AtRGL1 LSGWVESMLSDLDP-------------------------------TRIQEKPDSEYDLRAIPG-SAVYPRDEHV------------TRRSKRTRI-----ESELS---- 135
AtRGL2 LSNWVESMLSELNNPASSD-------------------------LDTTRSCVDRSEYDLRAIPGLSA-FPKEEEV---------FDEEASSKRIRL-GSWCESSD----- 162
AtRGL3 LSGWAQSMLSDLNYYPD---------------------------LDPNRIC------DLRPIT-------DDDECCS--------SNSNSNKRIRL-GPWCDSVTS---- 141
AtRGA  LYSWLDNMLSELNPPPLPASSNGL-------DPVLPSPEICG-F--PA------SDYDLKVIPGNAIYQFPAIDSSS--------SSNNQNKRLKS-CSSPDSMVTSTST 180
SLR1   LSSWVESMLSELNAPLPPIPPAPPAARHASTSSTVTGGGGSGFFELPAAADSSSSTYALRPISLPV--VATADPSAA--------DSARDTKRMRTGGGSTSSSSSSSSS 198
SLN1   LSSWVESMLSELNAPPPPLPPAPP-QLNASTSSTVTGGGG--YFDLPPSVDSSSSTYALRPIISPP--VAPADLSA---------DSVRDPKRMRTGGSSTSSSSSSSSS 190
RHT1   LSSWVESMLSELNAPPPPLPPAP--QLNASTSSTVTGSGG--YFDLPPSVDSSSSIYALRPIPSPAGATAPADLSA---------DSVRDPKRMRTGGSSTSSSSSSSSS 194
Motif     VHYNP                     Poly S/T                    IK/RXI                    Poly S/T
JP_SS
JP_ACC BB-BB-BBBBBB---B---B---B       ---B-----B- B  --      --B-B-BBB--BBB--B------         --------B-B B-----BBB----
NORS   NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

Figure 1.4: Top, cartoon of the MoRF site of GAI disordered N-terminal, from PDBID 2ZSH. Bottom, alignment of GAI homologs denoting the conserved MoRFs (DELLA, VHYNP) [57].

1.3 Short Peptidic Polymers

Peptides are short chains of amino acids that are distinct from proteins in terms of their size, structure, and function. The terminology derives from Emil Fischer's dipeptid, tripeptid, and polypeptid (1902–1903), and defines a molecular compound which has a polymeric amide bond with at least two monomers. While proteins are composed of long chains of amino acids, peptides typically consist of fewer than *ca.* 50 amino acids. Biologically, peptides are multifunctional, acting as hormones, neurotransmitters, and signaling molecules [58]. Additionally, they have therapeutic potential as antibiotics, antiviral agents, tools for cellular engineering, and critical components in immune responses [59–62]. Several peptide-based drugs are currently in use or in development. The peptide drug Exenatide, used to treat type 2 diabetes, is among several Glucagon-Like Peptide-1 (GLP-1) analogs that have gained popularity for off-label weight loss. The discovery of these drugs has been made possible by advances in peptide synthesis and production techniques, which have made it possible to create a wide range of peptide structures.

1.3.1 Peptides in Nature

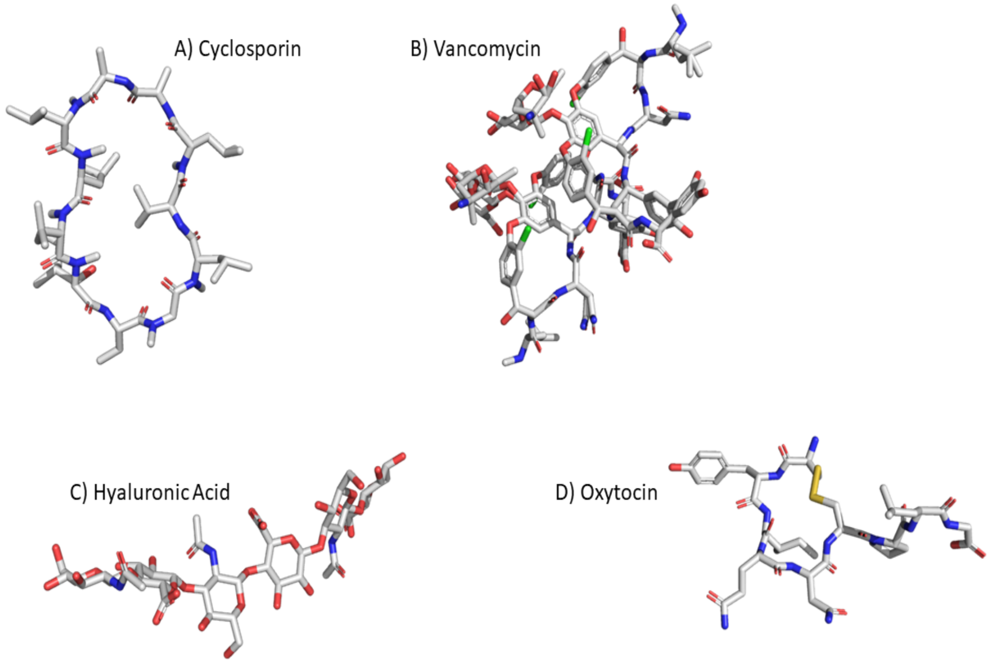


Figure 1.5: A) Cyclosporin, a calcineurin inhibitor used as an immunosuppressant medication. B) Vancomycin, a glycopeptide antibiotic medication. C) Hyaluronic acid, a glycosaminoglycan gel-like, polymer found in skin tissues. D) Oxytocin, the amorous hormone, is a peptide hormone and neuropeptide.

Peptides play pivotal roles in various biological functions found in nature. Their diverse range includes unique post-translational modifications resulting in di-peptide conjugates, such as insulin, and cyclized structures like oxytocin and cyclosporin. Some natural peptides undergo extensive modifications, deviating substantially from their original structure, as illustrated in Figure 1.5. For instance, triceptides can be so transformed that their structures structures look more complex than compounds like vancomycin or glycosaminoglycans [63].

While peptides exhibit significant structural diversity, they naturally possess traits that render them valuable in drug development. Their generally small size and inherent flexibility position them well for engaging with specific molecular targets. With rapid synthesis and degradation rates, peptides routinely facilitate regulation of biological activity. The therapeutic potential of peptides is increasingly acknowledged, with several peptide-based drugs, such as the Glucagon-Like Peptide-1 (GLP-1) for type 2 diabetes and the antimicrobial peptide Daptomycin, now in clinical practice [64, 65]. Beyond their direct therapeutic roles, peptides are being explored as vehicles for targeted drug delivery, aiming to enhance the efficiency and safety profiles of established medications [66, 67].

1.3.2 Peptides in Biotech and Medicine

As nature demonstrates, peptides are versatile tools to control proteins with specific characteristics to be useful signalling agents. Given that most pharmaceuticals modulate signaling mechanisms within the body, it is not surprising that peptide-based drugs have a long history in medical applications. One of the earliest and most celebrated instances is the discovery and therapeutic use of insulin, first isolated in the early 1900s [68]. This groundbreaking achievement was followed by its lifesaving administration to children with diabetes, garnering a Nobel Prize in 1923. Peptide-based drugs have had a multitude of applications such as the human growth hormone, the biomimetic peptide Enfuvirtide for human immunodeficiency virus (HIV), or the recent surge in GLP-1 based peptide drugs.

As pharmaceuticals, peptides prove indispensable for several applications [69–71]. They facilitate enzyme activation [72, 73], and are paramount in modulating protein-protein interactions [62, 74]. Peptide inhibitors halt biological processes ranging from cell signaling to viral propagation. Advanced methods now allow for large-scale peptide synthesis and efficient integration of peptides into living organisms [75]. Designing peptides for specific applications often entails drawing parallels with biological references to achieve optimal interactions with protein or other molecular targets. However, crafting a peptide with desired interaction attributes remains an endeavor, especially when aiming to interact with difficult or uncharacterized targets.

1.4 Computational Methods for Protein Design

Historically, plasmid insertions gave biochemists the ability to deduce functional sites and facilitate molecular engineering [76, 77]. Early computational methodologies developed for protein design primarily hinged on the concept of rational design. Approaches ranged from augmenting hydrogen bonds to solidify particular protein folds, fortifying stability via covalent disulfides, to even grafting entirely new functional domains onto recognized structured domains [24, 78]. Computationally, proteins were designed using intervals of known possible rotations to define the backbone conformations and sidechain repacking algorithms to design the sequence that fit the desired structure [79, 80]. Achievements of these methods included folding a small protein, repacking proteins, small binding peptides, eventually leading to the first ever de novo, from the beginning, designed protein [21, 81, 82].

As applications scaled to large proteins, implementations were made to crowd source the computational demand [83]. With the development of the internet and digital repositories for empirically collected structural and sequence data, knowledge-based design emerged as the more feasible approach. Pioneering contributions from renowned teams like those of Baker, Khulmann, Karlplus, and Zhang [84, 85, 85–87].

These strategies, complemented by improving computational capabilities, were employed to tailor-make proteins with bespoke attributes, novel folds, or enzymatic properties [88–90]. Concurrently, biological methodologies, like directed evolution, for protein design were deployed in many sectors of research and industry with reasonable efficacy; still heavily research due to byproducts, contaminants, and a variety of other living cell related issues [91, 92]. Processes like these are exemplified by the works of Anfenson and Arnold, harnessing the rapid evolutionary capabilities of organisms like bacteria or yeast [13, 93]. Directed evolution, another facet, remodels existing proteins genetically to enhance or innovate their functionalities utilizing the rapid evolution of bacteria or yeast to generate new protein sequences with desired properties. This method has been used to generate novel enzymes for industrial applications, among other achievements [94].

Computational protein design has witnessed steady advancements in recent decades, evolving from rudimentary rational designs using coarse grained modelling to sophisticated computational methods. Pioneering contributions include the development of the Rosetta software suite by David Baker's group, which facilitated the design of proteins with unprecedented functions. Notably, Kortemme and Baker's exploration into the computational design of protein-protein interfaces stands as a testament to these advances [95, 96].

Recently classical computational approaches have been increasingly supplanted or augmented by machine learning models. Given the large amount of empirical data now available, machine learning techniques offer an regressive and predictive modeling. Generically it can be stated that machine learning is expected to impact on fields such as medicine, structural bioinformatics, and protein design [97,98]. By generative design and predictions trained from the decades of existing work and empirical data, these algorithms are making the design process more streamlined and accurate, with respect to characterization ability [99, 100].

1.4.1 Molecular Dynamics

The basics of MD simulations regardless of detail level or implementation are: given initial positions and velocities for each elementary object (atom or group of atoms) in the system, all resultant positions and velocities are calculated at some time interval later. For an all-atom MD simulation, the molecular interactions are modeled by potential energy functions given by Equation 1.1. In computational chemistry often a single molecule has the formal bonding and valence electrons minimized using a variety of techniques and foundational algorithms to approximate the multi component systems. This can be used to calculate a variety of properties of a molecule and even be used to calculate transition states of intramolecular interactions [101]. Everything above the dynamics of electrons are typically mimicked in the simulation [102]. Key aspects of stability and packing characteristics of a protein are determined by van der Waals interactions, electrostatic interactions, hydrogen bonds, and various torsion movements [103]. These independent considerations can be summed to simulate the nanoscale molecular environment. All of these calculations are implemented into software that can perform them using clever calculation schemes on a range systems sizes and force fields [104, 105].

Bond Stretching Potential Energy:

$$U_{\text{bond}}(r) = \frac{1}{2}k_{\text{bond}}(r - r_{\text{eq}})^2 \tag{1.1a}$$

Bond Angle Potential Energy:

$$U_{\text{angle}}(\theta) = \frac{1}{2}k_{\text{angle}}(\theta - \theta_{\text{eq}})^2 \tag{1.1b}$$

Dihedral Angle Potential Energy (for torsional angles):

$$U_{\text{dihedral}}(\phi) = \frac{1}{2}V_n[1 + \cos(n\phi - \delta)] \tag{1.1c}$$

Van der Waals (Lennard-Jones) Potential Energy:

$$U_{\text{vdW}}(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{1.1d}$$

Electrostatic (Coulombic) Potential Energy:

$$U_{\text{elec}}(r) = \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r} \tag{1.1e}$$

Hydrogen Bonding Potential Energy:

$$U_{\text{hbond}}(r, \theta) = -\frac{A}{r^{12}} + \frac{B}{r^6} + \frac{C}{r} + k_{\text{angle}}(\theta - \theta_{\text{eq}})^2 \tag{1.1f}$$

Modern MD has achieved many feats, including simulating entire cells with a combination of computational details [106, 107]. Meta dynamics enhances molecular dynamics by guiding simulations to explore underrepresented regions in the free-energy landscape, allowing for extremely long time scales or to investigate some manifold of dynamics like unfolding [108–110]. The burgeoning interest in these exascale simulations is likely being driven by their pharmaceutical implications [111]. With correct considerations MD can recreate a variety of information about drug-protein interactions including estimating kinetic properties [43, 111]. Investigations have demonstrated the potential to estimate drug properties or determine entire mechanisms through MD based experiments, facilitating the rational design of drugs and proteins [108, 112–115].

1.4.1.1 Molecular Dynamics in the Context of Protein Design

Molecular dynamics (MD) simulations play a pivotal role in protein design, offering valuable insights into protein stability, dynamics, and function [116,117]. The application of MD to proteins began in in the late 70s with the dynamics of bovine pancreatic trypsin being simulated using a CHARMM force field; when the most common processor was a 8-bit 2MHz MOS Technology 6502 [118]. These simulations, rooted in classical mechanics, facilitate the study of molecular behavior over time. A common approach is to consider an all-atom model where the equations of motions for each atom is solved within the system that is comprised of the molecule(s) of interest and solvent atoms. Such detailed explorations can elucidate conformational changes during processes such as small molecular, a.k.a. ligand, binding, enzyme catalysis, or other dynamic events.

Predicting the stability of specific protein sequences stands as a formidable challenge within protein design [119]. Nevertheless, MD simulations can estimate the free energy changes upon mutation or ligand binding states and discern changes in stability in response to such events [120, 121]. Such insights are invaluable in guiding the design of novel proteins boasting enhanced stability. Beyond the confines of stability and ligand binding, MD simulations extend their utility to the study of protein-protein interactions to pinpoint crucial binding interfaces and determine the residues that are important to a given function [108]. Generically, the state of simulation lacks chemical dynamics and heterogeneous system accuracy. This coupled to computational resources required to simulate biological systems and timescales limits MD applications outside of case studies.

1.4.2 Protein Sequence Alignment

Protein sequence alignments are a pivotal component of the field of bioinformatics, providing critical insights into the relationships between biological sequences. Multiple Sequence Alignments (MSAs) quantify the degree of variability, or conservation, within DNA, RNA, and protein sequences facilitating the elucidation of evolutionary connections, functional motifs, and essential biological information [122]. Most alignments utilize some flavor of Needlman and Wunsch global dynamic programming alignment algorithm, often paired with divide and conquer strategies [123]. In a correct MSA, aligned residues, letters, should be maximally similar according to some appropriate metric, in some cases an expectation value is used. For evolutionary reconstruction, the residues must be homologous, meaning they correspond to the same residue from the last common ancestor of the analyzed sequences. Thus, making the alignment to a homolog statistically significant compared to a random sequence. Various metrics exist to measure accuracy of alignment that implement direct calculations, utilize the sum of weights of mutations in the sequence tokens, or other heuristics to pass some objective function for alignment [124].

MSA can identify conserved domains, functional motifs, and structural elements; elucidating the nuanced relationship between protein structure and function. Several databases and tools augment the analysis of protein sequences. The UniProt database, encompassing curated protein sequences, along with the NCBI Protein Database and the Protein Data Bank (PDB) containing structural and sequence information, are pivotal resources. Among the toolkit, the Basic Local Alignment Search Tool (BLAST) offers versatile comparisons of protein and nucleotide sequences against diverse databases to identify homologous regions and infer functional or evolutionary relationships [125]. Similarly, programs such as MAFFT, Clustal Omega, HMMER, Muscle, and T-Coffee, extend diverse alignment methodologies catering to both nucleotide and protein sequences [122].

Within the realm of sequence similarity search, the pursuit of more diverse outcomes, rather than exclusively top-similar results, becomes particularly relevant in cases where the goal is to identify all functional domains within a query sequence or to understand evolution. The scope of such diversification extends to identifying proteins with disparate functions, yet share sufficient similarity with the query sequence. Instances of diversifying sequences directly from a single sequence are scarce [126]. But alternative to protein BLAST, which operates by direct sequence similarity, Position-Specific Iterative (PSI)-BLAST harnesses profiles constructed through the consideration of evolutionary relationships, facilitating the detection of distant protein relatives. Although the notion of diversification remains niche within the context of biological sequence searchs, parallels can be drawn from the realms of information retrieval and recommendation systems, where diversity and novelty are desired.

1.4.3 Docking

An important aspect of design and function is understanding if proteins will interact with other proteins [127]. For example, an interaction with certain immunological proteins can induce an immunological response, which can be troublesome if medicinal utility is desired. There has classically been a strong interest in docking small molecules to proteins, to characterize or predict binding sites in proteins, and estimate binding affinities computationally [128]. However, predicting protein–protein interactions is particularly difficult because of the innate flexibility and size of binding regions on proteins, creating a high dimensional problem. This topic represents a current challenge in computational biology. Considering the possibly multiplicity of protein interactions to form large quaternary structures, it is likely the field as a whole is only scratching the surface of possible arrangements and structures that protein–protein interactions can achieve as assemblies [129].

Docking is divided into flexible and rigid docking, largely. Similar in implementation to MD, flexible docking allows movements consistent with molecular properties, but require calculations to find reasonable interactions. Rigid docking an be fast, but is limited by the conformation problem. Neither of these methods can provide free energy values or kinetic properties of these interactions. Most docking software for proteins are focused on finding where a small molecule will fit onto or inside the protein. Some example programs are AutoDock, SwissDock, Schrodingers' GLIDE program are classic small molecule docking to proteins. Many scientists rely on structure information being present for a interface of interest, or a close homolog to work from. On the scale of possible proteins the structural information currently present, *ca.* 200,000, is absurdly limited. For specifically protein protein interactions, which can include peptides, there is a sparse amount of software. HADDOCK is a partially flexible docking system, and can reveal extremely high detail of interactions and their near bound state dynamics [130]. This method requires a reasonable starting state for the docking, in other words it is not ab inito and requires a good initial approximation to a final state. Full structure docking is only tractable at a coarse grain level of detail between proteins. ZDOCK is a rigid docking method that was been used widely for its fast Fourier transform calculation used to arrive at interaction scores, similar in spirit to PatchDock [131,132]. Typically these kinds of methods utilize additive metrics to arrive at a score that optimizes the true positive from some dataset derived from the PDB. Recently MEGADOCK was released as another method directly implemented based off ZDOCK, but utilizing faster GPU hardware to perform the calculations [133].

1.5 Peptide Design

Peptides are versatile molecules that can specifically bind to receptors, even targeting sites that were once considered undruggable by small molecules. Entire books have been written on both the design of peptides for medicinal use [134,135]. Peptides can be used to interrupt biological processes ranging from cellular signaling to viral infection vectors industrial chemistry, or consumer use in cleaning agents, cosmetics and food preservatives [70,73,136–140]. The utility of peptides to promote biological functions is well established1 [58,72,136,141]. There has been great effort to design peptides for a variety of applications, particularly in medicine [61].

Instead of small molecule virtual libraries, often deployed in small molecule drug development, bioinformatic approaches are typically used to derive peptide libraries based on known information from the receptor and known binding partners. By leveraging structural homology and the propensity of certain amino acid sequences to form secondary structures, design provides a manageable subset of sequence space from millions, typically 20^n where n is the length of the peptide) to thousands of sequences [74, 142–145]. For the sake of clarity, an octamer peptide is over 1 billion possible sequences, 2.56e+10. This also excludes non-canonical amino acids, which are theoretically endless in number as the classical chemistry tool box is available offering literally endless possibilities [146]. In practice they can impart a variety of interesting effects are are widely used in circulating pharmaceuticals [147].

There are multiple works related to automated small protein and peptide design that involve the docking problem as a primary driving for the process. DynaRock, Rosetta FlexPepDock, and PepCrawler represent popular software for peptide modelling/docking [148–150]. More recently there has been an uptick in research regarding tools where the focus is seemingly a dichotomy of determining the binding affinity or generating the peptide [151–154]. Baker's group recently devised a backbone ab initio method in Rosetta for peptides that designs them and evaluates affinity [155]. Unfortunately, this method produces a limited set of peptides and requires intensive ab inito modelling.

1.5.1 Rational or De Novo?

Given a rigid structure the classic Rosetta tool set can identify collisions and improper angles in structural interactions. This is approximated using a series of empirically determined distributions of various protein related variables, like side chain rotamer spacing. Unfortunately for peptides, the conformation issue proves a extreme issue, which Rosetta has classically been weak towards [156]. Major limitations include the presence of multiple binding modes, plasticity in the receptor, large conformational changes in the binding molecule, highly charged systems, and comparison across different series of compounds

Rational design for peptides has demonstrated significant success in drug discovery, resulting in a plethora of marketed drugs that have emerged from this process. Notable examples include GLP1-based drugs, where the systematic exploration of the SAR of the GLP1 peptide has led to the development of effective treatments for diabetes [40]. The optimization of the GLP1 peptide and its receptor interactions has paved the way for innovative therapies targeting glucose regulation and metabolic disorders. Heparin, a widely used anticoagulant dominating pharmaceutical market for many years, showcases the impact rationally designed peptide-based drugs could have [157]. By elucidating the SAR of its binding interface with coagulation factors, researchers have been able to tailor heparin derivatives with improved efficacy and safety profiles. Another prominent example is Enfuvirtide, marketed as Fuzeon, which targets the HIV-1 virus by disrupting its fusion with host cells. The study by Chong et al. (2018) detailed the development of a novel T-20 sequence-based lipopeptide, exhibiting broad-spectrum activity against various HIV strains and related viruses, including those resistant to existing treatments [145]. Their comprehensive characterization of the SAR of T-20 derivatives highlighted the potential of rational design in utilizing peptide structural insights for advanced antiviral therapeutics. Novel membrane fusion inhibitors against HIV and other enveloped viruses is highly important in terms of the peptide drug T-20, which remains the only one for clinical use, even if it is limited by large dosages and resistance.

De novo design of peptides has classically been difficult. Successful works include tunable set of heterodimeric peptides that form coiled coils and based on the peptide sequence, have an array of different functions, from cellular localization to transcriptional activation. These peptides were used to construct a CRISPR-Cas9 transcriptional activator, which increased the cellular response to certain light and chemical stimuli [158]. Designs of peptides capable of binding to RNA, which focused mainly on structure based peptidomimetics [144]. Peptide designs meant to mimic the sequence and structure of proteins known to interact with RNA. Using stable well-known secondary structures like β-hairpins and α-helices to provide a backbone structure that, when combined with energetically favorable sidechains lead to greater binding affinity. Chevalier et al. published a study where they designed and tested 22,660 de novo mini protein binders and 6,286 control sequences [159]. The mini proteins were 37–43 residues in length, contained multiple hydrophobic residues and were designed to bind influenza haemagglutinin and botulinum neurotoxin β. The study identified 2,618 high affinity binders that are stable at high temperatures. Various groups have been using de novo designed peptides to both study the function of this virus as well as develop peptides that can interfere with the attachment of the spike protein to angiotensin 2 [160].

CHAPTER 2: Review of Machine Learning in Computational Biology

2.1 Machine Learning Paradigms and Applications to Proteins

Machine learning (ML) seeks to train models to accomplish specific tasks or actions. Methods rooted in discriminant analysis, formulated within the field of statistics, paved the way for modern ML techniques [161, 162]. Other functions of ML algorithms cover areas such as clustering, binary or multi-class classification, regression, generative modeling, natural language processing (NLP), and dimensionality reduction (DR). In contemporary computational studies, ML techniques can be broadly classified into three primary paradigms: unsupervised, supervised, and reinforcement learning. Unsupervised learning aims to identify patterns in data without predefined categorization. In contrast, supervised learning uses category labels, during training phase to to learn distinguishable, often nonlinear, characteristics and relationships. Reinforcement learning, on the other hand, adopts a distinct, agent-driven iterative approach. Here, the agent refines its actions based on a reward feedback system, tailoring responses to unique environmental cues. These three archetypes define a wide swath of the machine learning field, without going into model architecture and training. Only specific methods used in this work will be discussed at this level of detail. Table 2.1 provides an overview of current ML algorithms and their placement within the extensive ML landscape. It is also important to note that many models can be trained to accurately predict known data, then applied to predict various values from unknown data.

Table 2.1: Abbreviations of commonly used machine learning models.

Method	Task	Paradigm	Abbreviation
k-Means	clustering	Unsupervised	-
Agglomerative Clustering	clustering	Unsupervised	-
Spectral Clustering	clustering	Unsupervised	-
Self-Organizing Maps	clustering	Unsupervised	SOM
Principal Component Analysis	DR	Unsupervised	PCA
Time-Lagged Independent Component Analysis	DR	Unsupervised	tICA
t-Distributed Stochastic Neighbor Embedding	DR	Unsupervised	t-SNE
Supervised Projection Learning for Orthogonal Completeness	clustering/classification/DR	Supervised	SPLOC
Naive Bayes	classification/regression	Supervised	NB
Support Vector Machines	classification	Supervised	SVM
Random Forest	classification/regression	Supervised	RF
Gaussian Mixture Model	classification/regression	Supervised	GMM
Artificial Neural Network	classification/regression/NLP	Supervised	ANN
Deep Neural Network	classification/regression/NLP	Supervised	DNN
Convolutional Neural Network	classification/regression/NLP	Supervised	CNN
Recurrent Neural Network	classification/regression/NLP	Supervised	RNN
Autoencoder	DR/generative modeling	Supervised	-
Variational Autoencoder	DR/generative modeling	Supervised	VAE
Generative Adversarial Network	classification/generative modeling	Supervised	GAN
Message Passing Neural Network	classification/regression/DR	Supervised	MPNN
Graph Neural Network	classification/regression/DR	Supervised	GNN
Graph Convolutional Neural Network	classification/regression/DR	Supervised	GCNN

Methodological advancements have been steady across various ML subfields, recently being applied to computational biology. However, the deployment of ML techniques in computational biology presents unique challenges. Typically, a significant constraint in building ML models is the requirement of extensive datasets for training. As the volume of data and variables escalates, algorithmic efficacy often diminishes, dubbed the curse of dimensionality [163]. Other prevalent issues include over-fitting, where the model becomes overly tailored to training data fluctuations, or under-fitting from insufficient sampling, such as when training samples possess reduced variance relative to the entire population.

2.1.1 Large Language Models

Large language models (LLMs) are a subset of NLP techniques utilized for sequence analysis and generation tasks, including text analysis, speech recognition, and AI-driven chatbot models. Initially relying on recurrent models like LSTM for sequence analysis, the advent of transformer networks, exemplified by models like Generative Pre-trained Transformers (e.g. GPT-3) and Bidirectional Encoder Representations from Transformers (BERT), has driven substantial progress in text generation by enabling contextual learning [164–166]. Transformer networks incorporate attention mechanisms that facilitate understanding input order, yielding valuable context and long-range correlations. These advancements have been particularly instrumental in the realm of natural language processing in computational biology, aiding in the analysis of protein and DNA sequences. Typical large language models train either by masking tokens and fill in or reinforcement learning on tasks.

The training objective involved in a masked language modeling, where input sequences are corrupted by substituting amino acids with a designated mask token. The network is then trained to predict the masked tokens based on the corrupted sequence. This objective is formalized as:

$$L_{MLM} = E_x \sim_X E_M \sum_i -log\ p(x_i|x_M) \tag{2.1}$$

Where the loss for learning, L$_{MLM}$, is calculated from predicting tokens, or characters, for masked, or removed, tokens from the input. The input sequence, x, is measured characterise-wise across the masked positions for accuracy to known correct token, and a probability calculated inside the logarithm. The summation is performed over all positions i in the set M of masked positions in the sequence, creating a type of entropy measurement for bad predicted masked tokens. Recently there has been wide adaption of these kinds of models for generative tasks, such as learning sequences that can represent proteins.

2.1.2 Machine Learning Applications for Proteins

Protein functionality is both designed for and depends on the folded topology. Accurate protein structure prediction (PSP) offers designing novel protein functions or enriching our understanding of mechanisms through simulation and experimentation. Until recently, the common computational approaches for PSP have been template-based and template-free modeling. Template-based modeling operates on the principle that amino acid sequences with similarities will exhibit analogous folds [97, 167, 168]. Consequently, empirically determined structures become templates to model unknown protein configurations, a process termed homology modeling. Often a single protein will have structures determined in the presence of a ligand, while others are derived without a ligand association. This reveals various structure conformations, some time multiple conformations with each being ligand-specific. This empirical data serves as a pivotal foundation for the refinement and testing of docking software, essential in drug design [169].

The critical assessment of methods of protein structure prediction (CASP) is a biennial contest where participants, through a double-blind process, receive amino acid sequences of structures already determined experimentally but not yet public. Competing research teams are expected to submit predicted models using their unique method of determination. Structures are evaluated using the Z-scores of backbone conformational similarity using multiple measures for accuracy. ML methods such as AlphaFold and AlphaFold2 have produced substantially better results than previous years of CASP. Many other groups have build similar sequence to structure models, like Baker and Zhang groups. As seen, ML has made great strides in the field of PSP, but there are competitive methods like Feig-R2, which used physics-based refinement through MD simulations [170]. The complexity of implementation and resources often dictates practical usage, however.

Design applications involve mutating residues to modify protein function or achieve a specific molecular property. Historically, biochemists first had to deduce functional sites to facilitate molecular engineering by allowing the organism to mutate in a directed fashion to some functional goal [76, 77]. Computationally assisted protein design has used combinatorial approaches, threading along backbones, and statistical knowledge-based methods have dominated the field [93, 171–177]. Early attempts to apply ML to design applied dimension reduction techniques, SVM, or Random Forest to predict mutation effects [178–180]. To assess protein stability empirical mutational databases, like ProTherm, catalogue mutations with attributes such as ΔΔG and changes in melting temperature. Various strategies, including naive Bayes classifier, K nearest neighbor, partial least squares, artificial neural networks, and deep neural networks have shown reasonable predictive power on this dataset [181–183]. However, the majority of these models exhibit limited accuracy on new sequences, primarily due to their sensitivity to the distinct characteristics of proteins in the dataset, predominantly lysozyme. Training datasets over-representing specific protein families leads to poor generalization to other families. It has been noted that generalizing a mutational stability model must be done with great care, as protein families may express many methods of stabilization [184].

Advancements in predicting mutation induced stability changes of a protein include hallucinations of diffusion models applied to protein structures [185]. These were found empirically to generate novel composite single and multi-domain globular proteins. Improving stability of structural space from the relationship to sequence space was performed using a message-passing neural network, with empirically deduced success [186]. Methods that minimize the number of mutations needed to shift stability are available, which can also predict solubility factors [187, 188]. Methods using ML in conjunction with methods such as alphafold, Rosetta, or MD simulation evaluation to assist in finding possible realistic folding proteins [189]. Even empirical characterization methods are assisted by ML. A Gaussian Mixture Model that approximates the electron density map, which correlates to atom probability, produced from cryo-EM which can detect distributions of conformations of structures [190].

2.1.3 Machine Learning for Functional Dynamics

Protein function analysis often deals with high-dimensional data and low statistics; a difficult situation for any method. With increasing hardware capability the data sizes generated are hitting a transition to a convergence between ML and computational biology. Application towards MD is a natural step in research.

To maintain reliability in the dynamically changing network of chemical reactions, proteins evolved to maintain a functionally relevant conformational ensemble that satisfies thermodynamic and kinetic stability conditions [28]. While protein properties inherently depend on their environment, focusing on intrinsic properties is often effective, as most Earth-bound life operates under similar molecular constraints, regardless of extremities like hot springs or deep-sea conditions. This can be observed for a given family of proteins from organisms living in diverse environments that share the same biomolecular function [191]. It is likely that these family members also share some dynamic processes that are inherently caused by the structure they inherit.

At its core, ML intertwines deeply with statistics, leveraging automated procedures for statistical inferences. For instance, the principle of projection pursuit underpins a vast array of ML methods, such as principal component analysis (PCA), independent component analysis (ICA), factor analysis, and linear discriminant analysis (LDA). Clustering is another crucial facet of ML, with numerous methods designed based on varying objectives. Presently, numerous ML techniques are intertwined with statistical analyses, optimizing automated processes to derive meaningful statistical insights [192,193]. Unsupervised machine learning methods have played an important role in the analysis of MD trajectories. Principal component analysis (PCA) applied to MD data provides dimension reduction (DR) which can characterize the essential dynamics of macromolecules such as proteins [194–196] while reducing the degrees of freedom (df) for the data matrix. PCA identifies large-scale motions that are assumed critical to function; consequently, functional motions will be misidentified as noise if the dynamics have a smaller variance than what is contained in the top PCA modes [197]. Clustering algorithms are often combined with DR and feature extraction techniques, such as PCA, in order to identify key conformations that facilitate molecular function [198, 199]. Supervised machine learning is capable of identifying functional dynamics by associating experimental and simulation data [200] in binary classification. In literature, supervised ML techniques for discriminant analysis [201,202] such as linear/quadratic discriminant analysis (LDA/QDA), SVMs, and gradient-based NNs have perform poorly on MD data. As such a method that focus on streams of data and separating entire modes of motion might perform better at classifying functional proteins from dynamics.

2.2 Generative Machine Learning for Proteins

ML approaches have been widely applied in the field of medicinal protein biologics. Typically these methods use pattern recognition algorithms to discern relationships between empirical observations, such as protein secondary structure prediction, drug repositioning, and drug design [203]. These methods have garnered considerable interest from computational and medicinal chemists, many reviews related to the applications of ML or deep learning in drug design and discovery have been published [204–212].

Popular generative models currently include generative adversarial networks, diffusion DNNs, and variational autoencoder (VAE) [213]. Essentially, these methods discerns patterns from the training data and subsequently predict novel data points from learned relationships embedded in the network weights. One of the most impactful applications for these generative ML models are sequence to structure predictions.

Template-free modeling generates protein structures without the use of homologs in the PDB [214]. This method typically relies on physics-based energy functions and is much more computationally intensive than template-free modeling. Knowledge-based approaches dominate the field, yielding deterministic designs for stable proteins or homology-based structural predicts [173, 215, 216]. Of course, there are algorithms that combine the these approaches. Template-based and template-free modeling have become more seamless, and combinations of the two approaches are more commonplace [97, 217]. It is worth noting that many proteins cannot be crystallized because they have intrinsically disordered regions [218]. But with structures to work from mechanisms of action can be explored using computational methods allowing rapid therapeutic development, also known as computer-aided structure-based drug discovery (SBDD) [219–222].

2.2.1 Sequence to Structures

Debuting at the Critical Assessment of protein Structure Prediction (CASP)-Round XIV, in 2021, AlphaFold2 from DeepMind has raised the bar by achieving prediction accuracy near the experimental limits, far better than their competitors [223]. AlphaFold2 is a completely reworked model of the AlphaFold method presented at CASP13 [224, 225]. The success of AlphaFold2 comes from the unique implementation of deep neural networks and MSA derived evolutionary history [217,223,226–228]. The introduction and promise of the success of AlphaFold2 has elicited similar works such as RoseTTAfold, a “three track“ neural network based off the N-C-C backbones of proteins [229] or an AlphaFold reproduction using popular libraries like PyTorch, such as Openfold [230]. Still many areas of improvement are needed, such as accounting for the role of the environment or predicting structures without the need for multiple sequence alignments (MSAs), as seen in OmegaFold [231]. There are many ML inspired protein structure related works appearing. Using ML models to guide physics based modelling or to predict realistic structures from rapid diffusion based

ML models are a drop in the continuous research being produced by this applied niche.

2.2.2 Structure to Sequences

Original methods for structure to sequence prediction seem to have boiled down to empirically testing new sequences either via inferring from mutation scanning or just trial and error with expression and characterization. Computational methods were limited to sequence based methods or atomistic simulations of mutated structures. An early adopter of this process was the method GREMLIN, where a statistical model of a protein families attempted to interpret both conservation and co-evolution patterns to predict residue–residue contacts using evolutionary covariance information. However, these methods require large numbers of evolutionarily related sequences to assess the extent of residue covariation, and the larger the protein family, the more likely that contact information sufficient [232]. Once again, reinforcing the structure bias present in a variety of databases.

There are multiple ML based structure to sequence models, typically trained features are either sequence based or structure based, but the output is always possible homological sequences for the given structure. ProteinMPNN is a message passing neural network (MPNN) that predicts protein sequences in an autoregressive manner from N to C terminus using protein backbone features, distances of Ca-Ca atoms, relative backbone frame orientations, and a variety of other geometric properties. Recent work with ProteinMPNN has also shown the ability to rescue de novo designs as well as allowing the design of complex self organizing oligomers that form nanoscale structures [186,233]. Recently there has been a rise in large language models (LLMs), partly due to advances in model architectures like transformers with attention [234]. Many researchers have attempt to modulate sequence to sequences [235], some also include structure in this consideration. The Evolutionary Scale Model (ESM) is a LLM trained on protein sequences, and can achieve inverse folding using a sequence-to-sequence transformer with invariant geometric input processing layers. Reported native sequence recovery on masked backbones and buried residues are better than compared methods in sequence generation [236]. The ESM model also generalizes to a variety of more complex tasks including design of protein complexes, partially masked structures, and binding interfaces. [237]

2.2.3 Full Designs

The computational biology underpinnings of in silico structural based platforms that proceed the current state of protein design have been detailed and largely dealt with pseudo physics or homology centered approaches [93, 171–177]. Computational fixed backbone designs have typically lacked predictive power for design, possibly due to dynamics or flexibility of the protein [95]. Early attempts to solve this problem applied dimension reduction techniques or SVM [178,179]. Application of ML models to the curated ProTherm dataset, empirical mutational database that includes the attributes of $\Delta\Delta G$ and changes in melting temperature, have obtained fair predictive power for the dataset, when compared to naive Bayes classifier, K nearest neighbor, partial least squares, and an ANN [181, 182]. A deep neural network predictor of $\Delta\Delta G$ using this same training dataset outputs $\Delta\Delta G$ values for mutated sequences, as determined by multiple sequence alignments [183]. This model works well for represented families, but poorly for unseen protein families, unsurprisingly. Optimizing protein interfaces through mutagenesis has been successfully achieved using a Random Forest approach [180], which compared favorably to other mutagenesis predictors such as SKEMPI. Unfortunately, most results to date use training datasets that over-represent specific protein families, which leads to poor translation to other families. It has been noted that generalizing a mutational stability model must be undertaken with great care for each specific problem addressed [184].

Most recent advancements in predicting stability changes of a protein due to mutation include hallucinations via diffusion models applied to protein structures, language models, or geometrical CNNs [185, 187, 188, 238]. Pipelines are being developed that piece together ML packages with evaluation methods such as alphafold, Rosetta, or MD simulation to find possible realistic folding proteins, some open sourced like ColabDesign [189]. There are a few groups that focus entirely on design application. Baker group has created two variations of design using ML that hinge on sequence to structure models and fully design sequence and structure. A sequence based CNN that hallucinates and refines sequences, trRosetta [239]. Then a more formalized and classic Rosetta style RFDesign using RoseTTAFold [185]. Most recently, a fashionable diffusion method implemented in RFDiffusion [240]. Which can be combined with structure to sequence models to generate a variety of candidates. Furthermore, the generated sequences can be passed into classification models based off sequences to refine the outputs and increase success rates of empirical testing [241]. Full design methods also extend to entirely sequence based, such as the progen model [235].

CHAPTER 3: pepStream

3.1 Motivation

The pepStream software is a pipeline of bioinformatics methods and analysis that produce peptide candidate that bind to a protein targets. The original proposal for the software came from a cooperative effort between UNCC and Plant and Food research in New Zealand. In their research they had come to find a need for binding peptides for a specific target. Their target, however, lacked a robust set of structural data due to being a MoRF inside a disordered segment of this protein. The structural signature of a MoRF within protein-protein interaction is often a crucial step in controlling important biological functions [56]. Therefore, the development of a peptide to bind to a target MoRF is hypothesized to be an effective way to interfere with certain protein-protein interactions through competitive binding.

3.1.1 A Beta-Strand Library

The initial inspiration for pepStream was the work from the Vendruscolo lab, where a pipeline of designed beta-strand motifs were generated from a library of beta strand sequences. This library was constructed simply from from the PDB with a 90% identity threshold, then taking the DSSP labeled beta-sheet components and creating a library of sequence pairs that could be searched against the target subsection [242].

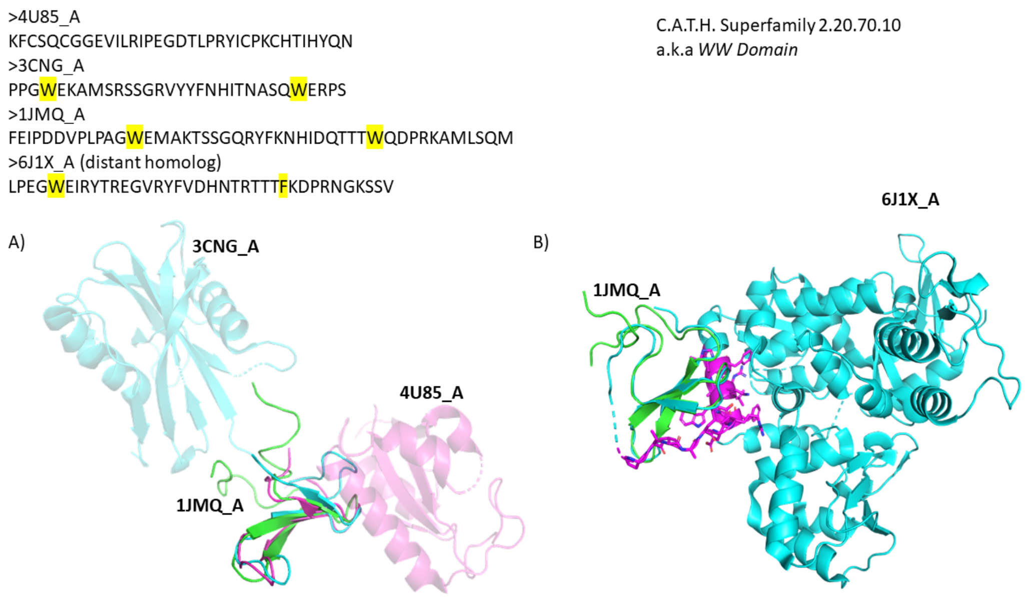
Antibodies are Y-shaped proteins with variable regions at the tips of their arms, where loops, called variable loops, specifically recognize and bind to foreign substances like pathogens. These beta-sheet sequences were subsequently mutated into binding loops of antibodies where the beta-strand motif was design against a particular segment of a protein. They showed that this library approach worked for making a specifically targeting interaction to the loop of the target. In a full size antibody a small segment of protein forming a beta-strand to a target can be stabilized by the rest of the structure, as antibodies tend to have large interaction surface areas at their binding domains. These large binding domains and structures can enable tight binding, micro to pico-molar affinities, and low immunogenicity that make antibodies effective [243].

A few logical shortcomings were found in this methodology. Beta-sheets are almost always found in buried segments of protein, and in beta-barrels there is a dichotomy between the inner facing residues of the beta-sheets and the outside. Intrinsically these kinds of secondary structures rely on stabilizing interactions to the side chains to keep the backbone interactions stable, and are long range with respect to sequence and become more stable with longer lengths. For a binding peptide, it was thought that this constraint would make it difficult for the peptide to bind to. Additionally, restricting peptide structures to beta-strands only seemed difficult to generalize to any binding target on a protein. For example, a local substructure that forms an alpha helix has enthalpy in it's nearby secondary structure and will maintain this structure for relatively long time spans, even without any flanking residues [49, 244]. Perhaps the most salient issue; there was a lack of sufficient information to reproduce the library created for this process nor was the library provided in the published work, a common issue in bioinformatics.

So taking the fundamental approach, the core functionality behind this library was basically looking for deep sequence homology in multiple structures. In the case of the beta-strand library, a small subsection of the target is used to search for a portion found in a beta-strand secondary structure, where the complement strand is used to create the binding segment. This is performed multiple times over a window to create the full length binding segment of interest. If considered as a search problem, in this instance the target sequence is used to inform the search for a set of residues that is complimentary to the subset, combining all the subsets to form a final binding motif as a straightforward problem [242]. The most obvious difficult challenge to generalize this kind of approach is creating a library of all kinds of secondary structures, or lack-thereof, to perform this kind of operation over seemed infeasible.

3.1.2 Approach of pepStream

Many design strategies struggle dealing with sequences distantly related to human proteins. The initial stage of peptide design rests heavily on accurate structure prediction, a feat that often necessitates multiple crystal structures to discern the intricate conformations of proteins. Historically, the foundational tools for protein design like MODELLER operated on direct homologous mapping, while others like Rosetta and I-TASSER used threading approaches. These tend to lack accuracy for proteins that diverge from the PDB. Although tools like Rosetta offer a myriad of capabilities, they come with their own set of challenges, such as their commercial restrictions, heavy reliance on initial structures for design, known complexities with peptides, and the imperative need for specialized knowledge to adapt the tool to novel challenges.



```
>4U85_A
KFCSQCGGEVILRIPEGDTLPRYICPKCHTIHYQN
>3CNG_A
PPGWEKAMSRSSGRVYYFNHITNASQWERPS
>1JMQ_A
FEIPDDVPLPAGWEMAKTSSGQRYFKNHIDQTTTWQDPRKAMLSQM
>6J1X_A (distant homolog)
LPEGWEIRYTREGVRYFVDHNTRTTTFKDPRNGKSSV
```

Figure 3.1: The WW domain is expressed widely in the proteome and is often found at the interface between monomers. A) Very distantly related sequences can still hold same function shape, shown by no sequence identity match in the 4U85 structure, but the WW domain is aligned and functions. B) Example of a distant homolog (PDBID 6J1X), which would not align well, having a protein–protein interaction not present in the original structure (PDBID 1JMQ).

Figure 3.1 graphically shows that a structure can be similar in conformation and functionality to another with low sequence identity. In both cases the distant homolog looks nearly identical in terms of fold, but lacks sequence characteristics conserved in the majority. A fragment in contact with the distant homolog likely interacts favorably, at least in the native protein. Taken out of context there is still a possibility that the fragment could bind to the original target. And given the hypothesis that these fragments might bind, as evolution tends to find the optimal binders, by predicting the sequence fragments as a peptide and then attempt an analysis to assay which are best. Computational protein docking is reasonable to test the hypothesis.

In the knowledge base approach from Vendruscolo et al., the entire database must be searched for every sequence window that is searchable for the target. This creates an increasing complexity as the target itself becomes larger, leading to a limiting behavior dependant on the time to search a single window and any subsequent blending of the resultant sequences. Instead, the target sequence itself could be turned into homological subunits to be used for searching inside an empirical database. This is particularly helpful for the cases where the target has little or no structural information itself, which tends to occur for disordered proteins. This approach postulates that exploring distant homolog targets in sequence space might unveil evolution design sequence that bind to a specific segment of protein. Data mining these distant targets in sequence space and searching for related structures allows simple peptide generation that can offer a reasonable binding to the target segment.

3.1.3 pepStream Version 1 Tests

A comprehensive peptide prediction process was developed using an integrated pipeline, predominantly implemented through Linux and Python, which automates the generation of peptides for various target sites on proteins through scripts that manage tool usage, data formatting, and analysis. The pipeline utilizes publicly available software such as BLAST, I-TASSER, and ZDOCK as well as custom-designed approaches to extract fragments of proteins that interact with particular regions with similarity to our target sites from protein structural databases. This prediction pipeline typically results in hundreds of possible binding peptides and therefore a mechanism to filter these peptides by ranking them was required to make workflows manageable. Collation of the resultant data on the candidate peptides enable the implementation of a ML approach to improve peptide binding ranking predictions for the multi-classification.

This was performed in the following way: in addition to developing the more specific docking scores developed directly from our pipeline, a series of other predicted biophysical properties were also calculated for each of the subset of selected predicted peptides including factors such as solubility, stability, charge at neutral pH and the isoelectric point (pI) of the peptide. Multiple dimensions of clustering were then performed with all of this data using principle component analysis in an unsupervised ML exploration of candidate peptides. By plotting both wet-lab “successes” (specific inhibition) and the different types of “failures”,non-specific binding and non-inhibition, onto these multi-dimensional frameworks we implemented supervised ML method to learn and predict peptide functionality classification in the output, Figure 3.2. Multiple systems were evaluated using this method. In both systems the target(s) of interest were usually MoRFs in disordered domains of signalling hub proteins DELLA, p53, and partners of p53.

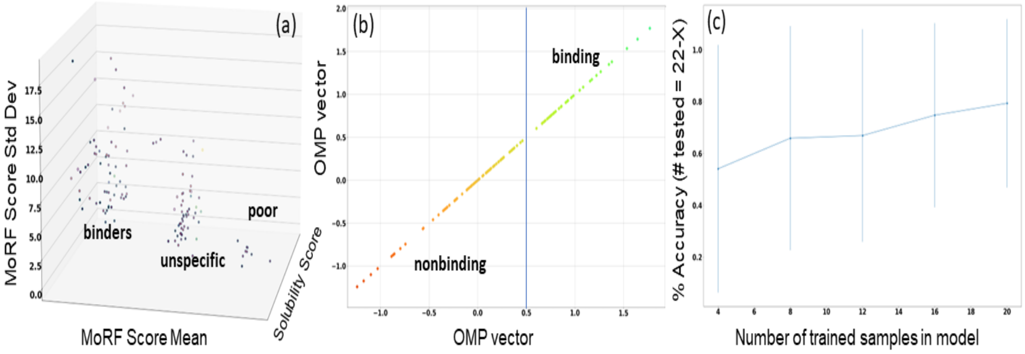


Figure 3.2: A) the raw pepStream output for the predicted peptides plotted in three dimensions to demonstrate the predisposition of separation in the dataset. B) A single dimension projection of a trained OMP model on the same dataset. C) Accuracy as the training set is increased during training of the OMP model for classifying the predicted peptides.

3.2 Methods

Casting the peptide design problem into a homology problem connected the concept of the twilight zone of sequence homology [245–247]. There are some instances of proteins having the same fold with as little as 20% sequence identity. In some instances the interactions that allow a fold to stabilize might be shared, affording a reasonable peptide interface to a different homolog target. While it may not always be the case, it is an easier search than the exhaustive combinatorics 20^n for all possible sequences of a protein length n. This also allowed the problem of finding a binding peptide to an arbitrary protein target to become an inverse problem, as possible answers are in the PDB but which is best is unknown. Alleviating the need to create an additional database of some sort to produce binding peptides.

3.2.1 MSA to Diversify Target Sequence

Typically, directed evolution experiments deal with trying to evolve a protein against some manufactured pressure. To do so, a sequence library gets created by generating a large number of mutations in the protein sequence, which results in a diverse population of protein variants [248,249]. This library is then screened for variants with improved function or desired properties, allowing for the selection of the best-performing variants. To maximize the results only certain parts of a protein are allowed to mutate, while the other parts are kept fixed. Determining these mutation-allowed regions is commonly performed by producing MSAs of the protein, identifying conserved and variable regions, then targeting the variable regions for mutation.

Considering physio-chemical similarities when mutating amino acids, the natural mutation range in protein sequences is used to identify a diverse set of homologs in structural databases. This approach has proven useful when applied to protein homology. In particular, a 30% sequence similarity is often sufficient to maintain similar secondary structures [245–247, 250]. The twilight zone of the sequence-structure relationship points to a transition region where sequence-structure relationships begin to fail. Although there is no reason to expect the sequence-structure relationship exhibited in proteins to apply to peptide versions, we wanted to investigate whether or not these distant homolog motifs will reveal complimentary interface structures and conserve favorable interactions. Therefore, the diversification of the query sequence is performed using open tool set BLAST, Basic Local Alignment Sequence Tool, along with specified libraries of protein sequences. This expands the input by many folds, allowing a wider sequence search space.

A limiting factor was finding enough diversity to escape the target sequence itself. In the initial trails of this running MSA with reasonable settings produced a low diversity of sequence for the high conserved MoRF. Sensibly, this is by design, the probability of matching to distant sequences is low when the MSA protocol is optimizing closeness of the sequences by identity or bit-wise likelihoods, like e-values in BLAST. Naturally the next step was to raise the allowable matches, in BLAST this is done using the e-value threshold. Typically this value is quite small to generate better than random matches. But performing this sort of search, only on a small subsection of a protein, created difficulties producing hits. I overcome this by providing BLAST an absurd e-value of 100-1000. It was found that running this over an entire sequence created long run times and produced fewer sequences than expected simply due to the difficulty for bit-wise matching of a longer sequence to any other sequence, even with a higher cut off. So using a small 20 to 40 residue target window was tested, effectively treating the target as if it were a protein structural domain, like the WW domain.

To assist in escaping the local sequence space of the target sequence a set of subsequences are generated. These footprints are described by a portion of the target sequence, selected from bifurcating the sequence recursively until a minimum length is reached. These footprints are used to assist in the creation of a mutational library for the target sequence, which will be used for structural searching. While a sliding window across the original target sequence was considered, creating an algorithm to stitch this together in the mutation stepped proved difficult to justify. So simple bifurcation of the sequence was applied to generate enough sequence diversity of mutant sequences from the original target sequence. All of these footprints have an MSA applied from returned hits, and then any gap in alignments filled with the original sequence. Effectively it has been observed that this expands a target sequence from what would be a shallow alignment of maybe a few hundred sequences to thousands or more. In the computation of creating a mutant sequence, a ad-hoc entropy can be calculated from the position specific scoring matrices, PSSMs, that come from position specific iterative (PSI) BLAST. This creates a metric to show clearly that as the sequences are generated farther from the base footprint they are more random. Understandably, only an upper portion of the sequences are used, typically something below 100,000 sequence.

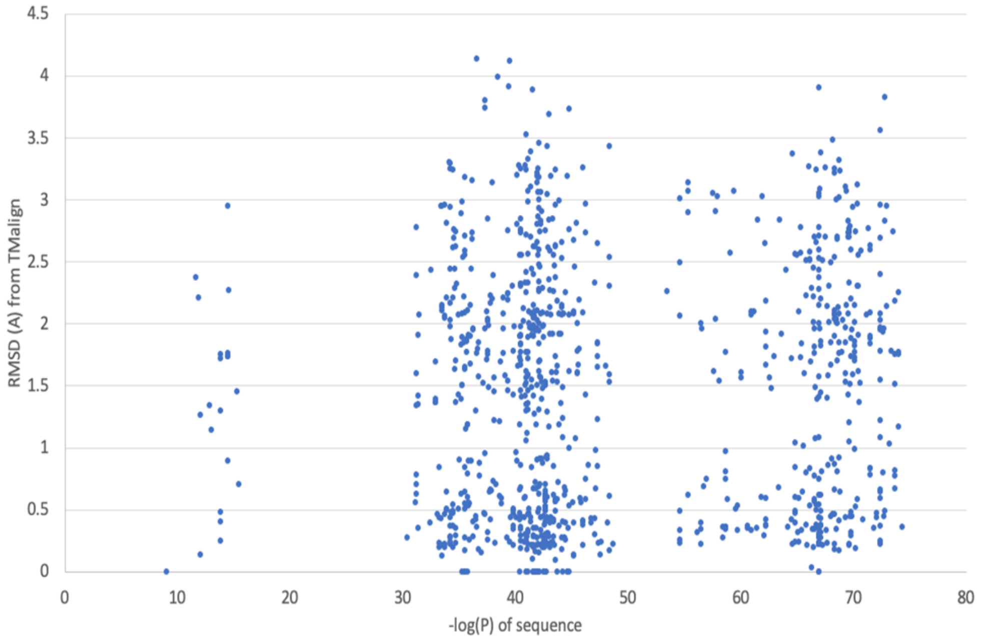


Figure 3.3: Plot of typical mutant footprint match to actual input structure RMSD versus the score, Σ-log(Probability of sequence), of the mutant sequence.

After taking the mutated sequences and retrieving aligned hits from the PDB, each matched distant homolog can be superpositioned to the original target sequence structure. This provides a metric to asses how close they are in secondary structure. To quantify sequence divergence, a simple score can be computed using the formula: Σ-log(Probability of sequence), where the probability is derived from the transition weights in the PSSM during the diversification phase. Observably there is a weak to no dependence of RMSD with respect to the scoring function for distance of the sequence to the original target sequence, Figure 3.3. The interpretation of this result is broken down into two considerations. Firstly, the expansion of the sequences yields similar structures across the distribution of mutated footprints. While these sequences gain more variance as the free energy of the sequence increases, or the sequence diverges from original input, similar structures are found that would otherwise have been missed without this expansion. This null relationship is discernible by the unique structures at the higher x-axis value keeping roughly similar RMSD in Figure 3.3. Secondly, this is complicated by possible low conserved areas or just a longer length of the input target sequence. This methodology is limited in sequence length for the input. For all targets tested in this diversification, the expansion allowed for two orders of magnitude more structural matches than what can be extracted by running the original sequence into BLASTp. As a result of this diversification, analogous structures suitable for contact mining peptides are generated.

Once the target sequence was expanded to a large number of related but distance sequences it was straightforward how to apply the inverse search. Given the mutated target, can a structural match be found? If so, then the protein residues around that match can be mined for potential binding peptides. A protein BLAST of all the mutated footprints was applied to only the sequences of PDB structures. Hits were pulled into a simple neighbor searching protocol to look for protein residues around the searched footprint-match only. Effectively this portion would be a distant homology, and the protein residues around it are fragments that could be binding peptides. Initially a size minimum of 9 residues was used to find potential binding peptides, or complimentary MoRF peptides; dubbed cMoRFs.

3.2.2 Generating Possible Peptides

Evolution favours combinations of molecular interactions that support biological function while simultaneously satisfying physical constraints that make the interactions possible. As such a search of interacting segments of protein from structures of interest has become a common practice in rationally designing peptides and is often effective [75, 152, 153, 155, 251]. Obtaining the interacting fragment from empirical data is the first step. Using the python package Biopython, a routine is employed that captures segments of amino acids that interact with a given reference of residues [252]. Thresholds to identify various levels of intermolecular interactions inside proteins incorporate user-defined cutoffs for finding contacts. There are opportunities to incorporate a process in this particular code to interpolate missing atoms, remove repeat region candidates, or create longer peptides.

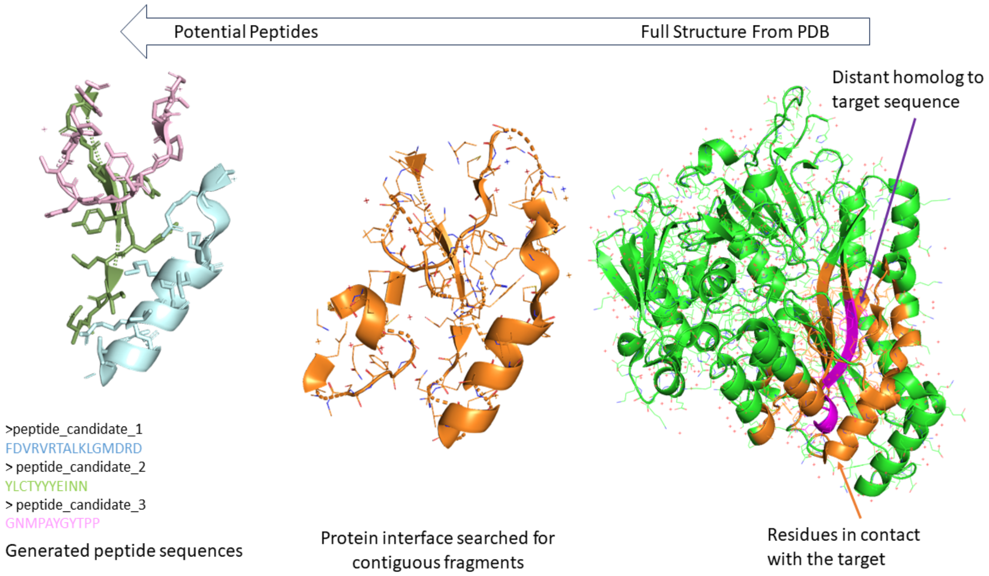

Potential Peptides
Full Structure From PDB
Distant homolog to target sequence
>peptide_candidate_1
FDVRVRTALKLGMDRD
> peptide_candidate_2
YLCTYYYEINN
> peptide_candidate_3
GNMPAYGYTPP
Generated peptide sequences
Protein interface searched for contiguous fragments
Residues in contact with the target

Figure 3.4: Automated peptide selection diagram from described python code in pepStream.

In Figure 3.4 an illustration of the output is given. Essentially the matched subsection, distant homolog, is the core for a n^2 neighbor search at heavy atom level. A radial algorithm returns all atoms interact with the matched sequence in a given structure. There are two variables that modulate detection of an interfacial interaction. The distance at which an atom will be included as being a contact and interface distance is a buffer distance to expand from contact atoms to capture backbone atoms or lengthen predicted peptides. Once the lists of interacting atoms are made the corresponding residues are then found, as highlighted in orange within Figure 3.4. The selected residues are then analyzed for contiguous length requirements, those that are long enough are passed as potential peptides as seen far left in Figure 3.4. At this point the sequence of the potential peptide is recorded.

3.2.3 Predicting Peptides and Docking

The tertiary interactions within proteins often stabilize specific conformations of small segments. When isolated from the larger protein context, these segments might adopt different conformations, or they may retain their original structure. The outcome is unpredictable. Therefore, when attempting to design a peptide based on such a segment, modeling its potential conformations becomes essential. Peptides are known to fluctuate in conformation, unfolding and folding rapidly in solution [244]. When modelling, more conformations is always be better. Typically this is best accompanied by some method to ensure the conformations are relevant to the free energy landscape or score function of interest.

For each contiguous peptide derived from the previous step a multi component modelling effort is automatically performed. The I-TASSER software has the advantage of both compute time and ease of input allows for a highly parallelized process, taking advantage of an HPC environment. The input is a fasta format file, defined from the empirical peptide found from the previous step. While smaller proteins tend to be more disordered, some conformations may be longer lived than others; accounting for conformation diversity during docking is important to ascertain an accurate binding affinity to the target binding site. Docking small molecules to proteins is a major research field, however open tools that can dock peptides to a target site on a protein are limited due to degrees of freedom between the pair can incur high cost of the calculation. In addition, a docking routine for multiple conformations interacting with a flexible interface is necessary, which often is not done in docking routines barring those that include MD. Instead, we used the I-TASSER software is used to create multiple structural models of the peptides, which constitutes all or part of the conformation sampling in the proposed approach. I-TASSER also offers variables to control for similarity or homology threading in the software's predictions. While this is a shallow exploration of conformational space, it is sufficient for early investigations and to demonstrate the pepStream workflow.

Docking a peptide against a protein is a complicated task that has been attempted by a few different studies. In Rosetta FlexPepDock a full conformational simulation is used to ascertain docking potential, at the cost of time. In the beta design of pepStream we decided ZDOCK was a suitable docking software to apply the routine described below [131,139,253]. This was largely due to the computation time and the combinatorics sizes of targets against predicted peptides. Each model of peptide is docked against each model of the target structure. While this work is focused on the computational design and characterization of the binding peptides, the target structures can be provided and design by the user, or even taken directly from empirical structures. To overcome the rigid docking multiple models are used of both the whole structure of interest and the peptide predicted to bind to the target.

Accounting for unspecific binding is performed by docking peptides to the target proteins at added off-target binding sites for the ZDOCK simulations. In this effort a simply set of scores can demonstrate if the peptide of interest has a higher than normal affinity to the target residue subset versus a random segment elsewhere on the protein. This physical filter is performed on two levels of specificity. One running over the entire target to test for overall binding, performed against all output peptides as an initial sweep. This affords a ranking for top candidates, which are then used in a specific docking study. The next step breaks the target protein into segments and runs docking over each subsection to effectively generate a comparative map to identify relevant and non-relevant sections. This encompasses the entiretly of the physical filter portion of the pipeline. Ranking the best binding peptides to the target and testing the hypothesis that the candidate will bind specifically.

The initially implemented pepStream algorithm is shown below, Figure 3.5.

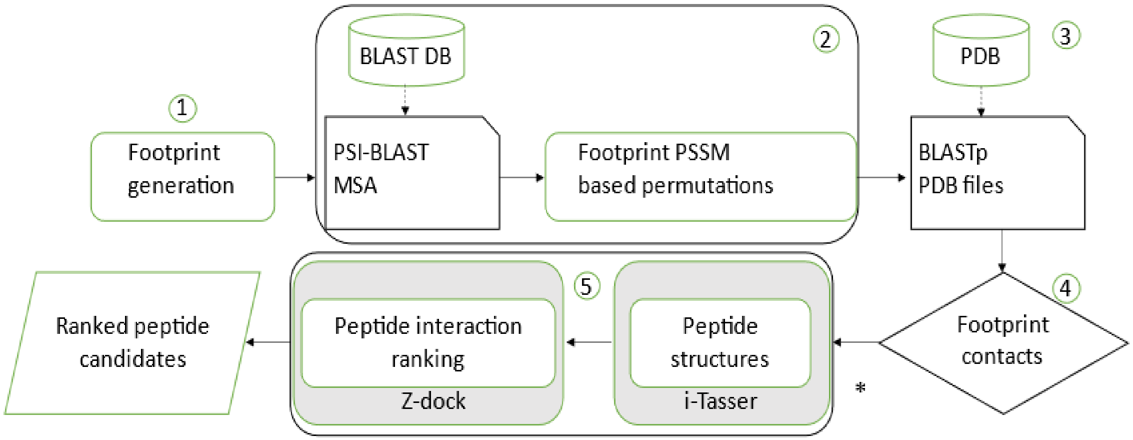


Figure 3.5: The original flowchart for pepStream pipeline. Software is primarily a mix of python and bash. *) filters are applied to any found sequences to avoid modeling issues.

(1) A binding target is specified as a sequence of residues input in fasta format. This is the initial footprint for the binding target. The inputted target sequence is partitioned multiple times before developing a sequence profile.

(2) PSI-BLAST with variable e-value thresholding and database is used to generate the mutant sequences and a position specific scoring matrix (PSSM). The mutant sequences, footprints, are a combination from actual hits with any gaps filled in with the original residue.

(3) These mutated footprints will be queried in BLAST against the PDB, or any specified subset of it.

(4) Structural hits will be analyzed in a routine to locate relevant sequences that interact with the matched footprint.

- Auxiliary filtering criterion based on sequence properties, length, and other problem dependent factors can be performed here.

(5) Sequences from structural interaction fragments are modeled in I-TASSER suite. The peptides structures are then docked onto the target structure using ZDOCK

against all unique pairs of generated unperturbed target structures and candidate peptides.

3.3 Results

3.3.1 Results on DELLA

Focus on applying a variety of methods to understand the dynamics of the 2ZSH structure, a complex of GID1A, in the apo and holo states, and DELLA protein interaction domain from a GAI family protein. This work included MD, mDCM, dimensional reduction and subsequent clustering, and then the modelling of peptides interacting with the disordered DELLA domain. There was mutual interest in using small peptides to induce a conformational change and disrupt the GID1A interaction. Similar endeavors have been highlighted in various studies, including research on a peptide binding to COVID [40, 136, 254].

The 2ZSH system, which is a protein complex consisting of GID1A with GA3 bound, and the DELLA segment of the GAI protein, exhibits a conformational space that is considerably larger than anticipated. The GAI protein is disordered, observable from various simulation, and even the ordered DELLA segment is extremely flexible. Careful modelling of the disordered protein was done for initial binding experiments by running multiple simulations from different initial conformational states. The cohort offered a set of experimentally characterized peptides in terms of binding towards the DELLA segment of GAI. Early results in ensembles of docked structures found rank-ordered scoring outputs that were closely aligned with binding affinities, thereby justifying quantification of potential peptide candidates in silico.

A series of peptides were predicted by the peptide pipeline. Using the original designed peptides with known classification as training, a ML method to predict peptide functionality classification in the output was trained. It was rapidly noticed that the feature space of the binding scores to the target protein segregated the peptide candidates. This was largely factored by the biophysical properties of the peptides.

In Figure 3.2 taking the dataset and training a projection pursuit model called orthogonal matching pursuit a low dimensional representation of the classification in a subset of the data can be used to project the entirely of the data into a classifier.

At this point a set of the peptides screened through pepStream were selected to test their bioactivity in vitro. Previous research had developed an antibody that can specifically bind to the target motif within the DELLA protein. The antibody is attached to a Biacore chip, and the chip can then measure the binding of the DELLA target to this antibody with a Biacore X surface plasmon resonance biosensor. A peptide that can bind to the same region should inhibit the antibody-DELLA binding reaction. The assay measures the ability of a ten-fold excess, by molarity, of supplied peptides to out-compete ,inhibit, the binding of the antibody to the DELLA protein motif. This biophysical data was used to develop the classification of the peptides as either non-inhibiting, non-specific binding, or specific inhibiting peptides. Surface plasmon resonance binding data was collected for the 10 top predicted peptides from pepStream and 3 bottom. It was shown that seven of then ten top candidates specifically bound, or successfully inhibited, and the other three seemed nonspecific binding. The three lowest scoring predicted peptides were found to be nonbinding.

Experiments were performed to explore how these same peptides might affect plant systems in vivo. For the exogenous peptide application assay 15 random leaf discs approximately 2 mm in diameter excised from fully expanded, non-senescent Arabidopsis thaliana rosette leaves were incubated under constant darkness followed by 2 days standard growth room conditions (16 hr light / 8 hr dark) prior to total chlorophyll extraction using acetone. The leaf discs were immersed in 450 μL of 10 μM synthetic peptide solution using HBS-N buffer as the diluent. A solution with GA3 was used as the positive control which promotes leaf senescence in A. thaliana via the GA-DELLA-GID1 complex was prepared to a working concentration of 10 μM using 100% EtOH followed by a final dilution using HBS-N buffer. Senescence of the leaf material was determined by measuring the total concentration of chlorophylls a and b post incubation with and without the small inhibiting peptide candidates. Essentially a refined solution of each well was measured spectroscopically, shown in Figure 3.6. Qualitative comparisons of the peptides to the control case made it clear that multiple peptide candidates inhibited the senescence in the leaflets.

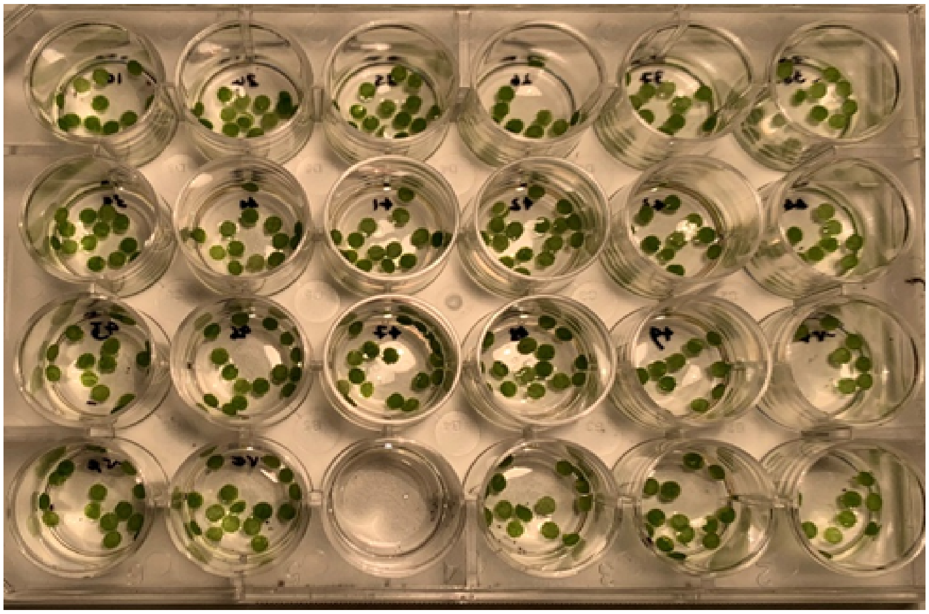


Figure 3.6: Experimental set up of exogenous peptide application to excised A. thaliana leaf discs to measure effect upon leaf senescence of peptide inclusion. Performed and pictured by Dr. Marion Wood of PFR.

3.3.2 Designing Interfering Peptides for p53 and Binding Partners

In continuation of developing the pepStream method a new model system was selected. The human tumor suppressing and transcriptional hub protein [251, 255, 256], an IDP also called p53, and has a significant effect on multiple types of human cancers and some virus infections [257]. The p53 protein has several MoRF sites that regulate different cell growth mechanisms for different protein-protein interactions, as shown in Figure 3.12. Published results for peptides binding to p53 demonstrates the feasibility of a peptide-based drug [251, 258]. Over-expression of any partner of p53 or having a misfunction, in either p53 or its partners, leads to various cancers.

In this project, mouse double minute 2 human homolog (MDM2) and the human deacetylase transcription regulation protein called sirtuin (SIRT) will be studied as well. The p53-MDM2 interaction is related to oncogenesis. The MDM2 protein inactivates p53 either by physical inhibition at the N-terminal, ubiquitination, or export from the nucleus. The signaling protein SIRT1 are ancient proteins found in most cellular life; recently gaining popularity as possible anti-aging targets due to their control of DNA repair. These proteins participate in essential pathways that lend themselves as targets for peptides that interfere with p53-MDM2 and p53-sirtuin interactions. Various methods for synthesizing peptides and introducing peptides into living organisms are available [59, 75]. A successful outcome was considered as the design of three or more peptides that inhibit each of the four target binding sites, 12+ peptides in total, involving two MoRF-mediated interfaces.

The hub protein p53 has 393 residues, with 37% intrinsically disordered mainly near the N-terminal transactivation and C-terminal regulatory domains [259–261]. The N-terminal partner MDM2 and the C-terminal partner SIRT1 have sufficient parts of their structure crystallized regarding the binding of p53, and are successfully used in silico and in vitro studies [251,262]. Thus, focus was placed on the binding sites of p53-MDM2 and p53-SIRT1, which have sufficient structural coverage for docking methods. A similar situation occurs in the DELLA IDP, where only the first 110 residues are modeled, with the rationale that the targets have characteristics of the MoRFs previously studied, which are known to be activated by peptides. The abundance of research on these proteins, together with the relative ease of purchasing p53 and many of its binding partners, form the basis for why the p53 protein was identified as a model system with its binding partners. Moreover, mutant variants can be considered for future studies aimed at extending specificity requirements. Specific mutated targets that can be controlled are an essential aspect of personalized medicine, and the p53 system can also be used to explore these ideas.

This work aims at obtaining potential peptide-based drug candidates that will disrupt certain out of control protein-protein interactions involving p53. Promoting pepStream as a platform capable to rationally design peptide-inhibitors interfering with a specified protein-protein interaction. The protein p53 is a cellular hub protein in humans that regulates the function associated with p53 binding to many binding partner proteins, which tend to control cell growth. Importantly, p53 and its cellular partners are found in cancers in nearly every cell type. Rescuing mutated p53 activity or blocking certain signaling interactions with p53 have been reported as an effective paradigm for setting up treatable targets for several cancers [51, 257, 258, 263].

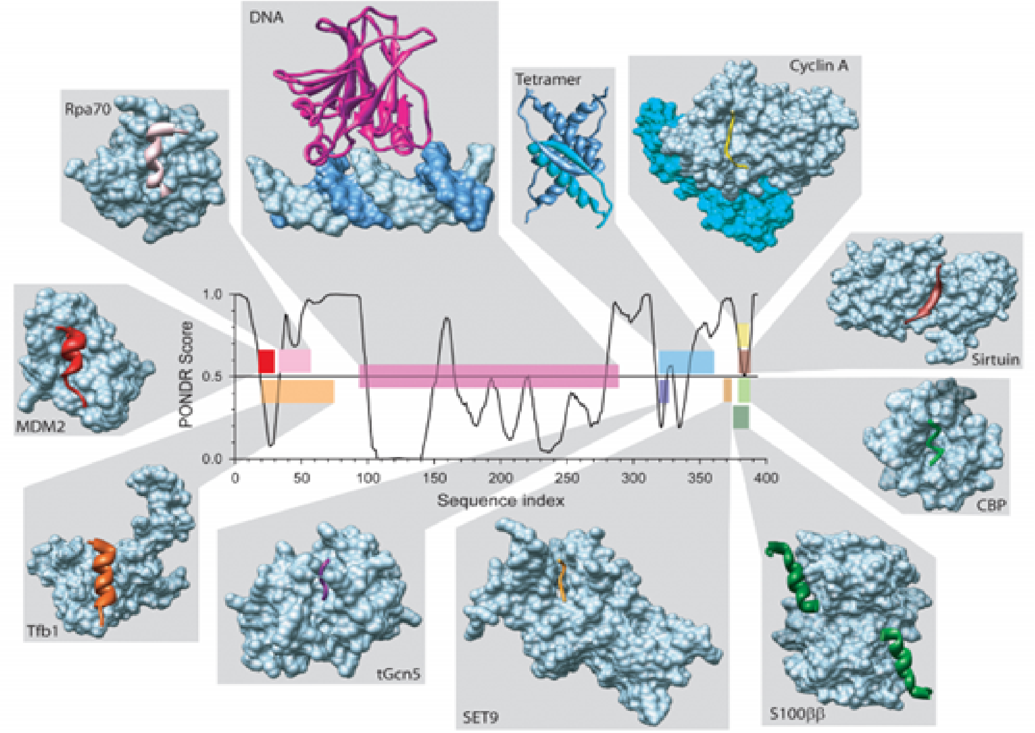


Figure 3.7: Disorder prediction for human p53. Residues with score above 0.5 are predicted to be disordered while those below are structured, which can be MoRFs. Those confirmed to be MoRFs have their associated interactions shown in grey.

The p53 protein has other protein binding partners and a DNA binding site, Figure 3.12, to design peptides for. Interference with other interactions, such as cyclin A, replication protein A or general control non-depressor 5 (tcgn5) at specific sites on p53 can be achieved. The p53 protein has multiple MoRF sites that regulate a variety of cell growth mechanisms through various protein-protein interactions as illustrated in Figure 3.12. There are several publications on inhibition of MDM2 and sirtuin proteins [51, 262, 264]. These proteins participate in essential pathways and make excellent targets for designed peptides to interfere with p53-MDM2 and p53-sirtuin interactions. The p53 IDP is known to bind to peptides, and its counterparts have also been shown to bind to peptides. Often crystal structures of MDM2 are stabilized with snippets of p53 or peptides that offer high tumor cell proliferation suppression [265]. Our previous success with an analogous system and this literature review give us confidence that our proposed computational approach is appropriate to apply to p53 and MDM2.

3.3.2.1 Results of SIRT2 and MDM2 Modeling

Experimental structure 4ZZJ was used as a reference for the binding of p53 to SIRT2. The actual binding site is discontinuous across the structured core of SIRT. As such, the longer component, residues 410-421, were selected as the target for these design runs. The scores from ZDOCK were found to mildly scale with size of interactions. The scores were normalized by the length of the binding site to help compare binding results between the target and alternative docking sites on the target protein, Figure 3.8. An ideal outcome would show the target site, used for the sequence diversification and subsequent peptide creation, would have the highest scores. Observably in SIRT2 a portion of the flexible N-terminal (site 3), has the highest scores for all but the highest ranked candidate peptides where they are similar. While specificity of binding is considered when ranking the predicted peptides, this particular site is actually near the target binding site, determined from the p53 peptide in 4ZZJ. Considering a portion of the real binding site was excluded from the docking measurements in Figure 3.8, it is likely that affinity of these peptides to the binding site are good. There is also reasonable separation between these scores and the other subsections selected as alternate sites from across the structured SIRT2 core. Meaning in three dimensional space these peptide prefer binding to the area intended.

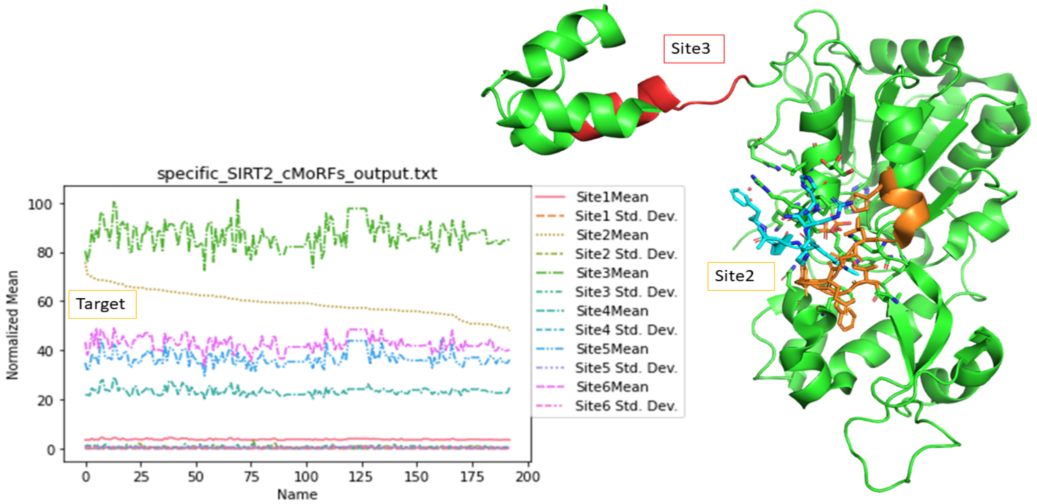


Figure 3.8: Docking score normalized by target length for each site. Data sorted on score for Site 2, the binding site for p53 to SIRT2. The cartoon of 4ZZJ show SITE 2 and SITE 3 along with the p53 peptide (cyan).

Experimental structures 4HFZ was used as a basis for the binding of p53 to MDM2. The MDM2 binding site for p53 is found along a MoRF section in a disordered segment. As such, rigid docking will rely on a correct conformation. This proved difficult for I-TASSER to model. To yield initial structures for the pipeline AlphaFold2 was used to generate full length structures of human MDM2, uniprot P04637.

The computation yielded fewer candidate peptides than previous runs. The binding site selected was residues 14 to 28. It could be this segment of sequence was too well conserved throughout the sequences used for the structure search. As before, scores are normalized by the length of the subsection being targeted during docking. Figure 3.9 shows the target site has higher mean score than most of the other sites tested. The site tested with the highest mean score for all (Site 4) is the first 14 residues of the protein that lead up to the binding site, but not including it. The peptides generated did show some selectivity over the C-terminus (site 3), and reasonable specificity to the structured components of MDM2.

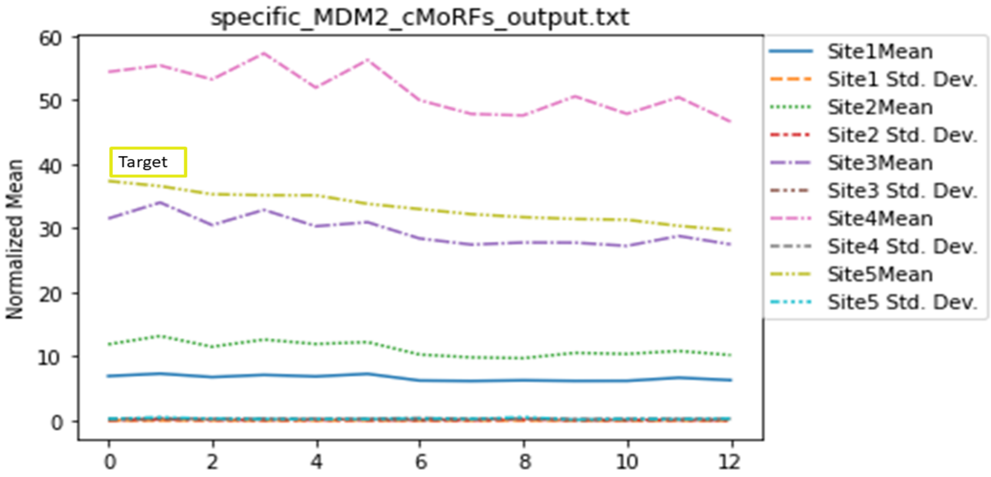


Figure 3.9: Docking score normalized by target length for each site. Data sorted on score for Site 5, the binding site for p53 to MDM2.

The disorder nature of MDM2 implies that rigid docking is likely insufficient for ascertaining bind ability. Recently it has appeared that AlphaFold2 multimer model has reasonable ability to predict protein–peptide conformations which can even predict binding affinity in certain cases [266–268]. Simple direct bound conformation, competition assays between known and unknown binders, as well as foundational benchmarking have been done for this model. All indications suggest this is a reasonable orthogonal method to ascertain possible interactions. Limitations of this method include precise backbone conformations being reproduced, as determined from solid states structures. So this is potentially a bias in the empirical data. Regardless, a wide variety of peptide structure and protein–peptide interactions can be predicted using alphafold in either a single sequence mode, relying on recycling the structure embedding during prediction to arrive at a solution, or the normal MSA mode.

Utilizing this methodology MDM2 was tested against p53 in a competitive binding assay, Figure 3.10. The MDM2 sequence from the 4HFZ structure was used, the p53 sequence from the same structure, and the predicted peptide sequence. This experiment shows a variety of classifications for the resultant bound states, Figure 3.11.

In some cases the p53 peptides was predicted nearly identical to the 4HFZ structure, with the peptide binding poorly or not at all. Successful cases that predicted the peptide to supplant the p53 peptide specifically at the binding site. And unspecific binding where the predicted peptide is clearly binding, but to a random part of the structure. Running the multimer prediction with just the peptide and MDM2 showed a correlation between the competitive predictions and mean distance of the single interacting predictions. This variation of the experiment can only demonstrate binding ability, without the clarification of specific or unspecific binding.

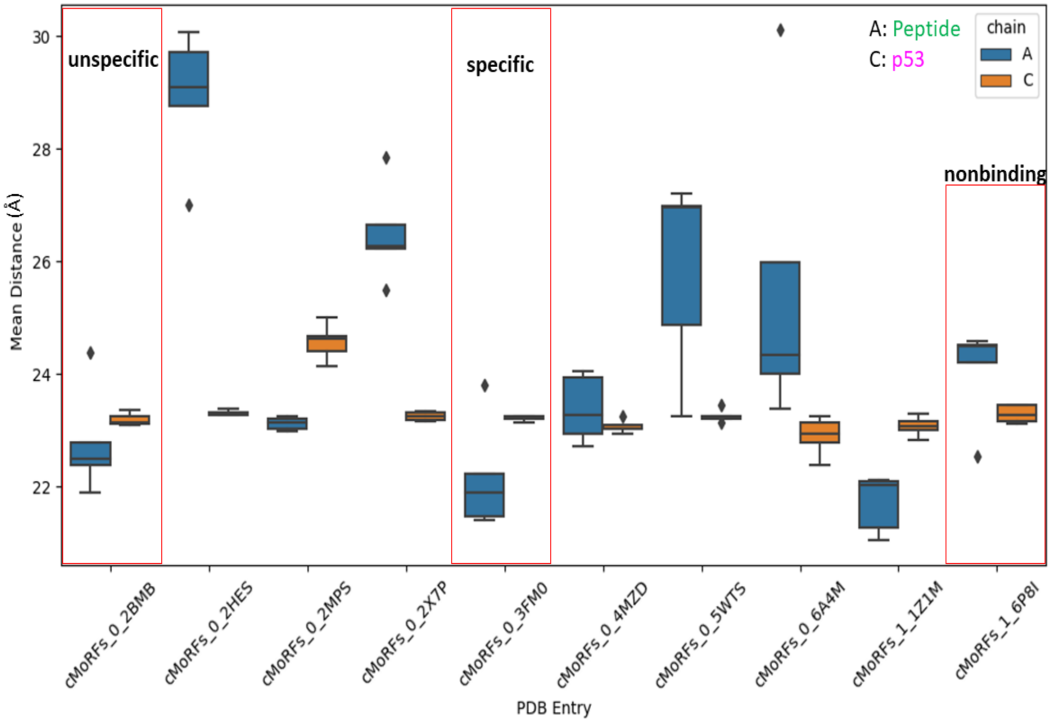


Figure 3.10: Mean Distance, Cα-Cα, between peptide and receptor for MDM2–p53 binding site top peptides.

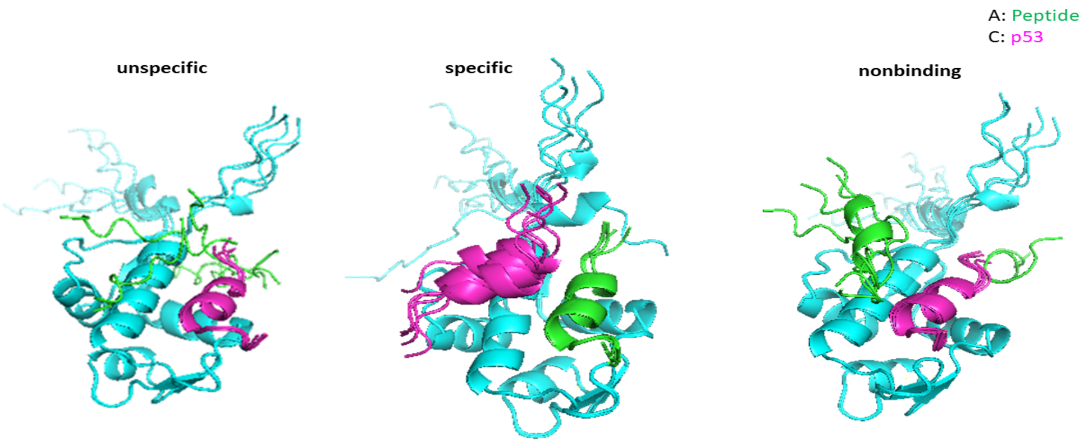


Figure 3.11: Example structures from the AlphaFold multimer prediction. Typucially three classes of binder appear: specific, unspecifc, and nonbinding. In magenta, the native p53 peptide that binds to MDM2, cyan, similar to the nonbinding pose. In green, the predicted peptide from pepStream.

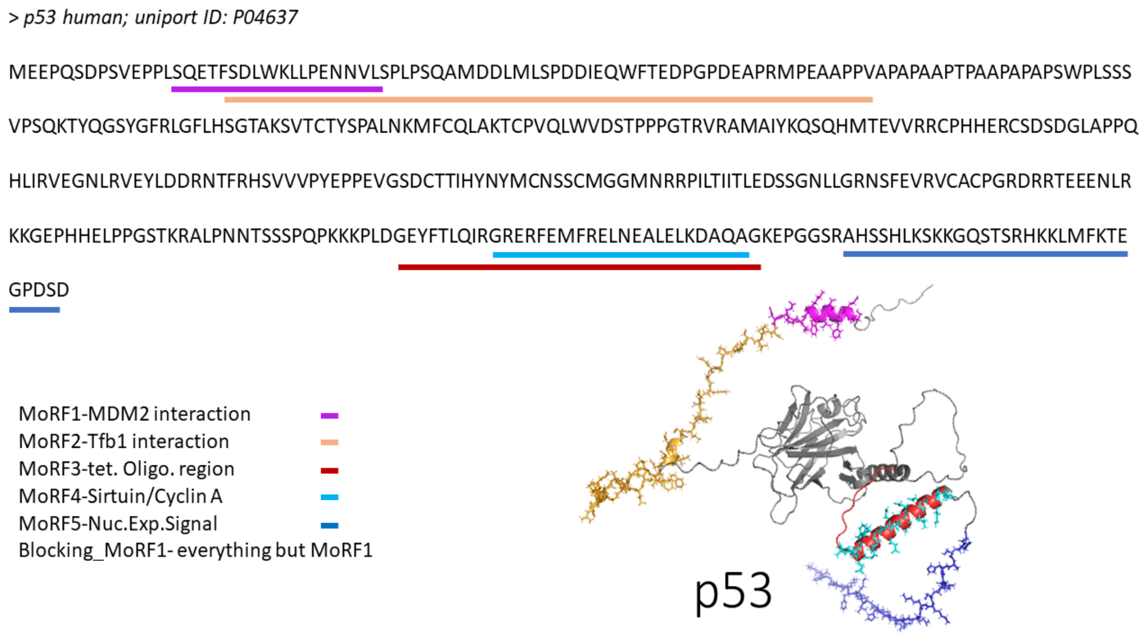
This orthogonal analysis was helpful to add a selection criterion for corresponding binding measurements, where a small number of peptide will be tested initially for practical reasons. Table 3.1 shows an example of the raw data being used to evaluate which candidates. Practical considerations of laboratory resources requires a subselection of candidates to test initially. The N terminal of MDM2 can be limited to the first 100 residues for this binding experiment, as the disordered segment will undergo a rapid conformation change to bound state and will prevent possible false positives. Poorly soluble peptide will likely require co-solvents or extremely low concentrations, so trying to select reasonable peptides for the aqueous experiment is also a reasonable requirement. To refine the top peptide set the overall docking score to the target site from ZDOCK, the predicted gravy scores, and the mean distance predicted between the peptide and the N-termini of MDM2 were used to select peptides to perform MST on.

Table 3.1: Docking Scores and Peptide Characteristics

Name	Sequence	Docking Score Mean	Docking Score Std. Dev.	Gravy (- is soluble)	AF2 Mean Dist. to Target
cMoRFs_1_6P8I	SDGDLTLIYFGGD	597.9841	4.399017	0.030769	**21.896299**
cMoRFs_0_2MPS	GGTTFEHLWSSLEPD	585.2565	8.029428	**-0.64**	23.045256
cMoRFs_0_2BMB	WKLVPLPYRSG	565.1265	4.846744	**-0.29091**	**21.624031**
cMoRFs_0_5WTS	DLFGVPSFSVK	562.6523	3.746744	0.618182	23.014660
cMoRFs_0_4MZD	AGQYLMISGSYDDG	562.4132	3.688369	**-0.27857**	**21.562941**
cMoRFs_0_6A4M	FPQLTDVSFQSQNHTWDTVV	541.1369	3.746615	**-0.42**	22.176880
cMoRFs_1_1Z1M	HIVYCSNDLLG	528.0161	5.063502	0.554545	**20.689459**
cMoRFs_0_2HES	LQEHSQDVKHVIW	515.2622	4.150987	**-0.71538**	28.208982
cMoRFs_0_3FM0	AWAPSGNLLA	507.834	8.470817	0.58	**20.059511**
cMoRFs_0_2X7P	VKSPLLWAVSTGSNRD	503.2997	2.212397	**-0.225**	22.161436
cMoRFs_0_2Z72	KVVALSDVHGQYDVL	501.3251	4.276782	0.406667	21. 73202
cMoRFs_1_3FM0	CCATLEGHESTVWSLAFDPSGQ	486.0306	3.415333	-0.05	21.989141

3.3.2.2 Results of p53 MoRF Peptide Designs

In the p53 MoRF designs, identified experimental targets on the p53 alpha human sequence are depicted in Figure 3.12. Interestingly, some of these targets encompass either parts or entire other targets. Such overlaps were ascertained from a multitude of structural data derived from the p53 MoRFs [51]. One primary aim of this approach was to assess the precision with which the pepStream process could pinpoint specific peptides when comparing extensive segments of the same interaction. The selection of these targets was informed by their significance in cancer signaling and the fact that they are not situated on the structured DNA binding domain of p53. This criterion ensured that the chosen targets were disordered, making them ideal for pepStream. There are several recognized drugs and binding peptides associated with the DNA domain of p53. However, it was decided to exclude this region from our design efforts. The reasons for this exclusion are threefold. Firstly, this region is structured, making it less suitable for our primary focus. Secondly, though our pipeline could feasibly be applied to it, this was not the primary interest. And lastly, the binding interface in this domain is expansive and fragmented in terms of sequence space. This latter point highlighted a prominent limitation in our software, which became evident during the SIRT2 designs.



> p53 human; uniport ID: P04637

MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSS

VPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQ

HLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLR

KKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTE

GPDSD

MoRF1-MDM2 interaction 🟪
MoRF2-Tfb1 interaction 🟧
MoRF3-tet. Oligo. region 🟥
MoRF4-Sirtuin/Cyclin A 🟦
MoRF5-Nuc.Exp.Signal 🟦
Blocking_MoRF1- everything but MoRF1

Figure 3.12: Model of p53 with targeted MoRFs shown in colors corresponding to the colored blocks on the sequence above. Each MoRF site is correlated to a known binding interface, deduced largely from empirical structural data.

The results from running the pepStream pipeline over the p53 MoRFs selected are shown in Figure 3.13. Run-time for each MoRF was elongated to 3 to 4 weeks as the returned candidate peptide was many multiples larger than expected, slowing down the initial pass of the docking simulations. For MoRF 2 this was particularly difficult as the size created an extremely large number of mutant sequences, over 10^6, requiring implementation of cut off for the sequence diversification. The p53 structure, as with the MDM2 structure, was predicted from AlphaFold2. This large structure is heavily disorder, creating a very large surface area. The docking method is a FFT method that will likely scale with surface area, rather than strictly sequence legnth of a protein. There is a significant trend that all peptides were found to bind to the MoRF 1, N terminal MDM2 interaction site (site 2 in figures). In fact there is no re-ordering in any of the p53 MoRF design runs. For MoRF 1 it could be concluded that the pipeline generates a large gap in binding ability to the target designed for. With juxtaposition to other MoRF results it is clear there is either a failure to generate a diversity in sequence of peptides, resulting in consistent results, or some segments of p53 can bind to peptide more effectively than others. In the MoRF 5 results, Figure 3.13, there is a significant increase in the top ranked binding ability compared to the bulk of top results. However, this is mirrored in binding score to MoRF 1.

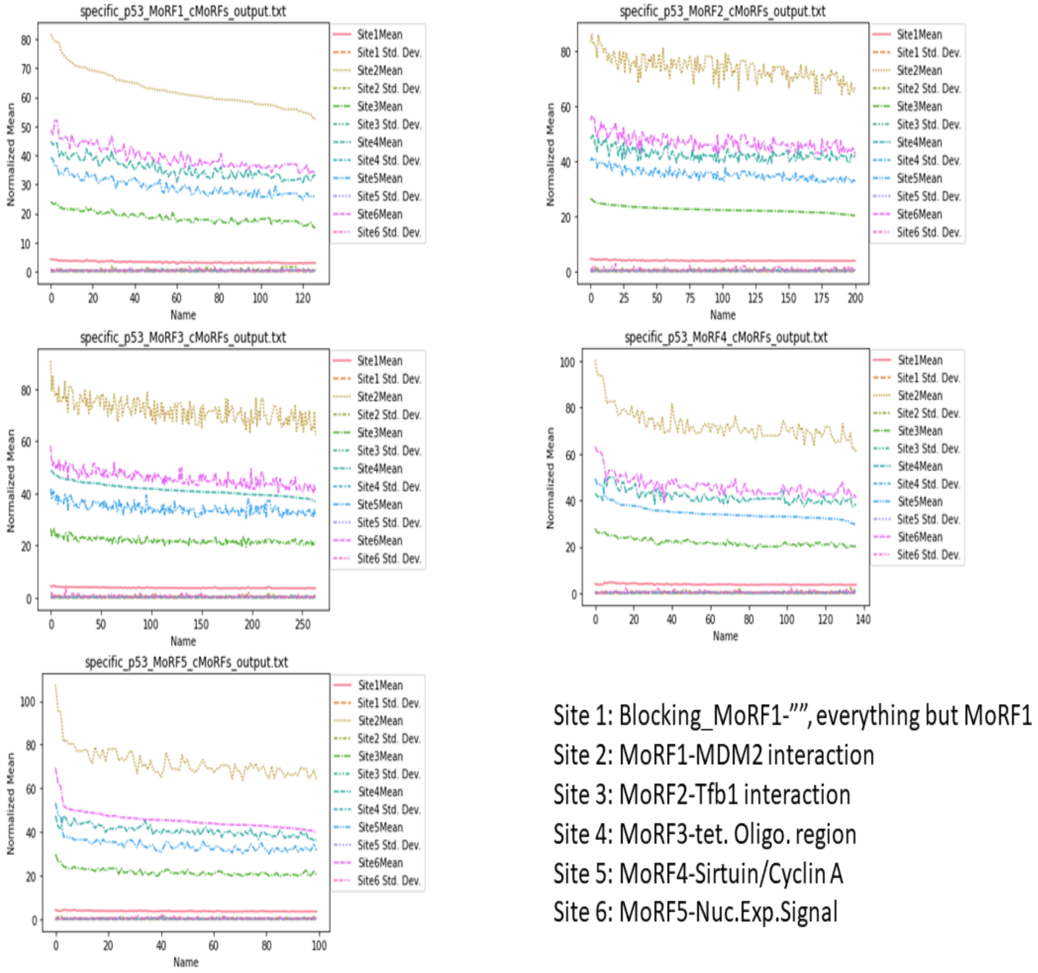


Figure 3.13: Mean of the MoRF length normalized docking scores from ZDOCK for each of the target sites on the disordered segments of p53 plotted against the rank ordered peptides for the given MoRF site of p53, denoted in bottom right corner.

3.3.2.3 Analysis of Candidate Peptides

As done with MDM2, using AlphaFold2, a computational competition simulation was conducted to predict the structures of receptors when paired with two peptides. Without knowing a precise bound conformation that p53 should take, a RMSD to a known structure is not possible. Instead 3.14, a simple mean of the distance between Cα of the two chains are taken for qualitative assessment. Notably, the greatest variation in average structural distance was observed in MORF 3, while sub-optimal binding, having the largest mean distance approximated at 30 Å, was exhibited by MoRF 5. Conversely, the most promising binding predictions were seen in MoRFs 3 and 4, with instances revealing mean distances close to 20 toward the p53 N and C termini. However, challenges were faced due to a marked lack of specificity, discerned from docking measurements, and the pronounced issue of conformation induction in p53 when interacting with binding partners—a complexity that the current iteration of pepStream was found to struggle with.

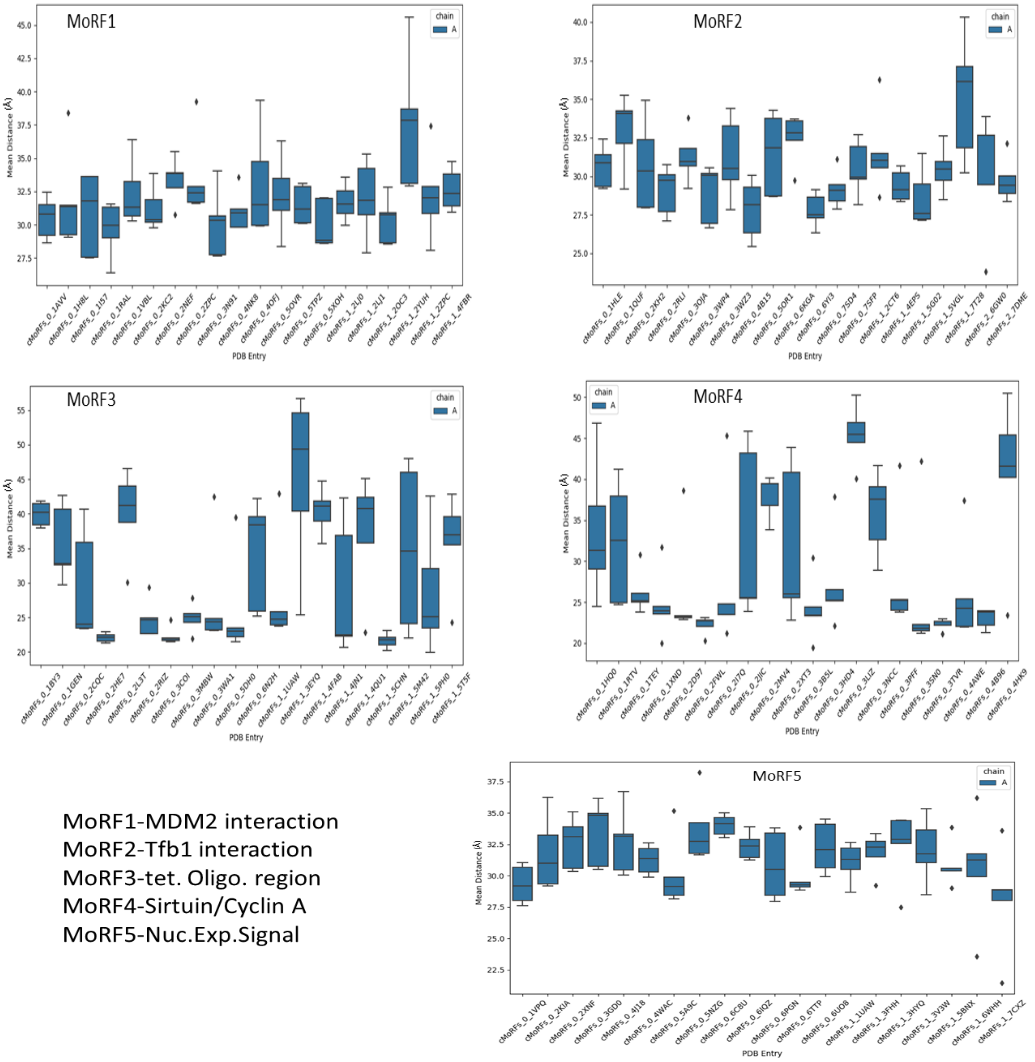


Figure 3.14: Mean Distance, Cα-Cα, between the peptide and the respective termini, first or last 100 residues, of the given MoRF target.

To explore the generic trends of the predicted peptides a simple biophysical predictions of the resultant peptides for each of the targets shown for the p53, MDM2, and SIRT2 systems. This sequence based prediction is plotted in Figure 3.15 was performed using Biopython. Observably there is little clustering of any of the pep-

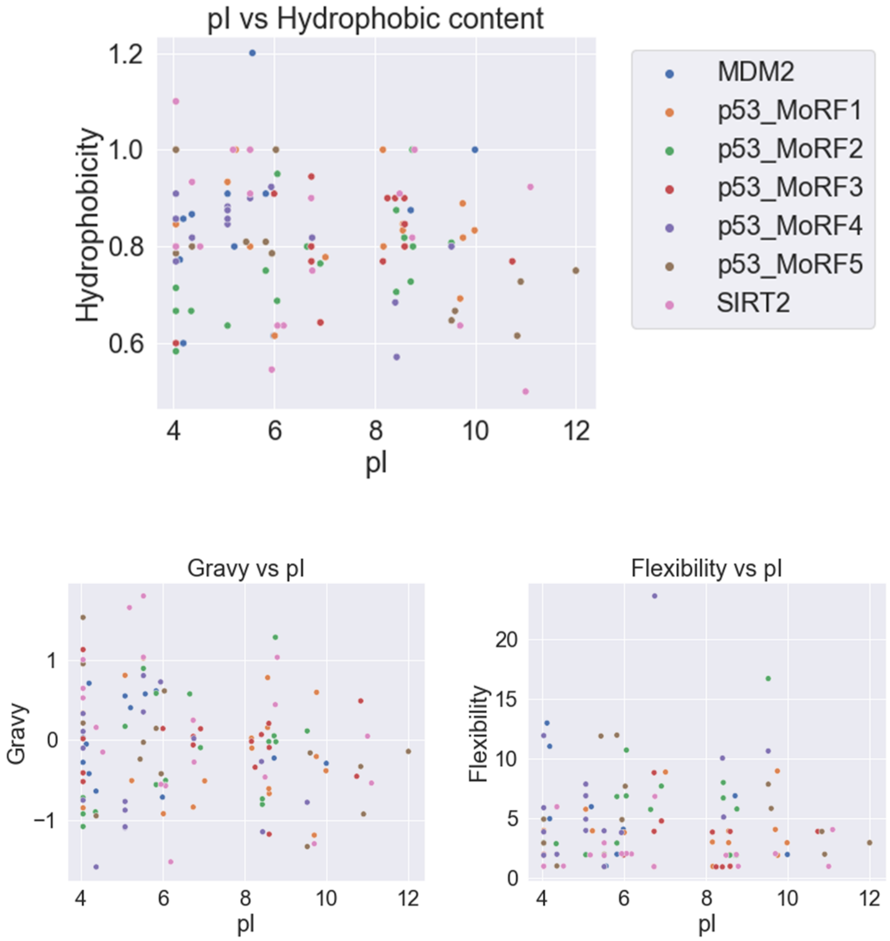


Figure 3.15: Biophsyical properties top 20 sequence of each system tested using pepStream v2. Properties calculated from the sequence alone using Biopython [252].

tides. There is a preference for peptides to hold a charge, as denoted by the gap of pI around 7. A small favouring of flexible peptides for MDM2, MoRF4, and MoRF2 seems to be present as well. Both gravy and hydrophobicity deal with solubility of these peptides, which seems to have been selected randomly in these metrics.

CHAPTER 4: pepStream Advancements and Machine Learning

4.1 Motivation

The advancements in ML in the last four years have reshaped the definition of cutting edge practices in protein simulations, structural prediction, design, and other facets of structural bioinformatics. The early success of pepStream can likely be attributed to two pivotal aspects. Firstly, our cohort had streamlined the problem to a granularity that made it manageable. Secondly, past failures had prompted our cohort to explore alternative design strategies. This afford both the interest and acute level details needed to address the particular problem.

4.1.1 pepStream version 2

The original architecture of pepStream was grounded in dated software. This predicament is not uncommon in computer science projects and is resonant with the notion of "technical debt". Technical debt generally refers to the future costs of rectifying the limitations of choosing a simpler solution over a comprehensive one that may initially be more time-consuming. In this case a short turn around time framed by funding also prevented using untested methods, thus slightly older software comprising the original pepStream. The unavailability of the original raw code for recompilation further exacerbates the challenges associated with this debt.

Recognizing these challenges and buoyed by the synergy between hardware advancements and software evolution, especially those leveraging GPUs or ML, the pepStream pipeline was upgraded. This redesign was driven primarily due to technological advancements that either improve traditional approaches, like rigid docking, or obsolete them with new methodology. There was also a desire to improve the pipeline in technical details like file handling, attempt to reconcile the current technical debt for new debt, and explore new methodology in the bioinformatics software ecosystem. Particularly, structure prediction being dominated by machine learning tools was of paramount interest to incorporate in some efficient manner.

4.1.2 pepStreaML: pepStream Refashioned Using a LLM

The emergence of AlphaFold inspired machine learning applications that could go into pepStream, as seen in pepStream version 2. Recently, multiple language based models for protein sequences appeared, implemented for a variety of different purposes, but usually the sequence input relationship being learned to some output. In the previous verion, the PSP usage of sequence to structure was limited to a one time prediction which then would be utilized for coarse grain docking simulations. There was no feedback involved with the peptide and the structure prediction; the classic conformation problem.

A structurally responsive pipeline could be implemented. The enhanced functionalities of pepStream involve a systematic exploration of sequence homology to identify conserved motifs within the target sequence. These motifs are subsequently subjected to diversification, enabling the localization of analogous sequences within different protein contexts. Using a machine learning model to facilitate the expansion of sequences and the subsequent generation of peptides, as well as the prediction of potential peptide-protein structural interactions affords a comprehensive pipeline consolidated within a singular software package. This enables the seamless execution of all stages, effectively streamlining the entire process for peptide design.

4.1.3 Retrospection of pepStream

The original version of the pipeline suffered from both poor encapsulation of the core runtime code and in the ability to predict structures quickly. Unification of the entire code base and turning the process into a single executable was a large improvement to the pepStream pipeline, warranting a clear distinction in version. User interface to the pipeline was reduced to a single run script that uses a parameter text file to run. The usage of a input file offers reproducibility, ease of explanation for new users, and easier exploration of important variables in the pipeline. Version 2 also benefits from building a conda environment, a language agnostic management tool that manages dependencies needed to run software. Transferring the software was also a major issue in the first iteration of the platform, largely due to complexities of installing third party software which rarely has software support. Building in easily deployed environment allows a single file produced from the developed conda environment to duplicate the same packages and software on any computer. Removal of knowledge based structure prediction I-TASSER and replacement with a machine learning structure prediction model was a heavily desired upgrade shortly after the public release of AlphaFold2. This lead to the development of an entirely ML based version of the pipeline as well.

4.2 Methods

4.2.1 pepStream version 2 Methods

The protein structure prediction pipeline AlphaFold was experimented with as a replacement, however the run times were longer than desired. An alternative Omegafold uses a large language model based off the PaLM language model [269] trained on protein sequences in tandem with a geometrical transformer to produce a protein structure prediction. Omegafold offers extremely fast structure prediction, with reasonable accuracy, even for large structural predictions above 1000 residues. One issue with Omegafold is there is only 1 output structure, without rerunning and changing the model parameters for the artificial MSA.

To compensate a framework rigidity optimized dynamic algorithm (FRODA) process is applied to generate lots of conforms, then refine down to a manageable number [194, 270]. FRODA is a constraint-based geometric simulation technique that speeds up the search for native like topologies by accounting only for geometric relationships between atoms instead of detailed energetics like in MD. This creates a large number of conformations. These are reduce to a sparse set of representative points, preserving as much of the original structure or variance as possible. This is achieved by an interesting application of dimensional reduction and simple radial basis functions and math.
The algorithm is as follows:

For each point i not previously categorized, the distances between that point and all other points are computed.

$$d_{ij} = |x_i - x_j| \tag{4.1}$$

These distances are then squared and inversed (u) to create a measure of how close or far other points are from point i.

$$d_{ij}^2 = d_{ij} \times d_{ij} \tag{4.2}$$

$$u_{ij} = \frac{1}{d_{ij}^2} \tag{4.3}$$

A summation of the largest spread of these inverted squared distances is stored in *a(i)*. It can be inferred that larger values of *a(i)* imply that point i is central to many other points.

$$a_i = \sum_{j \in \text{ top } m \text{ closest to } i} u_{ij} \tag{4.4}$$

$$\{x_i : a_i > T\} \tag{4.5}$$

The four closest points to point i are temporarily marked to prevent their immediate removal. Points are sorted based on their importance values (a) and least important points are marked for removal. After each iteration, temporary markers are reset allowing reconsideration of any point. These expanded conformations, typically a set of 10, are further refined by performing an energy minimization using GROMACS to ensure the generated peptide conformations represent physically realizable structures. The minimization occurs in explicit TIP3P solvent, with Na and Cl ions added to neutralize the net charge inside the simulation box. The minimization uses steepest descents for 20000 steps or until the maximum force on the protein is less than 500 kJ/mol. Lastly, MEGADOCK is a implementation of ZDOCK that utilizes GPU processing to speed up the FFT-grid-based docking with MPI [133]. Beyond dropping some calculation for their energy score the software and outputs are identical, but can perform thousands of docking calculations in a few minutes.

In Figure 4.1 the outline of the updated pepStream version 2 pipeline is shown.

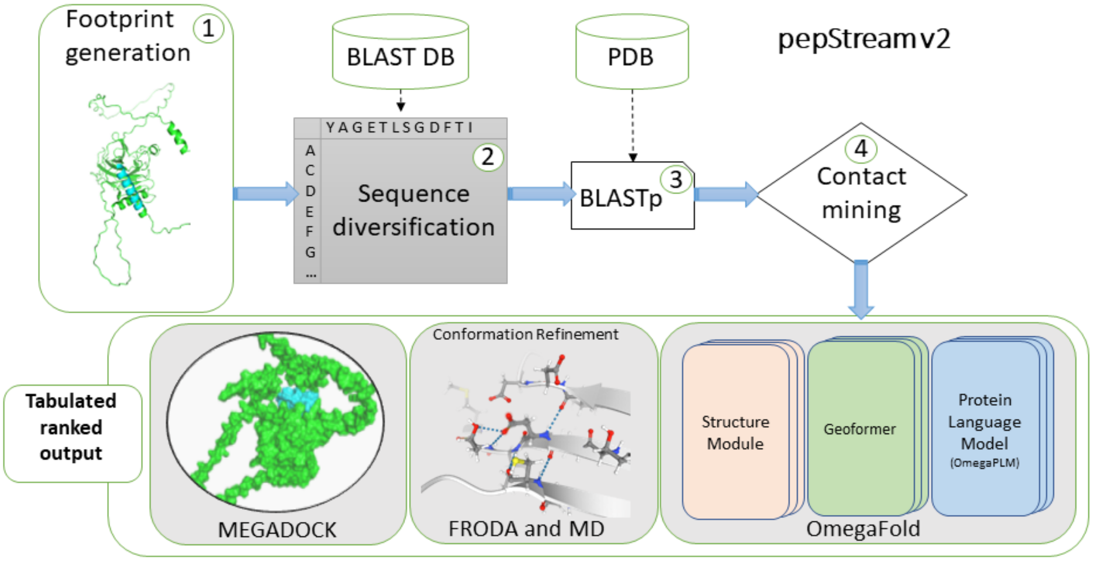


Figure 4.1: Diagram of the improved pepStream, version 2. The entire pipeline is run from a single script, and no prior data beyond sequence of target protein is needed to run. Typically runs fit on a single GPU of 10+ Gb VRAM.

4.2.2 pepStreaML Methods

As with previous versions of pepStream, a user defined target on a protein has a peptide optimized against it. In pepStreaML this is done utilizing features generated from structures predicted and sequences generated in an optimization loop. It was found a single python package, fair-esm, contained most of the needed prediction or generative models needed to performed a guided optimization of a peptide [271]. The python package Biotite was used to measure various aspects of predicted structures to generate quantitative data to optimize peptides for binding properties to the user defined target [272].

4.2.2.1 Evolutionary Scale Model

Encapsulating biological properties from sequence data is a logical step toward generative and predictive ML for biology. Using 250 million sequences from the UniParc database, a deep contextual language model with unsupervised learning to sequences spanning evolutionary diversity can be performed. The Evolutionary Scale Model (ESM) finds without prior knowledge, information emerges in the learned representations on fundamental properties of proteins such as secondary structure, contacts, and biological activity [271]. Embedded representations are useful across benchmarks for remote homology detection, prediction of secondary structure, long-range residue–residue contacts, and mutational effect. Protiogenic protein sequences use a small vocabulary of 20 canonical elements, the modeling problem is more similar to character-level language models than word-level models. Like natural language, protein sequences also contain long-range dependencies, motivating use of architectures that detect and model distant context, like BeRT [165].

The BERT model uses unsupervised learning to train deep neural networks to model contextual language, originally for language applications. When trained on protein sequences the learned representations capture multiscale patterns, reflecting various levels of biochemical properties and remote homology across proteins. The representations also encode information about secondary and tertiary protein structures, which can be discerned through linear projections. Self-attention facilitates representation learning on the sequences by incorporating context from across the input sequence, making the architecture adept at representing residue–residue interactions as well as it represents word-word relationships in NLP applications.

Algorithm 1 Pseudocode of the pepStream ESM based workflow. Classes for the design mode, target or whole, and the inverse folding process are called and operate over input handled from the core run code.

```
procedure MAIN
    Initialize Parameters and Parse Command Line Options
    if not binding_site provided then
        exit program
    end if                                          ▷ Input structure handeling for uniprot or PDB IDs
    Prepare Output Directory
    Save user options as JSON                       ▷ Section 1: ESM-IF diversify structure
    Initialize ESM Inverse Folding
    if Inverse folding results exist then
        Use existing results
    else
        Perform inverse folding
        Calculate relative threshold for sequence identity
        Return N sequences per temperature
    end if
    Plot Results of ESM-IF
    Get Conserved Positions                         ▷ Section 2: Run time design using ESMFold
    for i in range(number of designs requested) do
        Initial Sequence Generation: using only target sub–sequence to predict peptide
        Convert the best found peptide to new design object
        Optimize the sequence using Simulated Annealing
    end for                                         ▷ Section 3: Outputs csv and pdb files
end procedure
END
```

The pepStream ESM based pseudocode is shown above, referred to as pepStreaML. Replacement of the diversification steps using the BLAST sequence database with a inverse folding process from ESM-IF offers a faster and more portable diversification method. This process was optimized for conservation of sequence by simply adding a threshold for the sequence identity produced from the structure to sequence prediction. This eliminates the need for any sequence searching performed via a database. A major limitation was the relative cut off, some structures cannot generate 50% sequence recovery using this method. As such a relative sampling routine was done on a batch of sequence generation before anything is saved. These are measured for the mean sequence identity produced to guarantee that the filter will not prevent the generation of the minimum number of new sequences, passed in by the user.

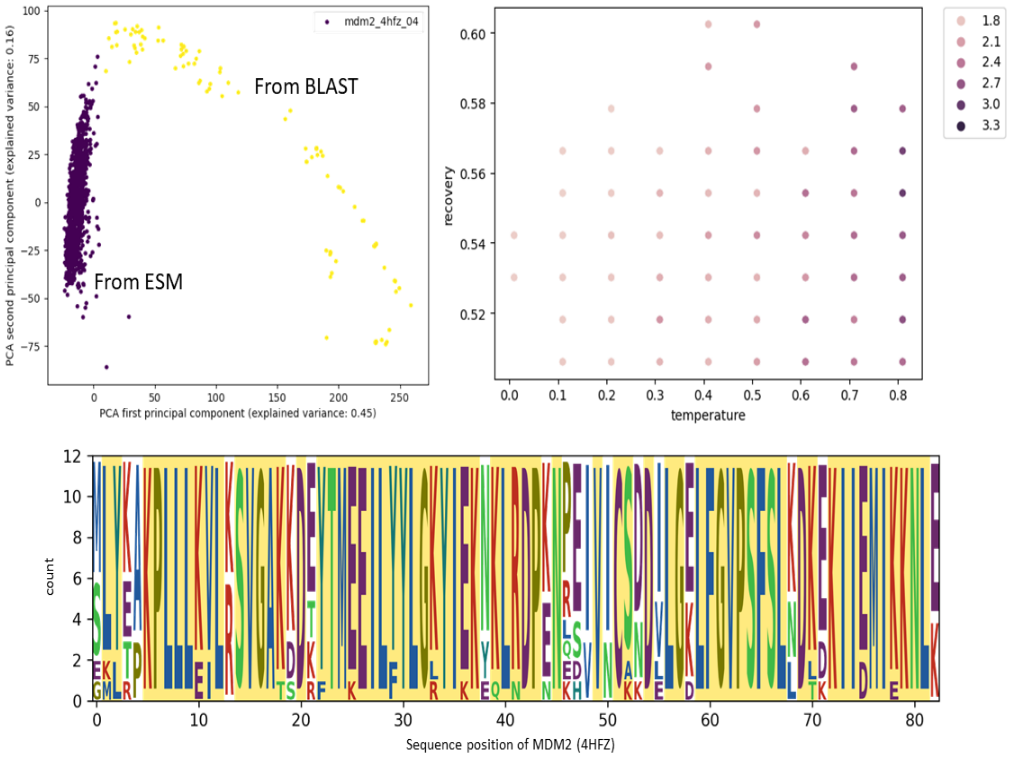


Figure 4.2: ESM Inverse Folding on MDM2 (PDBID 4HFZ) embedded with real MDM2 sequence produced from BLAST alignment.

From this, conserved residues likely important for binding events are known and alternative sequences for the target sequence could be derived easily. To generate a peptide simply applying the core ESM language model would fill in a set of masks attached onto the input sequence. It was obvious that this is a deterministic prediction unless gradient and drop outs are enabled for the model. These conditions would still converge rapidly. Sequence space was found to optimize poorly for structure space. Obviously an optimizer was needed to relate the goodness of the sequence to a relevant metric, i.e. binding of the peptide to the protein. Work affiliate with the original ESM research published an open source protein programming language that included a metropolis Hastings based Monte Carlo Markov Chain (MCMC) optimization that could use non-differentiable measurements during optimization time [237]. So instead of a gradient descent on properties that might be directed during prediction time, the optimizer will sample in a semi random fashion over the allowed mutations of a given encapsulated protein sequence. The properties are referred to as energy functions and include things like globularity, hydrophobic content, RMSD to reference, distances, among other structure based measurements. As optimization occurs it would be detrimental to continuously allow large portions of a sequence to change, as such a computational annealing method is applied to the sequence prediction as a temperature. This scales down during the optimization based off a defined rate, affecting both the number of residues mutated and the mutation prediction by ESM.

At this point a peptide could be optimized for a given protein interaction, as the ESMFold process can handle multi-chained complexes, being based off Openfold. This offered a simple path for measuring goodness of interaction between peptide and target. Approach design with the whole protein and hope target arrives, or only model sub domain with a reference constraint and allow peptide to design. It could also be the case that a structure should be deformed upon binding, so remove target structure restrictions is easily done. Getting the best initial peptide start became an obvious limitation for optimization during early trials. It was observed that poor initial sequences essentially have no optimization plane. During annealing the rapid mutation narrowing secludes the bad sequences into a useless optimization routine. To capture a reasonable optimization run, the initial peptide is calculated by running the sub–target only. After sampling this sequence space a best initial structural start is selected and the peptide sequence is swapped into a full model, if run in whole protein mode. This affords two design strategies: quick and local, target mode, or slower and global, whole protein mode.

4.3 Results

4.3.1 Results of pepStream version 2

During testing of the pipeline, reliable execution and base functionality were consistently achieved. It quickly became evident that the docking scores exhibited a strong correlation between length and score. To address this, raw docking scores are normalized by the length of both the target and the peptide being scored. Four protein systems have been initial tested as exploratory results for the new pipeline. Both MDM2 and the MoRFs for MDM2 and SIRT2 on p53 are extensions of the core work with pepStream version 1. The addition of a WW domain protein is to begin benchmarking the method as a peptide binding motif is known.

The NMR structure 1JMQ was used for a target sequence. Running the method on default setting required one GPU and approximately 24 hours of compute, a large speedup from the previous version. It was found that 20 percent of the inital fragments from distant homologs yielded the motif known to bind, PPXY or PPX in some cases [273]. The MDM2 sequence is selected from the 4HFZ structure again, restricting the target receptor to just the 100 residues from the crystal. For p53 MoRF 1 and 4 were selected for their importance to MDM2 and SIRT2 partners, should any of the outputs be of interest to test experimentally. These sequences are selected as before, and the entire human p53 sequence is used, as previous run from pepStream v1. The alternative binding sites selected for this trial are the remaining p53 MoRFs from previous computations described for this system. This likely will cause some issue with specificity as portions of the sequences are shared.

Using cross-site binding specificity scores, it is possible to measure the binding affinity to specific sites through restricted site docking, as done in ZDOCK. Docking scores typically have scores in the 1000's and a standard deviation in the single digits. This is because ten conformations of each partner are input into each docking procedure, leading to 100 docking pairs for every peptide and running 1,000 interaction poses for each pair. While specificity is achievable, certain targets, like the MDM2 run, can be especially challenging. To streamline the analysis, it's advisable that alternate docking sites don't overlap any residues with the primary target. Unlike what is done here for p53. Despite this, numerous instances have yielded successful outcomes based on specificity. This suggests that this version of the pipeline is capable of achieving specificity. However, it is worth noting that no empirical tests have been conducted on any of the results so far.

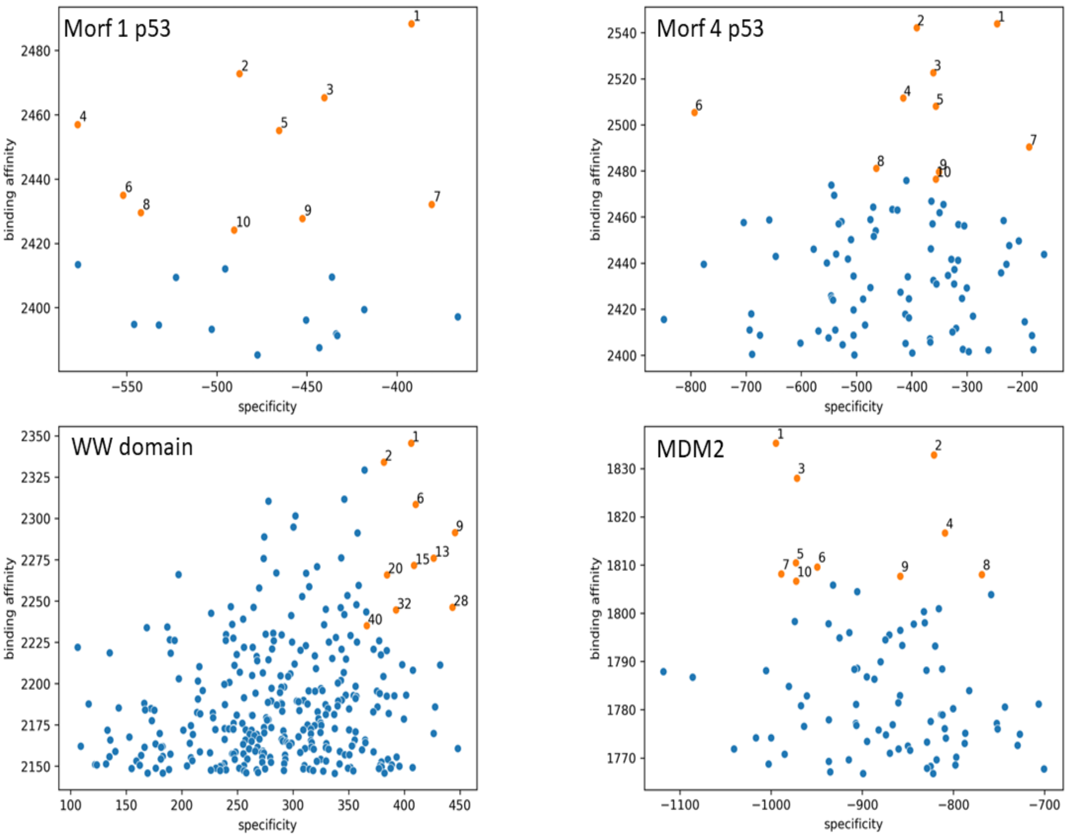


Figure 4.3: Affinity plotted against specificity of predicted peptides for various systems. The Affinity score is directly calculated from the docking scores. The specificity scores is calculated as a quotient of the target binding to alternate site binding, normalized by lengths of interaction protein segments.

Figure 4.3 shows the results of four different systems. It is clear that high scoring binders, determined from the docking scores, do not necessarily imply the specificity. In the previous iteration of pepStream the only cross comparison for specificity was by inspecting the docking scores for the various alternate sites on the protein of interest, as seen in Figure 3.13.

Binding affinity is described above, and specificity is calculated as the difference between the max alternate docking score and the target score. In Figure 3.13 only the WW domain run generated specifically binding peptides, as determined from these two metrics. For MoRF 1 of p53 the specificity calculated was negative, but this was determined to come from the MoRF 2 scores exclusively. Meaning MoRF 1 and MoRF 2 had the highest docking scores, showing that the peptide bound to the termini of interest with some specificity. In reality specificity is likely dominated by kinetic factors, meaning generated peptides can still be practically effective if they hold the correct biophysical attributed needed to interact correctly.

The previous version of pepStream produced low biophysical diversity in the peptides predicted from the method. Predicted peptide sequences produced from this version seem to show biophysical attributes preferences by the targets, Figure 4.4. For WW domain there is a strong preference for longer flexible peptides, which is interesting as the primary binding sit is a narrow gap between two alpha helices. The lowest scorer mdm2 seems to have a mild preference to low pI, and like the rest of the peptides, more rigid than flexible. The p53 MoRFs 1 and 4 had the higher target binding scores than the other two systems. Peptides with a low pI were preferred for for the MoRF 4, showing the highly charged nature the p53 segment. For MoRF 1 there is some abundance of neutral charged peptides, with pI around physiological pH, or perhaps with the spread across pI there is a different factor than just charge that selects for these peptides.

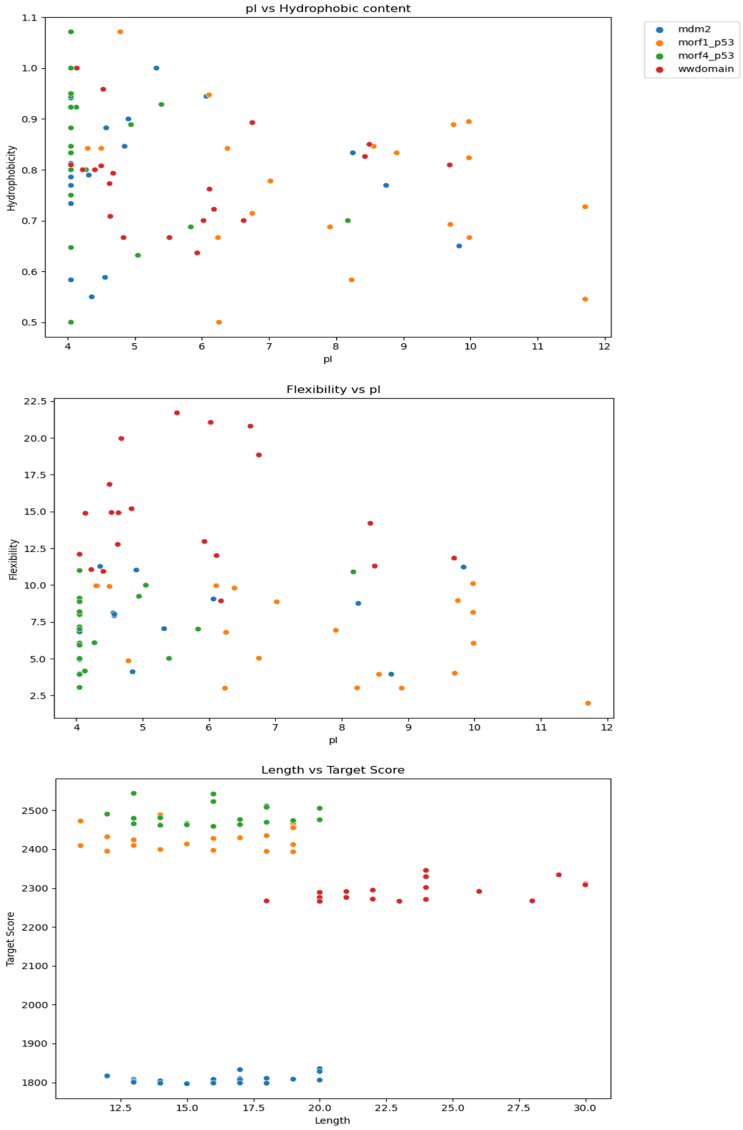


Figure 4.4: Biophsyical properties top 20 sequence of each system tested using pepStream v2. Properties calculated from the sequence alone using Biopython [252].

As before, an orthogonal computational analysis was desired for the candidates produced from version 2 of pepStream. Structural prediction of the complexes were done using AlphaFold2 from the ColabFold package [274].A single example from the WW domain peptide predictions shows a the top ranked pepStream peptide candidate modeled against the 1JMQ sequence, Figure 4.5. This exemplifies the specific binding capability from the predicted peptide for WW domain, seemingly a well behaved system for this methodology. Comparing multiple conformation show a similar pattern of interaction to the WW domain between the native and designed peptides.

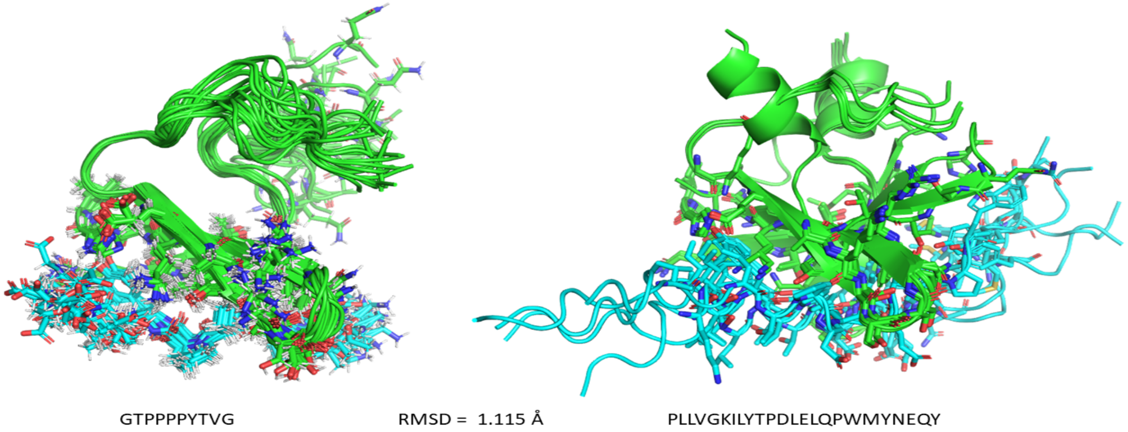


Figure 4.5: Top ranked peptide from pepStream v2 was modelled against WW domain sequence from PDBID 1JMQ. Juxtaposed are the 20 states of 1JMQ, left, aligned and the 5 states predicted from AlphaFold2 for predicted peptide, right. Each system as the peptide sequence printed below. Superposition calculated from first frame of each ensemble using Pymol.

The results derived from applying the AlphaFold2 multimer prediction to the top set of peptides across all systems are presented in Figure 4.6. The disparities between these predictions and those from the first version are noteworthy. The current predictions typically manifest shorter distances. Additionally, certain predictions display extremely limited variance, reflecting a high degree of confidence by AlphaFold2 in the predicted conformations. Furthermore, these results offer visual affirmation of the correct binding sites being identified. As a specific example, considering the WW domain, the native mean distance of 1JMQ is 17.19 Å. The predicted peptides offer numerous instances that are within 1 Å of this mean distance, including an instance with a measured distance of 17.45 Å.

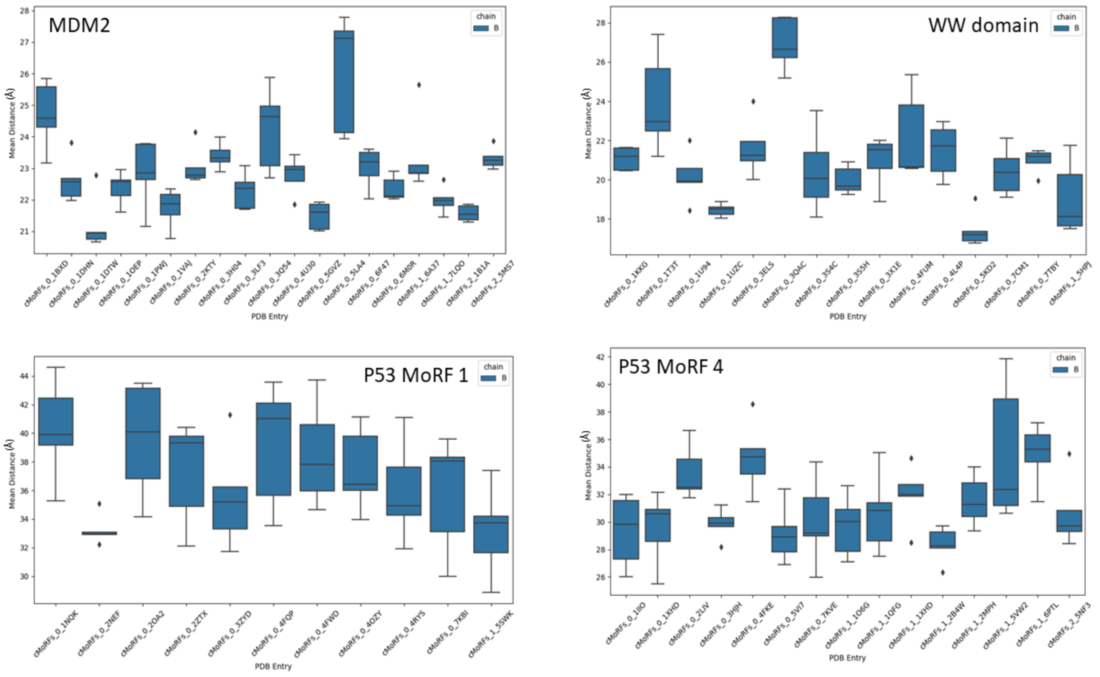


Figure 4.6: Top ranked peptide from pepStream v2 predicted in complex with respective target proteins, or termini from the protein in p53 targets.

4.3.2 Initial Set of Results for pepStreaML

The running speed for the design for sub 100 residues in total, sum of sequence length in target and peptide, on consumer hardware runs in less than a few minutes per design at hundreds of optimization steps. This affords rapid alpha testing and evaluation of possible protein–peptide prediction problems that the design method will arrive at known solutions. A variety of control cases were selected, where not just a structure of a protein peptide was known, but some general information about the peptide motifs that successfully bind. All design runs were performed against 1 structure from the PDB, unless otherwise stated. For each protein the target was determined from residues within 4 Angstroms of the native peptide interaction, and ten peptides designed. For the larger structures sub-targets were used to initially design peptides, then the whole protein was used during the design stage.

4.3.2.1 Method Validation Using Model Systems

Premise of each test is to get an idea on systematic runs on known peptide-protein systems. For each 10 sequences are predicted using the full 3B sequence model then refined for 300 steps using the default optimization parameters, described above. In all cases the two step process of local sequence related peptide prediction followed by an entire protein peptide structural optimization process. Success of a design is determined by conformation induction, motif presence in designed sequence, or binding mode specifics derived from empirical examples. In Figure 4.7 some different designs for known protein–peptide systems are shown.

The WW domain protein used for these designs was from PDBID 1JMQ YAP65 in complex with GTPPPPYTVG peptide. Typically WW domain proteins bind to peptides with polyproline sequences, often motifed as PPXY though some outliers can occur [273]. In this design the entire structure was given to the target mode so conserved residues would be used for score, to improve optimization. Resulting designs all were bound roughly to the known binding cite, within 2 Å of the 1JMQ peptide. Multiple sequences were found with polyproleins of either PPX or PPXY, 5 of 10 runs.

PDZ domain protein used in this design was PDBID 1BE9, bound to partially resolved peptide KQTSV. The binding mode involves a ionic interaction between the conserved carboxylates on the loop of the PDZ domain and the anchor site for aliphatic residues, like leucine or valine. The known [RK]-XXX-G-θ-G-θ motif, where X is any amino acid residue and θ is hydrophobic residues is typical for PDZ binding peptides. As always some variation in literature exists for exact make up or pattern [275]. In this computation some number of designs were actually designed off target, a common issue as there is no way to force the structural prediction model to place the peptide. Of the 10 designs 4 were offsite and 6 contained a reasonable binding motif.

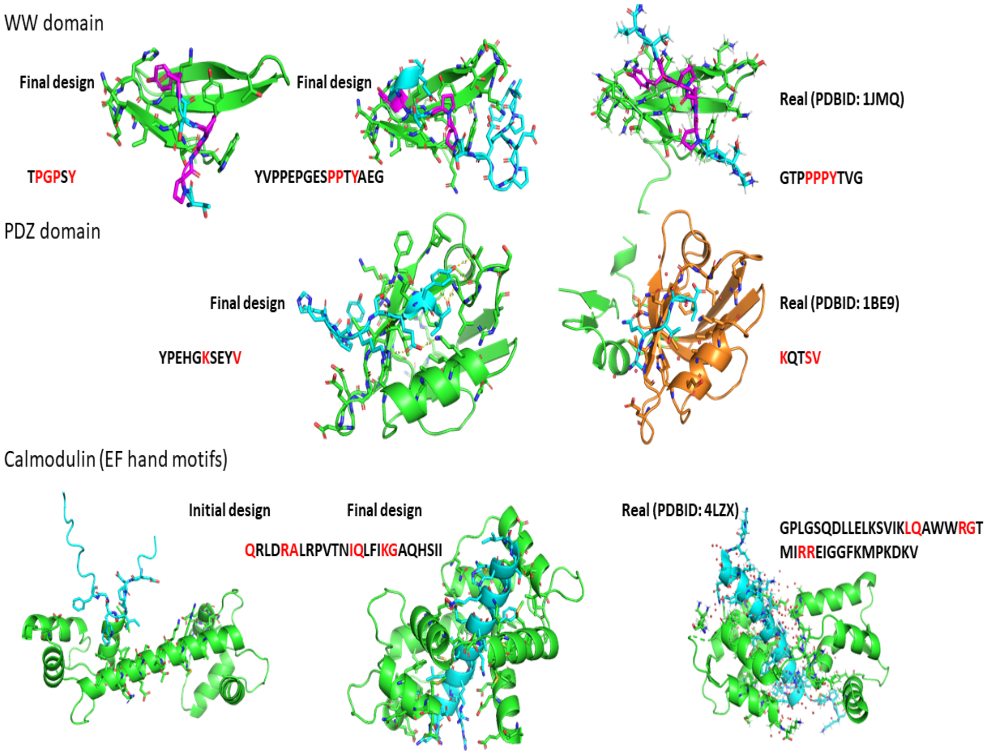


Figure 4.7: Design peptides and the known receptor peptide interactions. Each design was run for 300 steps of design using default weights on the optimization scores. Top of figure, exact matches to WW Domain peptide were generated showing exact match and near match that mimics native binding pose. Middle of figure, the PDZ domain binds to positive charged sequences that are followed by hydrophobic segments. In this case a near match is found for sequence, but the binding interaction that uses valine as an anchor is duplicated. Bottom of figure, calmodulin is a double EF hand motif protein that has significant conformation change upon binding to native peptides. Both correct peptide motifs and induction of bound conformation are generated from the design process.

Calmodulin is a primarily a calcium binding protein featuring dual EF hand motifs, typically involved in Ca2+ regulation of a wide range of physiological processes. This domain can also bind to peptides with or without its ion ligands. When this particular protein binds to a peptide, a very distinct conformation change occurs [276]. Of the 10 designs generated for this receptor 9 generated the bound conformation in calmodulin used from the PDBID 4LZX structure. It was noticed that the energy functions had steep drops during training time, suggesting the correct peptide sequence induced the conformation change. A factor that aids in this guidance is utilizing weighting on the RMSD of the receptor protein. For these designs this conformation of reference was in a bound position, meaning a sudden conformation change from open to closed, relative to bound to peptide, is positively scored.

Beta Lactamase and major histocompatibility complex (MHC) class I were difficult cases, partially. An important observation is that conformation or sequence really matters when utilizing this design process. The inability to capture relevant interactions, arrive at a known conformation, or overly compete ruins the design process. As such multiple conformations and sequences for target proteins might be a wise idea to incorporate in future developments. The method scheme was kept the same as previous test systems, favouring the reference conformational state.

MHC class I molecules present peptides derived from cytosolic proteins, the pathway of MHC class I presentation is often called cytosolic or endogenous pathway. This affords a passive immune system antigen presentation of proteins in the environment. The proteins have evolved to gregariously bind to peptide segments. Multiple studies have tried to evaluate a peptide motif, seemingly structure activity relationships only yield that certain segments of the peptide MHC interface must be anchored with peptide residues, at the boundaries of the helix interface. These proteins have two protein interfaces in practice. Typically they are crystallized bound to a microglobulin, a relatively small globular protein, Figure 4.8. The peptide interface is nested between two helices along the anterior surface of the MHC monomer, relative to the interface with microglobulin. It seems this has created a off target design interface when using the ESM model to generate peptides. Observably in the first set of 10 designs, using structure 2XPG, none were positioned at the normal peptide–protein interface. Instead all were designed at the interface with the microglobulin. The MHC I family has an abundance of structures, using 5AD0 instead of 2XPG designs were found to be at the correct peptide site, shown in Figure 4.8. Only two of 10 peptides were found off peptide interaction site. Overall, disordered extended peptides might be the hardest thing to design, which is sensible given the conformational variance.

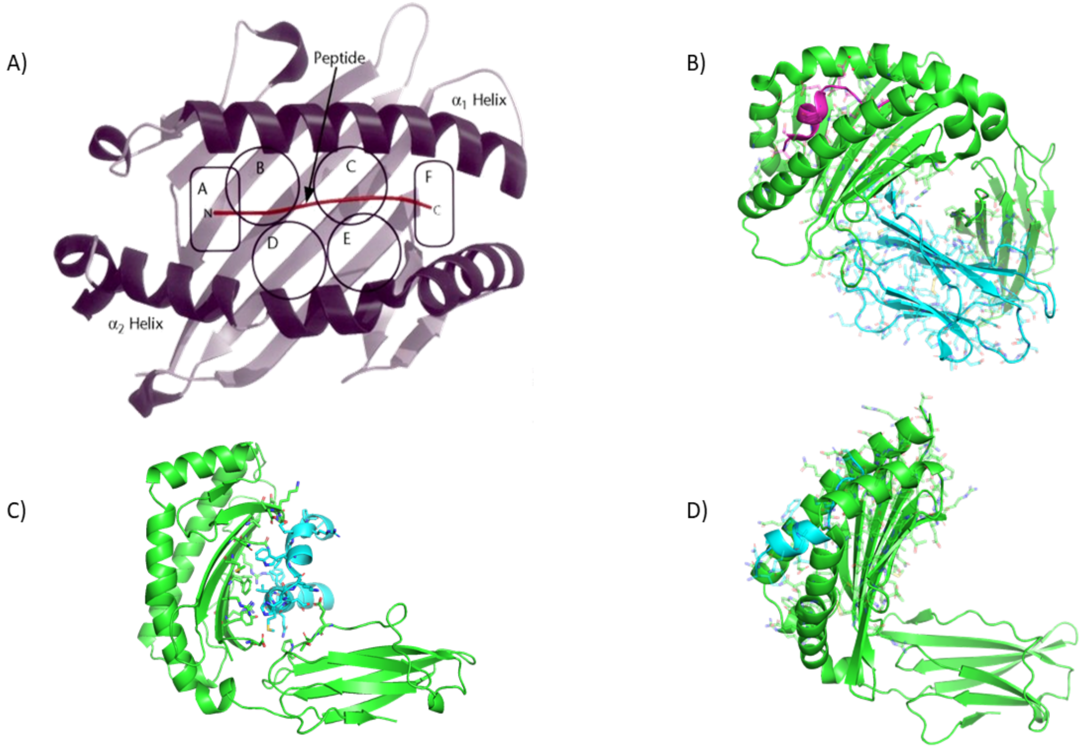


Figure 4.8: Designs for MHC Class I receptors. A) Breakdown of peptide interface on the family, zones A and F are generally considered the most important binding sites. B) Cartoon of 5DA0 chicken MHC class I (Green) bound to a beta-2-microglobulin (cyan) and 11mer peptide (magenta). C&D) Designs from pepStream for 4REX and 5DA0, respectively.

Beta Lactamase is a famously model protein system, the class A enzymes are probably the best-studied β-lactamase family with the description of TEM-1 dating back to the 1960s [277]. While there are no known structures of peptide bound to one of these proteins, the globular protein can serve as an example case for a truly unknown outcome. Generically designs did not extend to designated dynamic targets on the protein surface, determined from recent work identifying dynamic allosteric sites to the functional modes of these enzymes [278]. A single output of interest during the targeting of residue 90-120 of 1XPB showed a possible binding mode utilizing an disordered long interface, shown in Figure 4.9. The outlier result was passed into AlphaFold2 multimer prediction, Figure 4.9, where a near identical conformation between the protein and peptide, 0.2 Å RMSD, was predicted for this peptide and the 1XPB structure. Beta Lactamase is a soluble and highly globular protein, as such it is difficult to bind to the surface as there is a large penalty to enthalpy of solvation. Finding a structure that not only binds stably but also covers the enzymatic pocket is a promising result.

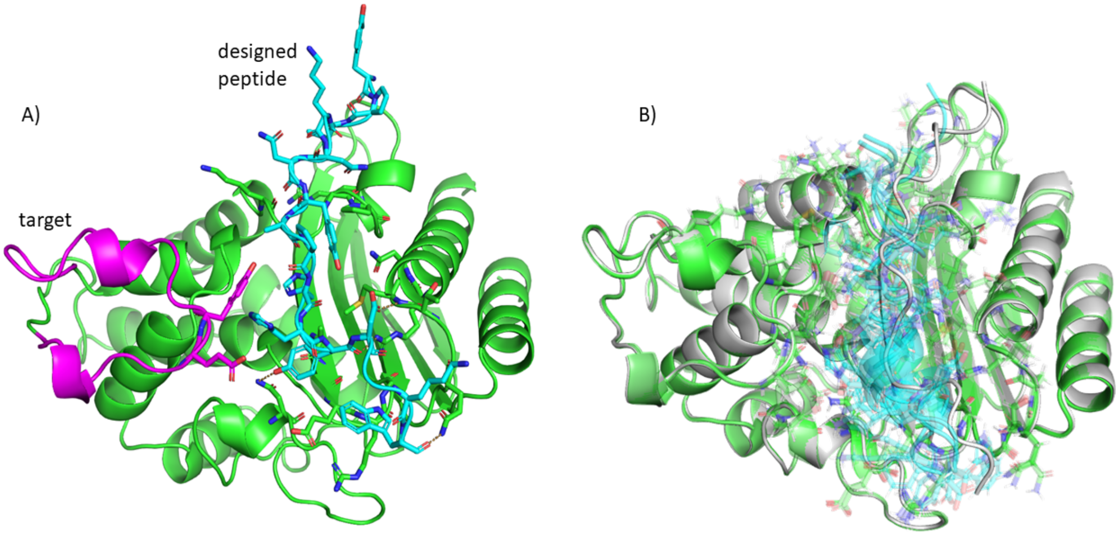


Figure 4.9: Designed peptide for 1XPB, targeting loop between helix 9 and 10, residues 90-120. A) the ESMFold predicted structure for the TEM-1 β Lactamase (green) and peptide (cyan) for a single design found to bind into the structure near the target site (magenta). B) AlphaFold2 multimer model v3 predictions for the 1XPB sequence and the designed peptide. Superposition with the original output in A (white) from pepStream.

Once again the disordered signalling hub protein p53 and disordered partner MDM2 are considered as interesting design targets. MDM2 has many peptide bound structures and even known motif, FXXXWXXL with synonymous variations [279]. Method set up was change partially to account for the conformation change in these systems; due to their inherent IDP nature this can be predicted to be a major factor of interaction. Only the termini of p53 and mdm2 were used, first or last 100 residues depending. In the MDM2 results observably some peptides were designed off target, but 80% designed to within 3 Å of the binding position of the p53 peptide in 4HFZ.

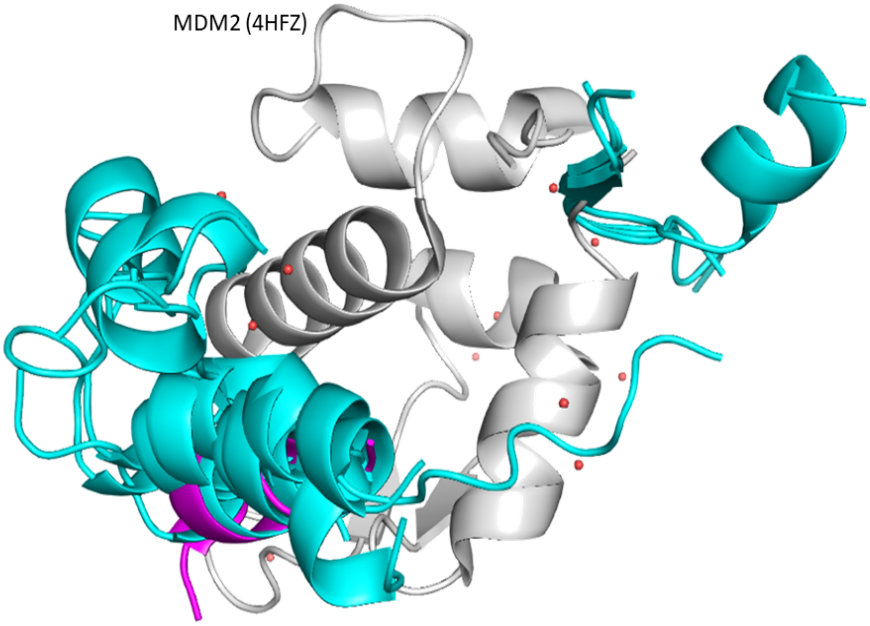


Figure 4.10: MDM2 alpha (uniprot P04637) designed peptides in cyan superimposed to PDBID 4HFZ, which has MDM2 in white and the bound peptide from p53 in magenta.

The p53 N-terminal is very infrequently folded in known structures. Under a conformation hypothesis, like DELLA there might be a specific conformation that we can target. If this is utilized as the initial conformation in the program perhaps this can overcome the disordered aspect, if that conformation is given as a constraint to the optimization. Using the PDBID 6XRE structure a segment of p53 near the N-terminal can be found folded and bound to RNA polymerase II (Pol II) Figure 4.11. Pol II, one of the three nuclear RNA polymerases in eukaryotes, is responsible for transcribing the genetic program, including protein-coding mRNAs and certain small non-coding RNAs. The transactivation domain of p53 N-ter binds the surface of Pol II's jaw in contact with DNA. Simply put, p53's functional domains directly regulate DNA binding activity of Pol II to mediate transcription.

Purposefully utilizing a "bound" conformation from 6XRE, we can see a clear struc-

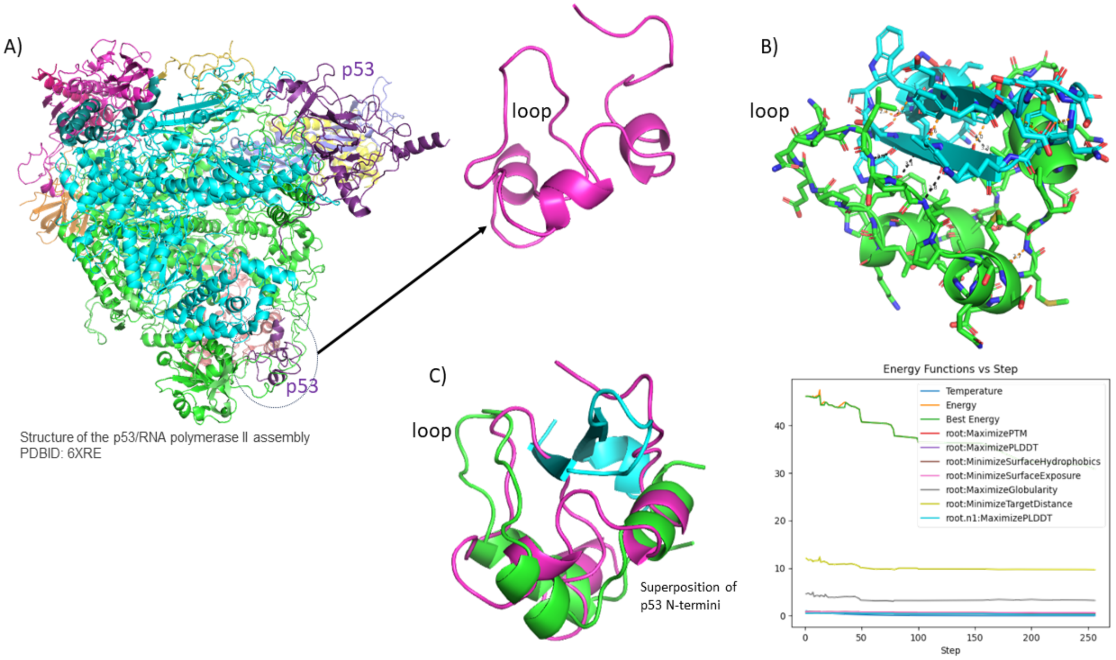


Figure 4.11: Designed peptide for p53 N-ter domain (MoRF1 - MDM2 interaction). A) PDBID 6XRE with the folded p53 N-ter in cartoon. The loop of the p53 structure is denoted for reference. B) A design from pepStream where the loop has been opened by a peptide interaction; the conformation induction can be seen occurring in the plotted meta data from the design process underneath by the sheer drops in the overall energy (green). C) Superposition of 6XRE N-ter and designed peptide complex of the same p53 residues.

ture interactions predicted. Importantly, some of these actually specifically seems to mimic the bound state of 6XRE. An anchored Beta-sheet formation of the peptide opens loops of TAD, preventing fit into Pol II shown in Figure 4.11. This peptide could feasibly bind into the bound state, as it is bound to Pol II receptor. It is notable that a large piece-mill drops in the optimization energy is likely from exporing other conformations possible between the p53 N-terminal and the designing peptide.

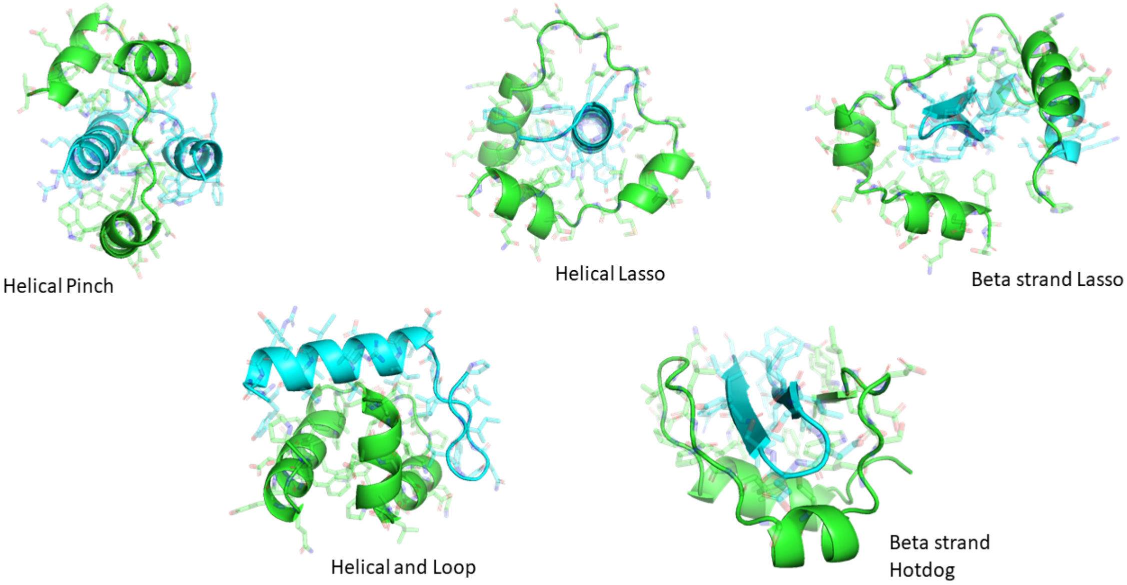


Figure 4.12: Demonstration of the diverse set of results obtained from predicting binding peptides (cyan) to the p53 N-terminal (MoRF1 – MDM2 interaction and TAD), shown in green.

Many possible interaction conformations are predicted in response to the peptide candidate, Figure 4.12. Utilizing structure to sequence methods a set of different interaction mode peptides could produce a large number of lead sequences to interact with the p53 N-termini. This kind of data would be applicable to expression systems for directed evolution experiments, where a diversity in initial sequences yields better successful outcomes [248].

There are not any bound or folded structures for the p53 C-terminal that were found in searching the PDB, barring examples of the tetramerization domain complexes. The interactions site of interest, denoted MoRF 4 previously, interacts with SIRT signaling proteins. From the design runs with the N-terminal a conformation change was desired, but no reference to direct it existed. So alteration to the core scoring methods inside the optimization of pepStream was done to remove the initial receptor conformation as a bias. The p53 C terminus was allowed to be designed freely, optimizing by prediction likelihoods and metrics like distance between peptide–protein interface.

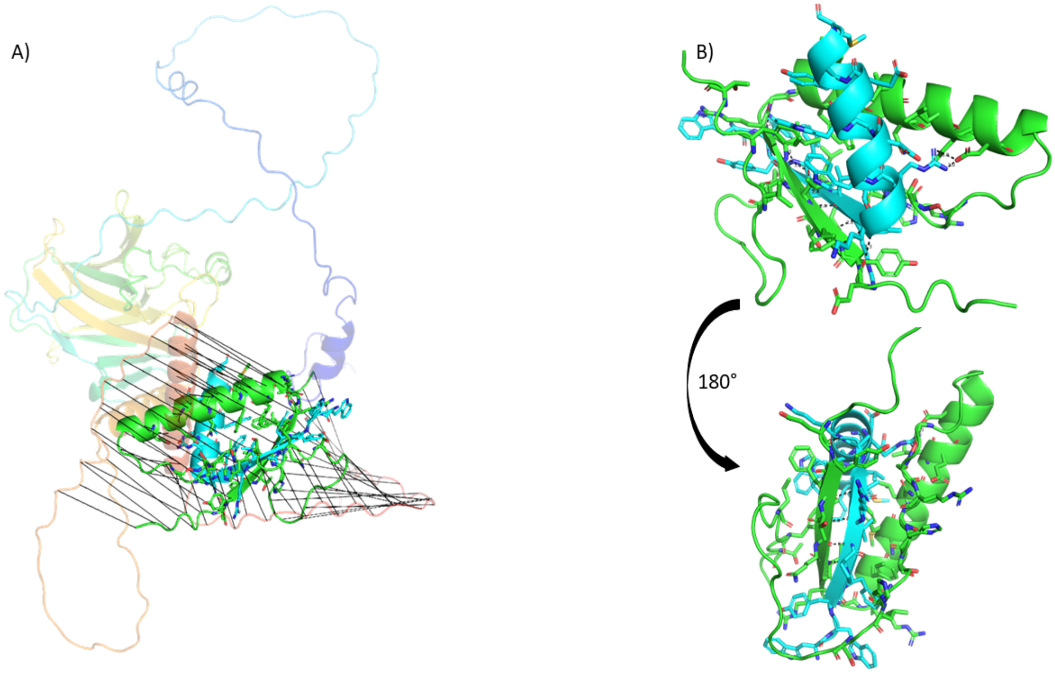


Figure 4.13: A) Uniprot P004637 p53 alpha canonical sequence. B) Example designed peptide anchored via β-sheet formation with the p53 monomer.

There were fewer successful examples of design for this site of p53 (5/10), determined by degree of folded p53 in the interaction. However, without the bias towards initial conformation rational binding peptides, Figure 4.13, were designed from a disordered predicted structure (Uniprot P004637). Notably, a similar signal in the optimization energy is found in the meta data plots produced from design of the successfully designs as the successful p53 and calmodulin designs.

4.4 Future and Ongoing Related Work

The ESM based pepStreaML can handle: discontinuous binding interfaces, large induced conformational changes, a variety of possible design rules, multi-design options, as well as immediately understandable outputs that can be passed into other structural analysis like MD or a Rosetta routine. There are benefits and draw backs to this method, such as inability to force a specific location for a peptide. Improvements to pepStreaML could be built directly from the ESM embeddings, or combine other models to improve the designs. For example ProteinMPNN could be used to guide the mutation process, a ML based MD simulation could be performed, but the inclusion of physics informed models will be of utmost importance to work towards a full chemical modeling of these biomolecules. In the coming years it is likely even more sophisticated architectures for machine learning will be devised, such as the forward forward model from Hinton that learns local objective functions during training time [280], that will shift levels of accuracy, implementation, or even trivialize certain issues discussed here. There is more work to be done in the computational structural biology field as a whole.

4.4.1 Refining the ESM Guided Design

The design of the classes within the pipeline facilitates mutability within encapsulated object types. Extending the energy functions is straightforward, involving the addition of relevant methods to generate metrics. Incorporating valuable energy functions like hydrophobic contacts and hydrogen bonding between the peptide and the target protein, with potential weighting of conserved residue interactions to enhance their significance in design, can be seamlessly integrated. Implementing these calculations is made easy through the utilization of a structural analysis package such as MDTraj. Additional energy functions, such as dipole moment and contact calculations, can be incorporated using the MDTraj package for precise measurements.

Tying in electrostatics for a design process likely would improve successful enzyme design, were this desired, or improve Columbic driven peptide interactions [281].

Currently two design modes are enabled and run in tandem, depending on user commands. Adding new epochs for design could be achieved by varying the weights of energies at different epochs, reinforcing various singular aspects during design. This could look like initializing a design to a local area, then apply against whole protein with only hydrogen bond scores, then introduce shape and probability scores, then end with purely pLDDT for structural accuracy maximization. Constant or variable size of the peptide is another design consideration that could be included easily to modulate outputs, particularly if a mutational library is desired for displays or other biological work. Passing in other references for design constraints would be easy to modulate a design mode with. And interesting design mode might include playing a game of predicting a peptide, using ESM-IF on the peptide, to reposition the sequence landscape being optimized. It could also be interest to cross seed a receptor using ESM-IF to generate a peptide that can bind to conserved aspects of the protein, similar in spirit to how antibodies are matured for influenza virus, targeting most conserved protein segments.

Conformation will always be a primary driving force of peptide interactions. Adding alternative steps, like FRODA, MD, or maybe diffusion based model step to the problem trying to account for conformation could benefit. In some cases flexibility information could be gleaned from the dynamics modelling, potentially offering estimates for entropy of the binding event. The only hindrance would be speed of calculation. Both diffusion based and mechanical based, like FRODA, could offer rapid enough calculations to add dynamic considerations to the predicted interface.

There are a variety of optional outputs that could be of interest to expose to the user such as saving a trajectory of designs or even logging the meta data associated with the design optimization more so that current is done; which only saves the energy values currently. With some work running on laptop level resources could be demonstrated for this version of pepStream, which would effectively bring this tool to any interested parties. Often in the software community sharing of python code can be done by sharing both the code and the environment using specialized tools like the anaconda software, which the entire pepStreamL software is developed in. While distribution is fairly easy, a higher end consumer grade GPU is required to generate the predictions. Not all researchers have access to recent hardware locally or via a HPC like resource. It maybe be feasible to scale down parts of the model, such as the core sequence model being used to score and predict the peptide sequence, to reduce local runtime demands. The openfold library can modulate computational demands using normal resource detection, this is the core software of the structure prediction being performed inside ESMFold [282, 283].

The original pepStream took over a week to run, the second version can run much faster and on fewer resources, but the ESM pepStream pipeline can run in minutes on a single GPU. There are great reasons to use a explainable pipeline like pepStream version 2, people tend to like to track how everything was selected and possibly use that information for further development. But the speed and possible elimination of major conventional issue, like induced conformations, is too tantalizing not to develop both generative ML and classical algorithm method. More so, due to the speed and ease of running pepStreaML multiple baseline tests could be conducted, which clearly show a degree of accuracy in terms of design ability that was not demonstrated before in the pipeline. The ML adaption does show real promise, but the best path forward will likely be a fusion of ML with classical techniques like MD and docking.

4.4.2 Models Building Models

Interesting aspect of a large unsupervised model is that there are hidden relationships in the embedding space of a 3 billion parameter model, unsurprisingly. High dimensional univariate vectors can be generated from sequences using the model, subsequently these can be used to train other kinds of models on a variety of problems. The original authors demonstrated this with the beta-lactamase classification problem from sequences alone [284]. This process is simple as the PCA of the embedding output itself can be used to train smaller models, such as a Random Forest model, that can achieve reasonable accuracy due to the embedding space spreading the functional components over many variables, instead of stuck in a string of letters.

The ESM model can embed sequences into a coherent high dimensional space. This is performed by dropped the final output layer and simply storing the last hidden layer values for the respective sequence. Zero-shot transfer is an interesting capability of large scale language models, and represents a major point of departure from the unsupervised learning methods that are the basis for current state-of-the-art inference of protein structure and function. The capability for zero-shot transfer implies that a model can be trained once and then applied to perform inference for many tasks [284].One of the most salient features of good drugs is solubility in standard biological conditions, which could be trained in a variety of ways based off the existing model in the pipeline. Similar to work performed using the ESM language model to predict two objectives, solubility and practical usability for purification of proteins in *Escherichia coli*, which claims to obtain state-of-the-art performance. [285]. Additionally, this could be done for synthesis outcomes, or a variety of other important factors.

Understandable not to provoke an immune response and is generally considered to be an undesirable physiological response. Immunogenicity is an undesired trait of any drug, particular those that are protein based. Work utilizing the ESM-1b model, trained on UniRef50, ed that general protein language models can efficiently evolve human antibodies by suggesting mutations that are evolutionarily plausible, despite providing the model with no information about the target antigen, binding specificity or protein structure. Using this method they improved the binding affinities of four clinically relevant antibodies [286]. This would be another example of synergistic design feature in situ, enabled by the underlying foundational model used.

CHAPTER 5: Conclusions

The core of this work is to apply methods to problems that reside between computer science and other disciplines. Designing proteins is based off the fundamental hypothesis that they will interact in a predictable way, based off the rules of chemistry and physics. Fully representing this in silico is rather difficult. With advancement in computer science finally percolating up in the media, it is likely bleed over from popular methods will continue to drive developments in the structural bioinformatic field. This will extend to other complicated biological polymers like RNA, which already have shown such high value with mRNA vaccines. The value of being able to process, analyze, and create design hypotheses about these kinds of systems will likely become a more reliable and abundant technique.

5.1 Summary on pepStream and Peptide Design

The initial pipeline was thoughtfully rationalized and tested. The present versions incorporates theoretical advancements that remain to be validated, a promising prospect currently in progress. Initial DELLA experiments with pepStream worked quite well, for a design project. Essentially the tested peptides returned with a 70% specific binding classification. While it could not be determined if this was a more potent binding than previous peptides, it offers new scaffolds to modify using rational design. Success can be attributed to having experts on hand that helped rapidly arrive at the minimal necessary interaction needed. Observably, pepStream version 1 had great success with a disordered protein that had no real structural information, particularly for near homologs that show any kind of protein protein interaction. Relating the physical filter to known empirical measurements from the PFR cohort inspired confidence leading to attempt to design for p53 signalling system.

Results from MDM2 peptide generated from the pepStream showed competitive predicted interaction as compared to the native p53 peptide, showing a variety of possibly close binding peptide. The laboratory experiments will begin with MDM2 and a few carefully selected peptides to provide the best possible initial outcomes for successful outcomes. This will be iterated as results are generated from our cohort. It is interesting to note that some overlap of exact sequence in the version 2 MDM2 matched with original outputs, only 5 peptides though. Additionally, pepStream version 2 was developed and run against the key MoRFs of p53 that interact with MDM2 and SIRT2. In the next round of MST experimentation it is likely some of these candidates will be tested. In attempt to baseline the accuracy of pepStream, predictions were performed on a WW domain protein, a popular model domain for peptide designs as the binding motif is well documented. Using the 1JMQ structure is was determined that 93 of the 3196 possible candidates contained a diproline, PP, in the sequence. In the final result set of 319 peptides only 11 contained a PP in the sequence and 1 contained the canonical PPXY motif. Degenerate solutions to interfaces can exist, which seem to be the case as observed from predicting the pepStream version 2 WW domain peptide against the 1JMQ receptor. There are a variety of internal process that could recover the exact peptide motif of the WW domain, this will be a focus on improvements for this version of the pipeline. Using AlphaFold2 multimer prediction, predicted peptides from the first version and the second version of the pipeline against two MoRFs of p53. The second version on average had closer interfaces and candidates had biophysical properties that separated depending on the target given, which is expected for specifically binding peptides.

A machine learning embodiment of the pepStream pipeline was developed using the ESM, pepStreaML. The unsupervised LLM can be deployed to both predict sequences given structures as well as structures from sequences; and can optimized a peptide sequence against a protein target. Designs performed in this newest version included a range of protein–peptide interactions with known peptide motifs. The outputs captured successful results for all of the experimental systems tested, only partially failing to produce peptides with a unknown binding to the target protein for β Lactamase. Even for this difficult case the method produced a single feasible binding peptide, as determined by complex prediction using AlphaFold2. These results demonstrates an interesting potential solution to the limitation of the classical computational methodology for design; conformation induction. It was found that given correct initial conditions, a conformation induction in the target protein of a successful peptide design can occur. There are a few additional processes that should be added to this simplistic framework that has been implemented. Building simplistic models on top of the sequence prediction model used could add a rapid refining to the optimization and adding physics informed ML methods would greatly improve optimization. The optimization itself will be an important improvement, other algorithms and mixing with design epochs will be tested.

5.2 Additional Work

2023 Biophysical Society Annual Meeting 2023 Spring [poster] " Evolution Inspired Design of Binding Peptides for p53 in an Automated Workflow: pepStream."

2021 Biophysical Society Annual Meeting 2021 Spring (Virtual) [poster] “A tale of two binding pathways: Molecular dynamics study of the GID1A-GA-GAI system”

2019 Gibberellins 2019 (Olomouc, The Czech Republic) [poster] “Computational elucidation of the binding dynamics in the GA-GID1-DELLA complex”

2019 Biophysical Society Annual Meeting 2019 Spring (Baltimore, Maryland) [poster]

“Using QSFR to ascertain Beta Lactamase family functionality”

National Institutes of Health under award number 1R15GM146200-01. "Computationally designing peptides to interfere with p53-MDM2 and p53-sirtuin interactions"

The New Zealand MBIE (Ministry of Business, Innovation, and Employment) Endeavour programme grant C11X1804 to Erik Rikkernik October 2019

Bioinformatics Association of Students (BiAS) service (Vice president 2018, Treasurer 2019)

REFERENCES

[1] R. Roskoski Jr, “Nature’s robots: A history of proteins: Tanford, c., reynolds, j,” *Biochemistry and Molecular Biology Education*, vol. 30, no. 5, pp. 343–345, 2002.

[2] H. E. Sigerist, “The history of dietetics,” *Gesnerus*, vol. 46, no. 3-4, pp. 249–256, 1989.

[3] H. Hartley, “Origin of the word ‘protein’,” *Nature*, vol. 168, no. 4267, pp. 244–244, 1951.

[4] A. Piro, G. Tagarelli, P. Lagonia, A. Quattrone, and A. Tagarelli, “Archibald edward garrod and alcaptonuria: “inborn errors of metabolism” revisited,” *Genetics in Medicine*, vol. 12, no. 8, pp. 475–476, 2010.

[5] T. Wieland and M. Bodanszky, *Syntheses of Peptides. The First Epoch*, pp. 23–43. Berlin, Heidelberg: Springer Berlin Heidelberg, 1991.

[6] P. K. Robinson, “Enzymes: principles and biotechnological applications,” *Essays Biochem*, vol. 59, pp. 1–41, 2015. 1744-1358 Robinson, Peter K Journal Article Review England 2015/10/28 Essays Biochem. 2015;59:1-41. doi: 10.1042/bse0590001.

[7] A. O. Stretton, “The first sequence. fred sanger and insulin,” *Genetics*, vol. 162, no. 2, pp. 527–32, 2002. Stretton, Antony O W Biography Historical Article Journal Article United States 2002/10/26 Genetics. 2002 Oct;162(2):527-32. doi: 10.1093/genetics/162.2.527.

[8] J. B. Sumner, “The isolation and crystallization of the enzyme urease: Preliminary paper,” *Journal of Biological Chemistry*, vol. 69, no. 2, pp. 435–441, 1926.

[9] D. Eisenberg, “The discovery of the -helix and -sheet, the principal structural features of proteins,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 20, pp. 11207–11210, 2003.

[10] S. de Chadarevian, “John kendrew and myoglobin: Protein structure determination in the 1950s,” *Protein Sci*, vol. 27, no. 6, pp. 1136–1143, 2018. 1469-896x de Chadarevian, Soraya Biography Historical Article Journal Article United States 2018/04/03 Protein Sci. 2018 Jun;27(6):1136-1143. doi: 10.1002/pro.3417. Epub 2018 May 11.

[11] G. Némethy and H. A. Scheraga, “The structure of water and hydrophobic bonding in proteins. iii. the thermodynamic properties of hydrophobic bonds in proteins1,2,” *The Journal of Physical Chemistry*, vol. 66, no. 10, pp. 1773–1789, 1962. doi: 10.1021/j100816a004.

[12] H. Neurath, J. P. Greenstein, F. W. Putnam, and J. A. Erickson, "The chemistry of protein denaturation," *Chemical Reviews*, vol. 34, no. 2, pp. 157–265, 1944. doi: 10.1021/cr60108a003.

[13] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, pp. 223–230, 1973.

[14] A. R. Fersht, "From the first protein structures to our current knowledge of protein folding: delights and scepticisms," *Nature Reviews Molecular Cell Biology*, vol. 9, no. 8, pp. 650–654, 2008.

[15] P. D. Sun, C. E. Foster, and J. C. Boyington, "Overview of protein structural and functional folds," *Curr Protoc Protein Sci*, vol. Chapter 17, no. 1, p. Unit 17.1, 2004. 1934-3663 Sun, Peter D Foster, Christine E Boyington, Jeffrey C Journal Article Review United States 2008/04/23 Curr Protoc Protein Sci. 2004 May;Chapter 17(1):Unit 17.1. doi: 10.1002/0471140864.ps1701s35.

[16] H. M. Berman, T. N. Bhat, P. E. Bourne, Z. Feng, G. Gilliland, H. Weissig, and J. Westbrook, "The protein data bank and the challenge of structural genomics," *Nature Structural Biology*, vol. 7, no. 11, pp. 957–959, 2000.

[17] W. L. Nichols, G. D. Rose, L. F. Ten Eyck, and B. H. Zimm, "Rigid domains in proteins: an algorithmic approach to their identification," *Proteins*, vol. 23, no. 1, pp. 38–48, 1995. Nichols, W L Rose, G D Ten Eyck, L F Zimm, B H 1 PO1 HL48018/HL/NHLBI NIH HHS/United States GM11916/GM/NIGMS NIH HHS/United States GM29458/GM/NIGMS NIH HHS/United States Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. United States 1995/09/01 Proteins. 1995 Sep;23(1):38-48. doi: 10.1002/prot.340230106.

[18] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic, "Intrinsically disordered protein," *Journal of Molecular Graphics and Modelling*, vol. 19, no. 1, pp. 26–59, 2001.

[19] P. F. N. Faísca, "Knotted proteins: A tangled tale of structural biology," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 459–468, 2015.

[20] B. Gutte and R. B. Merrifield, "Total synthesis of an enzyme with ribonuclease a activity," *Journal of the American Chemical Society*, vol. 91, no. 2, pp. 501–502, 1969.

[21] B. Gutte, "A synthetic 70-amino acid residue analog of ribonuclease s-protein with enzymic activity," *Journal of Biological Chemistry*, vol. 250, no. 3, pp. 889–904, 1975.

[22] M. P. Frushicheva, J. Cao, and A. Warshel, "Challenges and advances in validating enzyme design proposals: the case of kemp eliminase catalysis," *Biochemistry*, vol. 50, no. 18, pp. 3849–58, 2011. 1520-4995 Frushicheva, Maria P Cao, Jie Warshel, Arieh R01 GM024492/GM/NIGMS NIH HHS/United States R01 GM024492-34/GM/NIGMS NIH HHS/United States R01 GM024492-35/GM/NIGMS NIH HHS/United States GM024492/GM/NIGMS NIH HHS/United States Journal Article Research Support, N.I.H., Extramural United States 2011/03/30 Biochemistry. 2011 May 10;50(18):3849-58. doi: 10.1021/bi200063a. Epub 2011 Apr 15.

[23] J. P. Schneider, A. Lombardi, and W. F. DeGrado, "Analysis and design of three-stranded coiled coils and three-helix bundles," *Fold Des*, vol. 3, no. 2, pp. R29–40, 1998. Schneider, J P Lombardi, A DeGrado, W F GM54616/GM/NIGMS NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. Review England 1998/05/05 Fold Des. 1998;3(2):R29-40. doi: 10.1016/S1359-0278(98)00011-X.

[24] D. N. Woolfson, "A brief history of de novo protein design: Minimal, rational, and computational," *Journal of Molecular Biology*, vol. 433, no. 20, p. 167160, 2021.

[25] J. T. Seffernick and S. Lindert, "Hybrid methods for combined experimental and computational determination of protein structure," *The Journal of Chemical Physics*, vol. 153, no. 24, p. 240901, 2020.

[26] S. K. Burley, A. Joachimiak, G. T. Montelione, and I. A. Wilson, "Contributions to the nih-nigms protein structure initiative from the psi production centers," *Structure*, vol. 16, no. 1, pp. 5–11, 2008.

[27] A. Bolje and S. Gobec, "Analytical techniques for structural characterization of proteins in solid pharmaceutical forms: An overview," *Pharmaceutics*, vol. 13, no. 4, p. 534, 2021.

[28] A. D. Robertson and K. P. Murphy, "Protein structure and the energetics of protein stability," *Chemical Reviews*, vol. 97, no. 5, pp. 1251–1268, 1997.

[29] S. K. Lüdemann, V. Lounnas, and R. C. Wade, "How do substrates enter and products exit the buried active site of cytochrome p450cam? 2. steered molecular dynamics and adiabatic mapping of substrate pathways11edited by j. thornton," *Journal of Molecular Biology*, vol. 303, no. 5, pp. 813–830, 2000.

[30] C. C. David and D. J. Jacobs, *Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins*, pp. 193–226. Totowa, NJ: Humana Press, 2014.

[31] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, "Essential dynamics of proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 17, no. 4, pp. 412–425, 1993.

[32] N. Perdigão, J. Heinrich, C. Stolte, K. S. Sabir, M. J. Buckley, B. Tabor, B. Signal, B. S. Gloss, C. J. Hammang, B. Rost, A. Schafferhans, and S. I. O'Donoghue, "Unexpected features of the dark proteome," *Proceedings of the National Academy of Sciences*, vol. 112, no. 52, pp. 15898–15903, 2015. doi: 10.1073/pnas.1508380112.

[33] C. M. Overall, "The human proteome: 90*Journal of Proteome Research*, vol. 19, no. 12, pp. 4731–4734, 2020. doi: 10.1021/acs.jproteome.0c00914.

[34] S. K. Burley, C. Bhikadiya, C. Bi, S. Bittrich, H. Chao, L. Chen, P. A. Craig, G. V. Crichlow, K. Dalenberg, J. M. Duarte, S. Dutta, M. Fayazi, Z. Feng, J. W. Flatt, S. Ganesan, S. Ghosh, D. S. Goodsell, R. K. Green, V. Guranovic, J. Henry, B. P. Hudson, I. Khokhriakov, C. L. Lawson, Y. Liang, R. Lowe, E. Peisach, I. Persikova, D. W. Piehl, Y. Rose, A. Sali, J. Segura, M. Sekharan, C. Shao, B. Vallat, M. Voigt, B. Webb, J. D. Westbrook, S. Whetstone, J. Y. Young, A. Zalevsky, and C. Zardecki, "Rcsb protein data bank (rcsb.org): delivery of experimentally-determined pdb structures alongside one million computed structure models of proteins from artificial intelligence/machine learning," *Nucleic Acids Research*, vol. 51, no. D1, pp. D488–D508, 2022.

[35] J. M. Chandonia, L. Guan, S. Lin, C. Yu, N. K. Fox, and S. E. Brenner, "Scope: improvements to the structural classification of proteins - extended database to facilitate variant interpretation and machine learning," *Nucleic Acids Res*, vol. 50, no. D1, pp. D553–d559, 2022.

[36] I. Sillitoe, N. Bordin, N. Dawson, V. P. Waman, P. Ashford, H. M. Scholes, C. S. M. Pang, L. Woodridge, C. Rauer, N. Sen, M. Abbasian, S. Le Cornu, S. D. Lam, K. Berka, I. H. Varekova, R. Svobodova, J. Lees, and C. A. Orengo, "Cath: increased structural coverage of functional space," *Nucleic Acids Res*, vol. 49, no. D1, pp. D266–d273, 2021.

[37] M. van Kempen, S. S. Kim, C. Tumescheit, M. Mirdita, J. Lee, C. L. M. Gilchrist, J. Söding, and M. Steinegger, "Fast and accurate protein structure search with foldseek," *Nature Biotechnology*, 2023.

[38] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, "Cath–a hierarchic classification of protein domain structures," *Structure*, vol. 5, no. 8, pp. 1093–108, 1997.

[39] J. M. Chandonia, L. Guan, S. Lin, C. Yu, N. K. Fox, and S. E. Brenner, "Scope: improvements to the structural classification of proteins - extended database to facilitate variant interpretation and machine learning," *Nucleic Acids Res*, vol. 50, no. D1, pp. D553–d559, 2022.

[40] B. Sun, F. S. Willard, D. Feng, J. Alsina-Fernandez, Q. Chen, M. Vieth, J. D. Ho, A. D. Showalter, C. Stutsman, L. Ding, T. M. Suter, J. D. Dunbar, J. W. Carpenter, F. A. Mohammed, E. Aihara, R. A. Brown, A. B. Bueno, P. J.

Emmerson, J. S. Moyers, T. S. Kobilka, M. P. Coghlan, B. K. Kobilka, and K. W. Sloop, "Structural determinants of dual incretin receptor agonism by tirzepatide," *Proceedings of the National Academy of Sciences*, vol. 119, no. 13, p. e2116506119, 2022.

[41] M. A. Anwar and S. Choi, "Structure-activity relationship in tlr4 mutations: Atomistic molecular dynamics simulations and residue interaction network analysis," *Scientific Reports*, vol. 7, no. 1, p. 43807, 2017.

[42] Y. Yang, J. Qin, H. Liu, and X. Yao, "Molecular dynamics simulation, free energy calculation and structure-based 3d-qsar studies of b-raf kinase inhibitors," *Journal of Chemical Information and Modeling*, vol. 51, no. 3, pp. 680–692, 2011. doi: 10.1021/ci100427j.

[43] D. B. Kokh, M. Amaral, J. Bomke, U. Grädler, D. Musil, H.-P. Buchstaller, M. K. Dreyer, M. Frech, M. Lowinski, F. Vallee, M. Bianciotto, A. Rak, and R. C. Wade, "Estimation of drug-target residence times by -random acceleration molecular dynamics simulations," *Journal of Chemical Theory and Computation*, vol. 14, no. 7, pp. 3859–3869, 2018.

[44] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, "The protein folding problem," *Annu Rev Biophys*, vol. 37, pp. 289–316, 2008.

[45] Y.-M. He and B.-G. Ma, "Abundance and temperature dependency of protein-protein interaction revealed by interface structure analysis and stability evolution," *Scientific Reports*, vol. 6, no. 1, p. 26737, 2016.

[46] S. L. Speer, W. Zheng, X. Jiang, I.-T. Chu, A. J. Guseman, M. Liu, G. J. Pielak, and C. Li, "The intracellular environment affects protein–protein interactions," *Proceedings of the National Academy of Sciences*, vol. 118, no. 11, p. e2019918118, 2021.

[47] I. D. Kuntz, K. Chen, K. A. Sharp, and P. A. Kollman, "The maximal affinity of ligands," *Proceedings of the National Academy of Sciences*, vol. 96, no. 18, pp. 9997–10002, 1999. doi: 10.1073/pnas.96.18.9997.

[48] B. Kuhlman and D. Baker, "Native protein sequences are close to optimal for their structures," *Proceedings of the National Academy of Sciences*, vol. 97, no. 19, pp. 10383–10388, 2000.

[49] R. Aurora and G. D. Rose, "Helix capping," *Protein Sci*, vol. 7, no. 1, pp. 21–38, 1998.

[50] Z. A. Levine and J.-E. Shea, "Simulations of disordered proteins and systems with conformational heterogeneity," *Current Opinion in Structural Biology*, vol. 43, pp. 95–103, 2017.

[51] V. N. Uversky, C. J. Oldfield, U. Midic, H. Xie, B. Xue, S. Vucetic, L. M. Iakoucheva, Z. Obradovic, and A. K. Dunker, "Unfoldomics of human diseases: linking protein intrinsic disorder with diseases," *BMC genomics*, vol. 10 Suppl 1, no. Suppl 1, pp. S7–S7, 2009.

[52] P. E. Wright and H. J. Dyson, "Intrinsically disordered proteins in cellular signalling and regulation," *Nature Reviews Molecular Cell Biology*, vol. 16, no. 1, pp. 18–29, 2015.

[53] V. N. Uversky, "Natively unfolded proteins: a point where biology waits for physics," *Protein Sci*, vol. 11, no. 4, pp. 739–56, 2002.

[54] C. J. Oldfield, Y. Cheng, M. S. Cortese, C. J. Brown, V. N. Uversky, and A. K. Dunker, "Comparing and combining predictors of mostly disordered proteins," *Biochemistry*, vol. 44, no. 6, pp. 1989–2000, 2005.

[55] H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," *Nature Reviews Molecular Cell Biology*, vol. 6, no. 3, pp. 197–208, 2005.

[56] A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker, and V. N. Uversky, "Analysis of molecular recognition features (morfs)," *J Mol Biol*, vol. 362, no. 5, pp. 1043–59, 2006.

[57] X. L. Sun, W. T. Jones, D. Harvey, P. J. B. Edwards, S. M. Pascal, C. Kirk, T. Considine, D. J. Sheerin, J. Rakonjac, C. J. Oldfield, B. Xue, A. K. Dunker, and V. N. Uversky, "N-terminal domains of della proteins are intrinsically unstructured in the absence of interaction with gid1/ga receptors," *Journal of Biological Chemistry*, vol. 285, no. 15, pp. 11557–11571, 2010.

[58] F. Shahidi and Y. Zhong, "Bioactive peptides," *Journal of AOAC INTERNATIONAL*, vol. 91, no. 4, pp. 914–931, 2008.

[59] K. F. Chai, A. Y. H. Voo, and W. N. Chen, "Bioactive peptides from food fermentation: A comprehensive review of their sources, bioactivities, applications, and future development," *Comprehensive Reviews in Food Science and Food Safety*, vol. 19, no. 6, pp. 3825–3885, 2020.

[60] R. E. W. Hancock and H.-G. Sahl, "Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies," *Nature Biotechnology*, vol. 24, no. 12, pp. 1551–1557, 2006.

[61] S. Sachdeva, "Peptides as 'drugs': The journey so far," *International Journal of Peptide Research and Therapeutics*, vol. 23, no. 1, pp. 49–60, 2017.

[62] B. P. Wallner and M. L. Gefter, "Immunotherapy with t-cell-reactive peptides derived from allergens," *Allergy*, vol. 49, no. 5, pp. 302–308, 1994.

[63] T. Q. N. Nguyen, Y. W. Tooh, R. Sugiyama, T. P. D. Nguyen, M. Purushothaman, L. C. Leow, K. Hanif, R. H. S. Yong, I. Agatha, F. R. Winnerdy, M. Gugger, A. T. Phan, and B. I. Morinaka, "Post-translational formation of strained cyclophanes in bacteria," *Nature Chemistry*, vol. 12, no. 11, pp. 1042–1053, 2020.

[64] F. M. Gribble and F. Reimann, "Metabolic messengers: glucagon-like peptide 1," *Nature Metabolism*, vol. 3, no. 2, pp. 142–148, 2021.

[65] A. Müller, M. Wenzel, H. Strahl, F. Grein, T. N. V. Saaki, B. Kohl, T. Siersma, J. E. Bandow, H.-G. Sahl, T. Schneider, and L. W. Hamoen, "Daptomycin inhibits cell envelope synthesis by interfering with fluid membrane microdomains," *Proceedings of the National Academy of Sciences*, vol. 113, no. 45, pp. E7077–E7086, 2016.

[66] M. Lindgren, M. Hällbrink, A. Prochiantz, and Langel, "Cell-penetrating peptides," *Trends in Pharmacological Sciences*, vol. 21, no. 3, pp. 99–103, 2000.

[67] R. Taheri-Ledari and A. Maleki, "Antimicrobial therapeutic enhancement of levofloxacin via conjugation to a cell-penetrating peptide: An efficient sonochemical catalytic process," *Journal of Peptide Science*, vol. 26, no. 10, p. e3277, 2020. https://doi.org/10.1002/psc.3277.

[68] S. Y. Tan and J. Merchant, "Frederick banting (1891-1941): Discoverer of insulin," *Singapore Med J*, vol. 58, no. 1, pp. 2–3, 2017.

[69] W. C. Kenney, W. McIntire, and D. J. Steenkamp, "Amino acid sequence of a cofactor peptide from trimethylamine dehydrogenase," *FEBS Letters*, vol. 85, no. 1, pp. 137–140, 1978.

[70] Y. Gao, F. Zhao, Q. Wang, Y. Zhang, and B. Xu, "Small peptide nanofibers as the matrices of molecular hydrogels for mimicking enzymes and enhancing the activity of enzymes," *Chemical Society Reviews*, vol. 39, no. 9, p. 3425, 2010.

[71] L. Sun, Z. Fan, Y. Wang, Y. Huang, M. Schmidt, and M. Zhang, "Tunable synthesis of self-assembled cyclic peptide nanotubes and nanoparticles," *Soft Matter*, vol. 11, no. 19, pp. 3822–3832, 2015.

[72] S. Hinuma, Y. Habata, R. Fujii, Y. Kawamata, M. Hosoya, S. Fukusumi, C. Kitada, Y. Masuo, T. Asano, H. Matsumoto, M. Sekiguchi, T. Kurokawa, O. Nishimura, H. Onda, and M. Fujino, "A prolactin-releasing peptide in the brain," *Nature*, vol. 393, no. 6682, pp. 272–276, 1998.

[73] A. Stavrakoudis, S. Makropoulou, V. Tsikaris, M. Sakarellos-Daitsiotis, C. Sakarellos, and I. N. Demetropoulos, "Computational screening of branched cyclic peptide motifs as potential enzyme mimetics," *Journal of Peptide Science*, vol. 9, no. 3, pp. 145–155, 2003. https://doi.org/10.1002/psc.441.

[74] S. Galdiero, A. Falanga, R. Tarallo, L. Russo, E. Galdiero, M. Cantisani, G. Morelli, and M. Galdiero, "Peptide inhibitors against herpes simplex virus infections," *Journal of Peptide Science*, vol. 19, no. 3, pp. 148–158, 2013. https://doi.org/10.1002/psc.2489.

[75] A. C.-L. Lee, J. L. Harris, K. K. Khanna, and J.-H. Hong, "A comprehensive review on current advances in peptide drug development and design," *International Journal of Molecular Sciences*, vol. 20, no. 10, p. 2383, 2019.

[76] F. Barany, "Single-stranded hexameric linkers: a system for in-phase insertion mutagenesis and protein engineering," *Gene*, vol. 37, no. 1, pp. 111–123, 1985.

[77] F. Kawai, A. Nakamura, A. Visootsat, and R. Iino, "Plasmid-based one-pot saturation mutagenesis and robot-based automated screening for protein engineering," *ACS Omega*, vol. 3, no. 7, pp. 7715–7726, 2018.

[78] N. Chennamsetty, V. Voynov, V. Kayser, B. Helk, B. L. Trout, and A. M. Klibanov, "Design of therapeutic proteins with enhanced stability," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 29, pp. 11937–11942, 2009.

[79] W. F. DeGrado, L. Regan, and S. P. Ho, "The design of a four-helix bundle protein," *Cold Spring Harb Symp Quant Biol*, vol. 52, pp. 521–6, 1987. DeGrado, W F Regan, L Ho, S P Journal Article United States 1987/01/01 Cold Spring Harb Symp Quant Biol. 1987;52:521-6. doi: 10.1101/sqb.1987.052.01.059.

[80] J. W. Ponder and F. M. Richards, "Tertiary templates for proteins. use of packing criteria in the enumeration of allowed sequences for different structural classes," *J Mol Biol*, vol. 193, no. 4, pp. 775–91, 1987. Ponder, J W Richards, F M GM22778/GM/NIGMS NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. Netherlands 1987/02/20 J Mol Biol. 1987 Feb 20;193(4):775-91. doi: 10.1016/0022-2836(87)90358-5.

[81] B. I. Dahiyat and S. L. Mayo, "De novo protein design: Fully automated sequence selection," *Science*, vol. 278, no. 5335, pp. 82–87, 1997.

[82] S. T. Walsh, H. Cheng, J. W. Bryson, H. Roder, and W. F. DeGrado, "Solution structure and dynamics of a de novo designed three-helix bundle protein," *Proc Natl Acad Sci U S A*, vol. 96, no. 10, pp. 5486–91, 1999. 1091-6490 Walsh, S T Cheng, H Bryson, J W Roder, H DeGrado, W F GM-54616/GM/NIGMS NIH HHS/United States R01 GM054616/GM/NIGMS NIH HHS/United States GM-35926/GM/NIGMS NIH HHS/United States R37 GM054616/GM/NIGMS NIH HHS/United States P30 CA006927/CA/NCI NIH HHS/United States CA-06927/CA/NCI NIH HHS/United States Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. United States 1999/05/13 Proc Natl Acad Sci U S A. 1999 May 11;96(10):5486-91. doi: 10.1073/pnas.96.10.5486.

[83] M. Shirts and V. S. Pande, "Computing: Screen savers of the world unite!," *Science*, vol. 290, no. 5498, pp. 1903–4, 2000. Shirts, M Pande, V S Journal Article United States 2007/08/31 Science. 2000 Dec 8;290(5498):1903-4. doi: 10.1126/science.290.5498.1903.

[84] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, "Design of a novel globular protein fold with atomic-level accuracy," *Science*, vol. 302, no. 5649, pp. 1364–1368, 2003.

[85] H. Zhou, Y. Zhang, and Z.-C. Ou-Yang, "Stretch-induced hairpin-coil transitions in designed polynucleotide chains," *Physical Review Letters*, vol. 86, no. 2, pp. 356–359, 2001. PRL.

[86] I. V. Korendovych and W. F. DeGrado, "De novo protein design, a retrospective," *Q Rev Biophys*, vol. 53, p. e3, 2020. 1469-8994 Korendovych, Ivan V Orcid: 0000-0001-8144-783x DeGrado, William F R35 GM119634/GM/NIGMS NIH HHS/United States R35 GM122603/GM/NIGMS NIH HHS/United States Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Review England 2020/02/12 Q Rev Biophys. 2020 Feb 11;53:e3. doi: 10.1017/S0033583519000131.

[87] T. Lazaridis and M. Karplus, "Effective energy functions for protein structure prediction," *Current Opinion in Structural Biology*, vol. 10, no. 2, pp. 139–145, 2000.

[88] A. Bazzoli, A. G. Tettamanzi, and Y. Zhang, "Computational protein design and large-scale assessment by i-tasser structure assembly simulations," *J Mol Biol*, vol. 407, no. 5, pp. 764–76, 2011. 1089-8638 Bazzoli, Andrea Tettamanzi, Andrea G B Zhang, Yang R01 GM083107-01A1/GM/NIGMS NIH HHS/United States R01 GM084222/GM/NIGMS NIH HHS/United States GM083107/GM/NIGMS NIH HHS/United States R01 GM083107/GM/NIGMS NIH HHS/United States R01 GM084222-01A1/GM/NIGMS NIH HHS/United States GM084222/GM/NIGMS NIH HHS/United States Journal Article Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S. Netherlands 2011/02/19 J Mol Biol. 2011 Apr 15;407(5):764-76. doi: 10.1016/j.jmb.2011.02.017. Epub 2011 Feb 15.

[89] R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, J. Dunbrack, R. L., R. Das, D. Baker, B. Kuhlman, T. Kortemme, and J. J. Gray, "The rosetta all-atom energy function for macromolecular modeling and design," *J Chem Theory Comput*, vol. 13, no. 6, pp. 3031–3048, 2017.

[90] P. Carbonell and J. Y. Trosset, "Computational protein design methods for synthetic biology," *Methods Mol Biol*, vol. 1244, pp. 3–21, 2015. 1940-6029

Carbonell, Pablo Trosset, Jean-Yves Journal Article Research Support, Non-U.S. Gov't United States 2014/12/10 Methods Mol Biol. 2015;1244:3-21. doi: 10.1007/978-1-4939-1878-2$_1$.

[91] M. Vandermies and P. Fickers, "Bioreactor-scale strategies for the production of recombinant protein in the yeast yarrowia lipolytica," *Microorganisms*, vol. 7, no. 2, 2019. 2076-2607 Vandermies, Marie Orcid: 0000-0003-0928-8955 Fickers, Patrick Journal Article Review Switzerland 2019/02/02 Microorganisms. 2019 Jan 30;7(2):40. doi: 10.3390/microorganisms7020040.

[92] S. R. Rudge and M. R. Ladisch, "Industrial challenges of recombinant proteins," *Adv Biochem Eng Biotechnol*, vol. 171, pp. 1–22, 2020. Rudge, Scott R Ladisch, Michael R Journal Article Germany 2019/12/19 Adv Biochem Eng Biotechnol. 2020;171:1-22. doi: 10.1007/10$_2$019$_1$20.

[93] F. H. Arnold, "Protein engineering for unusual environments," *Current Opinion in Biotechnology*, vol. 4, no. 4, pp. 450–455, 1993.

[94] F. H. Arnold, "Design by directed evolution," *Accounts of Chemical Research*, vol. 31, no. 3, pp. 125–131, 1998. doi: 10.1021/ar960017f.

[95] D. J. Mandell and T. Kortemme, "Backbone flexibility in computational protein design," *Current Opinion in Biotechnology*, vol. 20, no. 4, pp. 420–428, 2009.

[96] T. Kortemme, D. E. Kim, and D. Baker, "Computational alanine scanning of protein-protein interfaces," *Sci STKE*, vol. 2004, no. 219, p. pl2, 2004.

[97] B. Kuhlman and P. Bradley, "Advances in protein structure prediction and design," *Nature Reviews Molecular Cell Biology*, vol. 20, no. 11, pp. 681–697, 2019.

[98] C. Avery, J. Patterson, T. Grear, T. Frater, and D. J. Jacobs, "Protein function analysis through machine learning," *Biomolecules*, vol. 12, no. 9, 2022. 2218-273x Avery, Chris Patterson, John Grear, Tyler Orcid: 0000-0002-8319-8432 Frater, Theodore Jacobs, Donald J Orcid: 0000-0001-7711-1639 R15 GM146200/GM/NIGMS NIH HHS/United States R15GM146200/NH/NIH HHS/United States Journal Article Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S. Review Switzerland 2022/09/24 Biomolecules. 2022 Sep 6;12(9):1246. doi: 10.3390/biom12091246.

[99] G. Li, Y. Qin, N. T. Fontaine, M. Ng Fuk Chong, M. A. Maria-Solano, F. Feixas, X. F. Cadet, R. Pandjaitan, M. Garcia-Borràs, F. Cadet, and M. T. Reetz, "Machine learning enables selection of epistatic enzyme mutants for stability against unfolding and detrimental aggregation," *ChemBioChem*, vol. 22, no. 5, pp. 904–914, 2021.

[100] Z. Chen, R. D. Kibler, A. Hunt, F. Busch, J. Pearl, M. Jia, Z. L. VanAernum, B. I. M. Wicky, G. Dods, H. Liao, M. S. Wilken, C. Ciarlo, S. Green, H. El-Samad, J. Stamatoyannopoulos, V. H. Wysocki, M. C. Jewett, S. E. Boyken, and D. Baker, "De novo design of protein logic gates," *Science*, vol. 368, no. 6486, pp. 78–84, 2020.

[101] L. Piela, ed., *Chapter 8 - ELECTRONIC MOTION IN THE MEAN FIELD: ATOMS AND MOLECULES*, pp. 324–427. Amsterdam: Elsevier, 2007.

[102] M. Karplus and J. Kuriyan, "Molecular dynamics and protein function," *Proceedings of the National Academy of Sciences*, vol. 102, no. 19, pp. 6679–6685, 2005.

[103] K. Vanommeslaeghe, O. Guvench, and J. MacKerell, A. D., "Molecular mechanics," *Curr Pharm Des*, vol. 20, no. 20, pp. 3281–92, 2014.

[104] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1-2, pp. 19–25, 2015.

[105] S. A. Hollingsworth and R. O. Dror, "Molecular dynamics simulation for all," *Neuron*, vol. 99, no. 6, pp. 1129–1143, 2018. 1097-4199 Hollingsworth, Scott A Dror, Ron O R01 GM127359/GM/NIGMS NIH HHS/United States T15 LM007033/LM/NLM NIH HHS/United States Journal Article Research Support, N.I.H., Extramural Review United States 2018/09/22 Neuron. 2018 Sep 19;99(6):1129-1143. doi: 10.1016/j.neuron.2018.08.011.

[106] J. A. Stevens, F. Grünewald, P. A. M. van Tilburg, M. König, B. R. Gilbert, T. A. Brier, Z. R. Thornburg, Z. Luthey-Schulten, and S. J. Marrink, "Molecular dynamics simulation of an entire cell," *Frontiers in Chemistry*, vol. 11, 2023.

[107] R. E. Amaro and A. J. Mulholland, "Multiscale methods in drug design bridge chemical and biological complexity in the search for cures," *Nature Reviews Chemistry*, vol. 2, no. 4, p. 0148, 2018.

[108] J. D. Durrant, S. E. Kochanek, L. Casalino, P. U. Ieong, A. C. Dommer, and R. E. Amaro, "Mesoscale all-atom influenza virus simulations suggest new substrate binding mechanism," *ACS Central Science*, vol. 6, no. 2, pp. 189–196, 2020. doi: 10.1021/acscentsci.9b01071.

[109] A. Laio and M. Parrinello, "Escaping free-energy minima," *Proc Natl Acad Sci U S A*, vol. 99, no. 20, pp. 12562–6, 2002.

[110] A. Barducci, G. Bussi, and M. Parrinello, "Well-tempered metadynamics: a smoothly converging and tunable free-energy method," *Phys Rev Lett*, vol. 100, no. 2, p. 020603, 2008.

[111] M. De Vivo, M. Masetti, G. Bottegoni, and A. Cavalli, "Role of molecular dynamics and related methods in drug discovery," *Journal of Medicinal Chemistry*, vol. 59, no. 9, pp. 4035–4061, 2016. doi: 10.1021/acs.jmedchem.5b01684.

[112] M. Feig, G. Nawrocki, I. Yu, P.-H. Wang, and Y. Sugita, "Challenges and opportunities in connecting simulations with experiments via molecular dynamics of cellular environments," *Journal of Physics: Conference Series*, vol. 1036, p. 012010, 2018.

[113] R. Wadhwa, N. S. Yadav, S. P. Katiyar, T. Yaguchi, C. Lee, H. Ahn, C.-O. Yun, S. C. Kaul, and D. Sundar, “Molecular dynamics simulations and experimental studies reveal differential permeability of withaferin-a and withanone across the model cell membrane,” *Scientific Reports*, vol. 11, no. 1, p. 2352, 2021.

[114] I. D. Zelnik, B. Mestre, J. J. Weinstein, T. Dingjan, S. Izrailov, S. Ben-Dor, S. J. Fleishman, and A. H. Futerman, “Computational design and molecular dynamics simulations suggest the mode of substrate binding in ceramide synthases,” *Nature Communications*, vol. 14, no. 1, p. 2330, 2023.

[115] R. E. Amaro and A. J. Mulholland, “Multiscale methods in drug design bridge chemical and biological complexity in the search for cures,” *Nature Reviews Chemistry*, vol. 2, no. 4, p. 0148, 2018.

[116] J. D. Durrant and J. A. McCammon, “Molecular dynamics simulations and drug discovery,” *BMC Biology*, vol. 9, no. 1, p. 71, 2011.

[117] T. Yamashita, “Toward rational antibody design: recent advancements in molecular dynamics simulations,” *International Immunology*, vol. 30, no. 4, pp. 133–140, 2018.

[118] J. A. McCammon, B. R. Gelin, and M. Karplus, “Dynamics of folded proteins,” *Nature*, vol. 267, no. 5612, pp. 585–90, 1977. McCammon, J A Gelin, B R Karplus, M Journal Article Research Support, U.S. Gov’t, Non-P.H.S. Research Support, U.S. Gov’t, P.H.S. England 1977/06/16 Nature. 1977 Jun 16;267(5612):585-90. doi: 10.1038/267585a0.

[119] D. J. Jacobs and S. Dallakyan, “Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity,” *Biophys J*, vol. 88, no. 2, pp. 903–15, 2005.

[120] M. A. Anwar and S. Choi, “Structure-activity relationship in tlr4 mutations: Atomistic molecular dynamics simulations and residue interaction network analysis,” *Scientific Reports*, vol. 7, no. 1, p. 43807, 2017.

[121] T. Zou, V. A. Risso, J. A. Gavira, J. M. Sanchez-Ruiz, and S. B. Ozkan, “Evolution of conformational dynamics determines the conversion of a promiscuous generalist into a specialist enzyme,” *Mol Biol Evol*, vol. 32, no. 1, pp. 132–43, 2015.

[122] B. Chowdhury and G. Garai, “A review on multiple sequence alignment from the perspective of genetic algorithm,” *Genomics*, vol. 109, no. 5, pp. 419–431, 2017.

[123] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” *J Mol Biol*, vol. 48, no. 3, pp. 443–53, 1970. Needleman, S B Wunsch, C D Journal Article Netherlands 1970/03/01 J Mol Biol. 1970 Mar;48(3):443-53. doi: 10.1016/0022-2836(70)90057-4.

[124] M. Chatzou, C. Magis, J.-M. Chang, C. Kemena, G. Bussotti, I. Erb, and C. Notredame, “Multiple sequence alignment modeling: methods and applications,” *Briefings in Bioinformatics*, vol. 17, no. 6, pp. 1009–1023, 2015.

[125] G. M. Boratyn, A. A. Schäffer, R. Agarwala, S. F. Altschul, D. J. Lipman, and T. L. Madden, "Domain enhanced lookup time accelerated blast," *Biol Direct*, vol. 7, p. 12, 2012.

[126] E. Eser, T. Can, and H. Ferhatosmanoğlu, "Div-blast: diversification of sequence search results," *PLoS One*, vol. 9, no. 12, p. e115445, 2014. 1932-6203 Eser, Elif Can, Tolga Ferhatosmanoğlu, Hakan Journal Article Research Support, Non-U.S. Gov't United States 2014/12/23 PLoS One. 2014 Dec 22;9(12):e115445. doi: 10.1371/journal.pone.0115445. eCollection 2014.

[127] T. S. Chen and A. E. Keating, "Designing specific protein-protein interactions using computation, experimental library screening, or integrated methods," *Protein Sci*, vol. 21, no. 7, pp. 949–63, 2012. 1469-896x Chen, T Scott Keating, Amy E GM067681/GM/NIGMS NIH HHS/United States GM084181/GM/NIGMS NIH HHS/United States P50-GM68762/GM/NIGMS NIH HHS/United States Journal Article Research Support, N.I.H., Extramural Review United States 2012/05/18 Protein Sci. 2012 Jul;21(7):949-63. doi: 10.1002/pro.2096. Epub 2012 Jun 8.

[128] R. D. Taylor, P. J. Jewsbury, and J. W. Essex, "A review of protein-small molecule docking methods," *J Comput Aided Mol Des*, vol. 16, no. 3, pp. 151–66, 2002. Taylor, R D Jewsbury, P J Essex, J W Comparative Study Journal Article Research Support, Non-U.S. Gov't Review Netherlands 2002/10/05 J Comput Aided Mol Des. 2002 Mar;16(3):151-66. doi: 10.1023/a:1020155510718.

[129] J. Zhu, N. Avakyan, A. Kakkis, A. M. Hoffnagle, K. Han, Y. Li, Z. Zhang, T. S. Choi, Y. Na, C.-J. Yu, and F. A. Tezcan, "Protein assembly by design," *Chemical Reviews*, vol. 121, no. 22, pp. 13701–13796, 2021. doi: 10.1021/acs.chemrev.1c00308.

[130] C. Dominguez, R. Boelens, and A. M. J. J. Bonvin, "Haddock: a proteinprotein docking approach based on biochemical or biophysical information," *Journal of the American Chemical Society*, vol. 125, no. 7, pp. 1731–1737, 2003. doi: 10.1021/ja026939x.

[131] J. Mintseris, B. Pierce, K. Wiehe, R. Anderson, R. Chen, and Z. Weng, "Integrating statistical pair potentials into protein complex prediction," *Proteins*, vol. 69, no. 3, pp. 511–20, 2007.

[132] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, and H. J. Wolfson, "Patchdock and symmdock: servers for rigid and symmetric docking," *Nucleic Acids Res*, vol. 33, no. Web Server issue, pp. W363–7, 2005. 1362-4962 Schneidman-Duhovny, Dina Inbar, Yuval Nussinov, Ruth Wolfson, Haim J N01CO12400/CA/NCI NIH HHS/United States N01-CO-12400/CO/NCI NIH HHS/United States Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. England 2005/06/28 Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W363-7. doi: 10.1093/nar/gki481.

[133] M. Ohue, Y. Matsuzaki, N. Uchikoga, T. Ishida, and Y. Akiyama, "Megadock: an all-to-all protein-protein interaction prediction system using tertiary structure data,"

Protein Pept Lett, vol. 21, no. 8, pp. 766–78, 2014. 1875-5305 Ohue, Masahito Matsuzaki, Yuri Uchikoga, Nobuyuki Ishida, Takashi Akiyama, Yutaka Journal Article Research Support, Non-U.S. Gov't Netherlands 2013/07/17 Protein Pept Lett. 2014;21(8):766-78. doi: 10.2174/09298665113209990050.

[134] O. J. Dodd, A. M. Acevedo-Jake, A. R. Azizogli, V. K. Mulligan, and V. A. Kumar, "How to design peptides," *Methods Mol Biol*, vol. 2597, pp. 187–216, 2023. 1940-6029 Dodd-O, Joseph Acevedo-Jake, Amanda M Azizogli, Abdul-Rahman Mulligan, Vikram Khipple Kumar, Vivek A R01 DE031812/DE/NIDCR NIH HHS/United States R15 EY029504/EY/NEI NIH HHS/United States R21 AR079708/AR/NIAMS NIH HHS/United States Journal Article United States 2022/11/15 Methods Mol Biol. 2023;2597:187-216. doi: 10.1007/978-1-0716-2835-5$_1$5.

[135] M. Sarkar and R. Ranbhor, *Chapter 9 - Patents in peptide science*, pp. 255–275. Academic Press, 2023.

[136] C. E. Milligan, D. Prevette, H. Yaginuma, S. Homma, C. Cardwellt, L. C. Fritz, K. J. Tomaselli, R. W. Oppenheim, and L. M. Schwartz, "Peptide inhibitors of the ice protease family arrest programmed cell death of motoneurons in vivo and in vitro," *Neuron*, vol. 15, no. 2, pp. 385–393, 1995.

[137] L. Li, Y. Shi, M. J. Cheserek, G. Su, and G. Le, "Antibacterial activity and dual mechanisms of peptide analog derived from cell-penetrating peptide against salmonella typhimurium and streptococcus pyogenes," *Applied Microbiology and Biotechnology*, vol. 97, no. 4, pp. 1711–1723, 2013.

[138] H. Higashi, S. Yoshida, K. Sato, and T. Yamagata, "Interaction of ganglioside with specific peptide sequences as a mechanism for the modulation of calmodulin-dependent enzymes," *Journal of Biochemistry*, vol. 120, no. 1, pp. 66–73, 1996.

[139] B. G. Pierce, Y. Hourai, and Z. Weng, "Accelerating protein docking in zdock using an advanced 3d convolution library," *PLoS One*, vol. 6, no. 9, p. e24657, 2011.

[140] C. M. Rufo, Y. S. Moroz, O. V. Moroz, J. Stöhr, T. A. Smith, X. Hu, W. F. Degrado, and I. V. Korendovych, "Short peptides self-assemble to produce catalytic amyloids," *Nature Chemistry*, vol. 6, no. 4, pp. 303–309, 2014.

[141] O. Carny and E. Gazit, "A model for the role of short self-assembled peptides in the very early stages of the origin of life," *The FASEB Journal*, vol. 19, no. 9, pp. 1051–1055, 2005.

[142] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–637, 1983. Kabsch, W Sander, C Journal Article Research Support, Non-U.S. Gov't United States Biopolymers. 1983 Dec;22(12):2577-637. doi: 10.1002/bip.360221211.

[143] Y. Zhang and C. Sagui, "Secondary structure assignment for conformationally irregular peptides: Comparison between dssp, stride and kaksi," *Journal of Molecular Graphics and Modelling*, vol. 55, pp. 72–84, 2015.

[144] M. J. Walker and G. Varani, "Design of rna-targeting macrocyclic peptides," *Methods Enzymol*, vol. 623, pp. 339–372, 2019.

[145] H. Chong, Y. Zhu, D. Yu, and Y. He, "Structural and functional characterization of membrane fusion inhibitors with extremely potent activity against human immunodeficiency virus type 1 (hiv-1), hiv-2, and simian immunodeficiency virus," *J Virol*, vol. 92, no. 20, 2018.

[146] A. Rezhdo, M. Islam, M. Huang, and J. A. Van Deventer, "Future prospects for noncanonical amino acids in biological therapeutics," *Curr Opin Biotechnol*, vol. 60, pp. 168–178, 2019. 1879-0429 Rezhdo, Arlinda Islam, Mariha Huang, Manjie Van Deventer, James A R21 CA214239/CA/NCI NIH HHS/United States Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Review England 2019/04/12 Curr Opin Biotechnol. 2019 Dec;60:168-178. doi: 10.1016/j.copbio.2019.02.020. Epub 2019 Apr 8.

[147] A. M. Saleh, K. M. Wilding, S. Calve, B. C. Bundy, and T. L. Kinzer-Ursem, "Noncanonical amino acid labeling in proteomics and biotechnology," *Journal of Biological Engineering*, vol. 13, no. 1, p. 43, 2019.

[148] E. Donsky and H. J. Wolfson, "Pepcrawler: a fast rrt-based algorithm for high-resolution refinement and binding affinity estimation of peptide inhibitors," *Bioinformatics*, vol. 27, no. 20, pp. 2836–2842, 2011.

[149] I. Antes, "Dynadock: A new molecular dynamics-based algorithm for protein-peptide docking including receptor flexibility," *Proteins*, vol. 78, no. 5, pp. 1084–104, 2010. 1097-0134 Antes, Iris Evaluation Study Journal Article Research Support, Non-U.S. Gov't United States Proteins. 2010 Apr;78(5):1084-104. doi: 10.1002/prot.22629.

[150] B. Raveh, N. London, L. Zimmerman, and O. Schueler-Furman, "Rosetta flexpepdock ab-initio: Simultaneous folding, docking and refinement of peptides onto their receptors," *PLOS ONE*, vol. 6, no. 4, p. e18934, 2011.

[151] S. Lata, N. K. Mishra, and G. P. S. Raghava, "Antibp2: improved version of antibacterial peptide prediction," *BMC Bioinformatics*, vol. 11, no. 1, p. S19, 2010.

[152] I. Johansson-Åkhe, C. Mirabello, and B. Wallner, "Predicting protein-peptide interaction sites using distant protein complexes as structural templates," *Scientific Reports*, vol. 9, no. 1, p. 4267, 2019.

[153] J. M. Cunningham, G. Koytiger, P. K. Sorger, and M. AlQuraishi, "Biophysical prediction of protein–peptide interactions and signaling networks using machine learning," *Nature Methods*, vol. 17, no. 2, pp. 175–183, 2020.

[154] Q. S. Du, Y. Ma, N. Z. Xie, and R. B. Huang, "Two-level qsar network (2l-qsar) for peptide inhibitor design based on amino acid properties and sequence positions," *SAR QSAR Environ Res*, vol. 25, no. 10, pp. 837–51, 2014. 1029-046x Du, Q S Ma, Y Xie, N Z Huang, R B Journal Article Research Support, Non-U.S. Gov't England SAR QSAR Environ Res. 2014;25(10):837-51. doi: 10.1080/1062936X.2014.959049. Epub 2014 Oct 2.

[155] M. Garton, C. Corbi-Verge, Y. Hu, S. Nim, N. Tarasova, B. Sherborne, and P. M. Kim, "Rapid and accurate structure-based therapeutic peptide design using gpu accelerated thermodynamic integration," *Proteins: Structure, Function, and Bioinformatics*, vol. 87, no. 3, pp. 236–244, 2019.

[156] K. W. Kaufmann, G. H. Lemmon, S. L. DeLuca, J. H. Sheehan, and J. Meiler, "Practically useful: What the rosetta protein modeling suite can do for you," *Biochemistry*, vol. 49, no. 14, pp. 2987–2998, 2010. doi: 10.1021/bi902153g.

[157] A. Verrecchio, M. W. Germann, B. P. Schick, B. Kung, T. Twardowski, and J. D. San Antonio, "Design of peptides with high affinities for heparin and endothelial cell proteoglycans*," *Journal of Biological Chemistry*, vol. 275, no. 11, pp. 7701–7707, 2000.

[158] T. Lebar, D. Lainšček, E. Merljak, J. Aupič, and R. Jerala, "A tunable orthogonal coiled-coil interaction toolbox for engineering mammalian cells," *Nat Chem Biol*, vol. 16, no. 5, pp. 513–519, 2020.

[159] A. Chevalier, D. A. Silva, G. J. Rocklin, D. R. Hicks, R. Vergara, P. Murapa, S. M. Bernard, L. Zhang, K. H. Lam, G. Yao, C. D. Bahl, S. I. Miyashita, I. Goreshnik, J. T. Fuller, M. T. Koday, C. M. Jenkins, T. Colvin, L. Carter, A. Bohn, C. M. Bryan, D. A. Fernández-Velasco, L. Stewart, M. Dong, X. Huang, R. Jin, I. A. Wilson, D. H. Fuller, and D. Baker, "Massively parallel de novo protein design for targeted therapeutics," *Nature*, vol. 550, no. 7674, pp. 74–79, 2017.

[160] R. Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, and Q. Zhou, "Structural basis for the recognition of sars-cov-2 by full-length human ace2," *Science*, vol. 367, no. 6485, pp. 1444–1448, 2020.

[161] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.

[162] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[163] J. H. Friedman, "On bias, variance, 0/1—loss, and the curse-of-dimensionality," *Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 55–77, 1997.

[164] OpenAI, "Gpt-4 technical report," 2023.

[165] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[166] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[167] F. Wu and J. Xu, "Deep template-based protein structure prediction," *PLOS Computational Biology*, vol. 17, no. 5, p. e1008954, 2021.

[168] M. T. Muhammed and E. Aki-Yalcin, "Homology modeling in drug discovery: Overview, current applications, and future perspectives," *Chemical Biology amp; Drug Design*, vol. 93, no. 1, pp. 12–20, 2019.

[169] X. Li, Y. Li, T. Cheng, Z. Liu, and R. Wang, "Evaluation of the performance of four molecular docking programs on a diverse set of protein-ligand complexes," *Journal of Computational Chemistry*, vol. 31, no. 11, pp. 2109–2125, 2010.

[170] L. Heo and M. Feig, "High-accuracy protein structures by combining machine-learning with physics-based refinement," *Proteins*, vol. 88, no. 5, pp. 637–642, 2020.

[171] H.-H. Tsai, C.-J. Tsai, B. Ma, and R. Nussinov, "In silico protein design by combinatorial assembly of protein building blocks," *Protein Science*, vol. 13, no. 10, pp. 2753–2765, 2004.

[172] D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment," *Proteins: Structure, Function, and Bioinformatics*, vol. 23, no. 4, pp. 566–579, 1995.

[173] A. Roy, A. Kucukural, and Y. Zhang, "I-tasser: a unified platform for automated protein structure and function prediction," *Nature Protocols*, vol. 5, no. 4, pp. 725–738, 2010.

[174] M. Prokop, J. Damborský, and J. Koca, "Triton: in silico construction of protein mutants and prediction of their activities *," *Bioinformatics*, vol. 16, no. 9, pp. 845–846, 2000.

[175] D. Gilis and M. Rooman, "Popmusic, an algorithm for predicting protein mutant stability changes. application to prion proteins," *Protein Engineering, Design and Selection*, vol. 13, no. 12, pp. 849–856, 2000.

[176] C. Pasquier and S. Hamodrakas, "An hierarchical artificial neural network system for the classification of transmembrane proteins," *Protein Engineering, Design and Selection*, vol. 12, no. 8, pp. 631–634, 1999.

[177] J. S. Marvin, E. E. Corcoran, N. A. Hattangadi, J. V. Zhang, S. A. Gere, and H. W. Hellinga, "The rational design of allosteric interactions in a monomeric protein and its applications to the constructionofbiosensors," *Proceedings of the National Academy of Sciences*, vol. 94, no. 9, pp. 4366–4371, 1997.

[178] C. Zhang, S. Liu, Q. Zhu, and Y. Zhou, "A knowledge-based energy function for proteinligand, proteinprotein, and proteindna complexes," *Journal of Medicinal Chemistry*, vol. 48, no. 7, pp. 2325–2335, 2005.

[179] S. Lise, C. Archambeau, M. Pontil, and D. T. Jones, "Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods," *BMC Bioinformatics*, vol. 10, no. 1, p. 365, 2009.

[180] C. Geng, A. Vangone, G. E. Folkers, L. C. Xue, and A. Bonvin, "isee: Interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations," *Proteins*, vol. 87, no. 2, pp. 110–119, 2019.

[181] R. Nikam, A. Kulandaisamy, K. Harini, D. Sharma, and M. M. Gromiha, "Prothermdb: thermodynamic database for proteins and mutants revisited after 15 years," *Nucleic Acids Research*, vol. 49, no. D1, pp. D420–D424, 2021.

[182] L. Jia, R. Yarlagadda, and C. C. Reed, "Structure based thermostability prediction models for protein single point mutations with machine learning tools," *PLoS One*, vol. 10, no. 9, p. e0138022, 2015.

[183] H. Cao, J. Wang, L. He, Y. Qi, and J. Z. Zhang, "Deepddg: Predicting the stability change of protein point mutations using neural networks," *J Chem Inf Model*, vol. 59, no. 4, pp. 1508–1514, 2019.

[184] P. Huang, S. K. S. Chu, H. N. Frizzo, M. P. Connolly, R. W. Caster, and J. B. Siegel, "Evaluating protein engineering thermostability prediction tools using an independently generated dataset," *ACS Omega*, vol. 5, no. 12, pp. 6487–6493, 2020. PMID: 32258884.

[185] J. Wang, S. Lisanza, D. Juergens, D. Tischer, I. Anishchenko, M. Baek, J. L. Watson, J. H. Chun, L. F. Milles, J. Dauparas, M. Expòsit, W. Yang, A. Saragovi, S. Ovchinnikov, and D. Baker, "Deep learning methods for designing proteins scaffolding functional sites," *bioRxiv*, 2021.

[186] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker, "Robust deep learning–based protein sequence design using proteinmpnn," *Science*, vol. 378, no. 6615, pp. 49–56, 2022.

[187] Z. Harteveld, J. Bonet, S. Rosset, C. Yang, F. Sesterhenn, and B. E. Correia, "A generic framework for hierarchical lt;emgt;de novolt;/emgt; protein design," *bioRxiv*, p. 2022.04.07.487481, 2022.

[188] Z. Cang and G.-W. Wei, "Topologynet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions," *PLOS Computational Biology*, vol. 13, no. 7, p. e1005690, 2017.

[189] L. Moffat, S. M. Kandathil, and D. T. Jones, "Design in the dark: Learning deep generative models for de novo protein design," *bioRxiv*, p. 2022.01.27.478087, 2022.

[190] M. Chen and S. J. Ludtke, "Deep learning-based mixed-dimensional gaussian mixture model for characterizing variability in cryo-em," *Nat Methods*, vol. 18, no. 8, pp. 930–936, 2021.

[191] D. A. Liberles, S. A. Teichmann, I. Bahar, U. Bastolla, J. Bloom, E. Bornberg-Bauer, L. J. Colwell, A. P. J. de Koning, N. V. Dokholyan, J. Echave, A. Elofsson, D. L. Gerloff, R. A. Goldstein, J. A. Grahnen, M. T. Holder, C. Lakner, N. Lartillot, S. C. Lovell, G. Naylor, T. Perica, D. D. Pollock, T. Pupko, L. Regan, A. Roger, N. Rubinstein, E. Shakhnovich, K. Sjölander, S. Sunyaev, A. I. Teufel, J. L. Thorne, J. W. Thornton, D. M. Weinreich, and S. Whelan, "The interface of protein structure, protein biophysics, and molecular evolution," *Protein Science*, vol. 21, no. 6, pp. 769–785, 2012.

[192] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[193] T. Soni Madhulatha, "An overview on clustering methods," *IOSR journal of engineering.*, vol. 4, no. 2, pp. 719–725, 2012.

[194] C. C. David, E. R. A. Singam, and D. J. Jacobs, "Jed: a java essential dynamics program for comparative analysis of protein trajectories," *BMC bioinformatics*, vol. 18, no. 1, pp. 271–271, 2017.

[195] L. Skjaerven, A. Martinez, and N. Reuter, "Principal component and normal mode analysis of proteins; a quantitative comparison using the groel subunit," *Proteins: Structure, Function, and Bioinformatics*, vol. 79, no. 1, pp. 232–243, 2011.

[196] A. Amadei, A. B. Linssen, and H. J. Berendsen, "Essential dynamics of proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 17, no. 4, pp. 412–425, 1993.

[197] O. F. Lange and H. Grubmüller, "Can principal components yield a dimension reduced description of protein dynamics on long time scales?," *The Journal of Physical Chemistry B*, vol. 110, no. 45, pp. 22842–22852, 2006.

[198] J.-h. Peng, W. Wang, Y.-q. Yu, H.-l. Gu, and X. Huang, "Clustering algorithms to analyze molecular dynamics simulation trajectories for complex chemical and biological systems," *Chinese Journal of Chemical Physics*, vol. 31, no. 4, pp. 404–420, 2018.

[199] K. Lindorff-Larsen and J. Ferkinghoff-Borg, "Similarity measures for protein ensembles," *PloS one*, vol. 4, no. 1, 2009.

[200] M. D. Sangid, “Coupling in situ experiments and modeling – opportunities for data fusion, machine learning, and discovery of emergent behavior,” *Current Opinion in Solid State and Materials Science*, vol. 24, no. 1, pp. 786–797, 2020.

[201] G. S. Heck, V. O. Pintro, R. R. Pereira, B. d. æ, M. vila, N. M.B. Levin, and W. F. de Azevedo, “Supervised machine learning methods applied to predict ligand- binding affinity,” *Current Medicinal Chemistry*, vol. 24, no. 23, pp. 2459–2470, 2017.

[202] T. Hastie, R. Tibshirani, and J. Friedman, *Overview of Supervised Learning*, pp. 9–41. Springer New York, 2009.

[203] C. Shen, J. Ding, Z. Wang, D. Cao, X. Ding, and T. Hou, “From machine learning to deep learning: Advances in scoring functions for protein–ligand docking,” *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 10, no. 1, p. e1429, 2020.

[204] A. Lavecchia, “Machine-learning approaches in drug discovery: methods and applications,” *Drug discovery today*, vol. 20, no. 3, pp. 318–331, 2015.

[205] Y.-C. Lo, S. E. Rensi, W. Torng, and R. B. Altman, “Machine learning in chemoinformatics and drug discovery,” *Drug discovery today*, vol. 23, no. 8, pp. 1538–1546, 2018.

[206] L. Zhang, J. Tan, D. Han, and H. Zhu, “From machine learning to deep learning: progress in machine intelligence for rational drug discovery,” *Drug discovery today*, vol. 22, no. 11, pp. 1680–1685, 2017.

[207] F. Ghasemi, A. Mehridehnavi, A. Pérez-Garrido, and H. Pérez-Sánchez, “Neural network and deep-learning algorithms used in qsar studies: merits and drawbacks,” *Drug discovery today*, vol. 23, no. 10, pp. 1784–1790, 2018.

[208] A. S. Rifaioglu, H. Atas, M. J. Martin, R. Cetin-Atalay, V. Atalay, and T. Doğan, “Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases,” *Briefings in bioinformatics*, vol. 20, no. 5, pp. 1878–1912, 2019.

[209] Y. Jing, Y. Bian, Z. Hu, L. Wang, and X.-Q. S. Xie, “Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era,” *The AAPS journal*, vol. 20, no. 3, pp. 1–10, 2018.

[210] D. Dana, S. V. Gadhiya, L. G. St. Surin, D. Li, F. Naaz, Q. Ali, L. Paka, M. A. Yamin, M. Narayan, I. D. Goldberg, *et al.*, “Deep learning in drug discovery and medicine; scratching the surface,” *Molecules*, vol. 23, no. 9, p. 2384, 2018.

[211] V. D. Mouchlis, A. Afantitis, A. Serra, M. Fratello, A. G. Papadiamantis, V. Aidinis, I. Lynch, D. Greco, and G. Melagraki, “Advances in de novo drug design: From conventional to machine learning methods,” *International journal of molecular sciences*, vol. 22, no. 4, p. 1676, 2021.

[212] J. Peña-Guerrero, P. A. Nguewa, and A. T. García-Sosa, "Machine learning, artificial intelligence, and data science breaking into drug design and neglected diseases," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 11, no. 5, p. e1513, 2021.

[213] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[214] S. Dhingra, R. Sowdhamini, F. Cadet, and B. Offmann, "A glance into the evolution of template-free protein structure prediction methodologies," *Biochimie*, vol. 175, pp. 85–92, 2020.

[215] C. Bystroff and D. Baker, "Prediction of local structure in proteins using a library of sequence-structure motifs," *J Mol Biol*, vol. 281, no. 3, pp. 565–77, 1998.

[216] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker, "Protein structure prediction using rosetta," *Methods Enzymol*, vol. 383, pp. 66–93, 2004.

[217] F. Noé, G. De Fabritiis, and C. Clementi, "Machine learning for protein folding and dynamics," *Current Opinion in Structural Biology*, vol. 60, pp. 77–84, 2020.

[218] S. Zhu, A. Shala, A. Bezginov, A. Sljoka, G. Audette, and D. J. Wilson, "Hyperphosphorylation of intrinsically disordered tau protein induces an amyloidogenic shift in its conformational ensemble," *PLOS ONE*, vol. 10, no. 3, p. e0120416, 2015.

[219] G. Sliwoski, S. Kothiwale, J. Meiler, and E. W. Lowe, "Computational methods in drug discovery," *Pharmacological reviews*, vol. 66, no. 1, pp. 334–395, 2014.

[220] S. P. Leelananda and S. Lindert, "Computational methods in drug discovery," *Beilstein journal of organic chemistry*, vol. 12, no. 1, pp. 2694–2718, 2016.

[221] D. B. Kokh, T. Kaufmann, B. Kister, and R. C. Wade, "Machine learning analysis of tauramd trajectories to decipher molecular determinants of drug-target residence times," *Front Mol Biosci*, vol. 6, p. 36, 2019.

[222] A. N. Lima, E. A. Philot, G. H. G. Trossini, L. P. B. Scott, V. G. Maltarollo, and K. M. Honorio, "Use of machine learning approaches for novel drug discovery," *Expert Opinion on Drug Discovery*, vol. 11, no. 3, pp. 225–239, 2016.

[223] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[224] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult, “Critical assessment of methods of protein structure prediction (casp)-round xiii,” *Proteins*, vol. 87, no. 12, pp. 1011–1020, 2019.

[225] A. R. Fersht, “Alphafold - a personal perspective on the impact of machine learning,” *J Mol Biol*, vol. 433, no. 20, p. 167088, 2021.

[226] M. AlQuraishi, “Machine learning in protein structure prediction,” *Curr Opin Chem Biol*, vol. 65, pp. 1–8, 2021.

[227] M. Torrisi, G. Pollastri, and Q. Le, “Deep learning methods in protein structure prediction,” *Comput Struct Biotechnol J*, vol. 18, pp. 1301–1310, 2020.

[228] S. H. P. De Oliveira, J. Shi, and C. M. Deane, “Comparing co-evolution methods and their application to template-free protein structure prediction,” *Bioinformatics*, p. btw618, 2016.

[229] 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.

[230] G. Ahdritz, N. Bouatta, S. Kadyan, Q. Xia, W. Gerecke, and M. AlQuraishi, “OpenFold,” 11 2021.

[231] R. Wu, F. Ding, R. Wang, R. Shen, X. Zhang, S. Luo, C. Su, Z. Wu, Q. Xie, B. Berger, J. Ma, and J. Peng, “High-resolution <i>de novo</i> structure prediction from primary sequence.” 2022.

[232] H. Kamisetty, S. Ovchinnikov, and D. Baker, “Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 39, pp. 15674–15679, 2013.

[233] B. I. M. Wicky, L. F. Milles, A. Courbet, R. J. Ragotte, J. Dauparas, E. Kinfu, S. Tipps, R. D. Kibler, M. Baek, F. DiMaio, X. Li, L. Carter, A. Kang, H. Nguyen, A. K. Bera, and D. Baker, “Hallucinating symmetric protein assemblies,” *Science*, vol. 378, no. 6615, pp. 56–61, 2022.

[234] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023.

[235] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, and N. Naik, “Large language models generate functional protein sequences across diverse families,” *Nature Biotechnology*, vol. 41, no. 8, pp. 1099–1106, 2023.

[236] C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, and A. Rives, “Learning inverse folding from millions of predicted structures,” *bioRxiv*, p. 2022.04.10.487779, 2022.

[237] B. Hie, S. Candido, Z. Lin, O. Kabeli, R. Rao, N. Smetanin, T. Sercu, and A. Rives, “A high-level programming language for generative protein design,” *bioRxiv*, p. 2022.12.21.521526, 2022.

[238] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker, “Robust deep learning based protein sequence design using proteinmpnn,” *bioRxiv*, p. 2022.06.03.494563, 2022.

[239] C. Norn, B. I. M. Wicky, D. Juergens, S. Liu, D. Kim, D. Tischer, B. Koepnick, I. Anishchenko, F. Players, D. Baker, S. Ovchinnikov, A. Coral, A. J. Bubar, A. Boykov, A. U. Valle Pérez, A. MacMillan, A. Lubow, A. Mussini, A. Cai, A. J. Ardill, A. Seal, A. Kalantarian, B. Failer, B. Lackersteen, B. Chagot, B. R. Haight, B. Taştan, B. Uitham, B. G. Roy, B. R. de Melo Cruz, B. Echols, B. E. Lorenz, B. Blair, B. Kestemont, C. D. Eastlake, C. J. Bragdon, C. Vardeman, C. Salerno, C. Comisky, C. L. Hayman, C. R. Landers, C. Zimov, C. D. Coleman, C. R. Painter, C. Ince, C. Lynagh, D. Malaniia, D. C. Wheeler, D. Robertson, V. Simon, E. Chisari, E. L. J. Kai, F. Rezae, F. Lengyel, F. Tabotta, F. Padelletti, F. Boström, G. O. Gross, G. McIlvaine, G. Beecher, G. T. Hansen, G. de Jong, H. Feldmann, J. L. Borman, J. Quinn, J. Norrgard, J. Truong, J. A. Diderich, J. M. Canfield, J. Photakis, J. D. Slone, J. Madzio, J. Mitchell, J. C. Stomieroski, J. H. Mitch, J. R. Altenbeck, J. Schinkler, J. B. Weinberg, J. D. Burbach, J. C. Sequeira da Costa, J. F. Bada Juarez, J. P. Gunnarsson, K. D. Harper, K. Joo, K. T. Clayton, K. E. DeFord, K. F. Scully, K. M. Gildea, K. J. Abbey, K. L. Kohli, K. Stenner, K. Takács, L. L. Poussaint, L. C. Manalo, L. C. Withers, L. Carlson, L. Wei, L. R. Fisher, L. Carpenter, M. Ji-hwan, *et al.*, “Protein sequence design by conformational landscape optimization,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 11, p. e2017228118, 2021.

[240] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. D. Bortoli, E. Mathieu, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker, “Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models,” *bioRxiv*, p. 2022.12.09.519842, 2022.

[241] K. A. Khan, S. A. Memon, and H. Naveed, “A hierarchical deep learning based approach for multi-functional enzyme classification,” *Protein Science*, vol. 30, no. 9, pp. 1935–1945, 2021.

[242] P. Sormanni, F. A. Aprile, and M. Vendruscolo, “Rational design of antibodies targeting specific epitopes within intrinsically disordered proteins,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 32, pp. 9902–9907, 2015.

[243] J. Zalevsky, A. K. Chamberlain, H. M. Horton, S. Karki, I. W. Leung, T. J. Sproule, G. A. Lazar, D. C. Roopenian, and J. R. Desjarlais, “Enhanced antibody half-life improves in vivo activity,” *Nat Biotechnol*, vol. 28, no. 2, pp. 157–9, 2010.

[244] G. G. Wood, D. A. Clinkenbeard, and D. J. Jacobs, “Nonadditivity in the alpha-helix to coil transition,” *Biopolymers*, vol. 95, no. 4, pp. 240–253, 2011.

[245] B. Y. Khor, G. J. Tye, T. S. Lim, and Y. S. Choong, “General overview on structure prediction of twilight-zone proteins,” *Theoretical Biology and Medical Modelling*, vol. 12, no. 1, p. 15, 2015.

[246] B. Rost, “Twilight zone of protein sequence alignments,” *Protein Engineering, Design and Selection*, vol. 12, no. 2, pp. 85–94, 1999.

[247] S. Y. Chung and S. Subbiah, “A structural explanation for the twilight zone of protein sequence homology,” *Structure*, vol. 4, no. 10, pp. 1123–1127, 1996.

[248] C. Neylon, “Chemical and biochemical strategies for the randomization of protein encoding dna sequences: library construction methods for directed evolution,” *Nucleic Acids Res*, vol. 32, no. 4, pp. 1448–59, 2004.

[249] R. I. Corona and J. Guo, “Statistical analysis of structural determinants for protein–dna-binding specificity,” *Proteins: Structure, Function, and Bioinformatics*, vol. 84, no. 8, pp. 1147–1161, 2016.

[250] R. Nair and B. Rost, “Sequence conserved for subcellular localization,” *Protein Science*, vol. 11, no. 12, pp. 2836–2847, 2009.

[251] O. Schon, A. Friedler, S. Freund, and A. R. Fersht, “Binding of p53-derived ligands to mdm2 induces a variety of long range conformational changes,” *Journal of Molecular Biology*, vol. 336, no. 1, pp. 197–202, 2004.

[252] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon, “Biopython: freely available python tools for computational molecular biology and bioinformatics,” *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.

[253] N. S. Pagadala, K. Syed, and J. Tuszynski, “Software for molecular docking: a review,” *Biophysical Reviews*, vol. 9, no. 2, pp. 91–102, 2017.

[254] E. Guida, A. Bisso, C. Fenollar-Ferrer, M. Napoli, C. Anselmi, J. E. Girardini, P. Carloni, and G. Del Sal, “Peptide aptamers targeting mutant p53 induce apoptosis in tumor cells,” *Cancer Research*, vol. 68, no. 16, p. 6550, 2008.

[255] C. C. Chao, "Mechanisms of p53 degradation," *Clin Chim Acta*, vol. 438, pp. 139–47, 2015.

[256] K. D. Sullivan, M. D. Galbraith, Z. Andrysik, and J. M. Espinosa, "Mechanisms of transcriptional regulation by p53," *Cell Death Differentiation*, vol. 25, no. 1, pp. 133–143, 2018.

[257] K. H. Vousden and D. P. Lane, "p53 in health and disease," *Nature Reviews Molecular Cell Biology*, vol. 8, no. 4, pp. 275–283, 2007.

[258] N. Rasafar, A. Barzegar, and E. Mehdizadeh Aghdam, "Structure-based designing efficient peptides based on p53 binding site residues to disrupt p53-mdm2/x interaction," *Scientific Reports*, vol. 10, no. 1, 2020.

[259] P. May and E. May, "Twenty years of p53 research: structural and functional aspects of the p53 protein," *Oncogene*, vol. 18, no. 53, pp. 7621–7636, 1999.

[260] M. Wells, H. Tidow, T. J. Rutherford, P. Markwick, M. R. Jensen, E. Mylonas, D. I. Svergun, M. Blackledge, and A. R. Fersht, "Structure of tumor suppressor p53 and its intrinsically disordered n-terminal transactivation domain," *Proceedings of the National Academy of Sciences*, vol. 105, no. 15, pp. 5762–5767, 2008.

[261] H. D. Ou, F. Löhr, V. Vogel, W. Mäntele, and V. Dötsch, "Structural evolution of c-terminal domains in the p53 family," *The EMBO Journal*, vol. 26, no. 14, pp. 3463–3473, 2007. https://doi.org/10.1038/sj.emboj.7601764.

[262] Y. Jiang, J. Liu, D. Chen, L. Yan, and W. Zheng, "Sirtuin inhibition: Strategies, inhibitors, and therapeutic potential," *Trends in Pharmacological Sciences*, vol. 38, no. 5, pp. 459–472, 2017.

[263] B. J. Aubrey, G. L. Kelly, A. Janic, M. J. Herold, and A. Strasser, "How does p53 induce apoptosis and how does this relate to p53-mediated tumour suppression?," *Cell Death Differentiation*, vol. 25, no. 1, pp. 104–113, 2018.

[264] J. L. Avalos, I. Celic, S. Muhammad, M. S. Cosgrove, J. D. Boeke, and C. Wolberger, "Structure of a sir2 enzyme bound to an acetylated p53 peptide," *Molecular Cell*, vol. 10, no. 3, pp. 523–535, 2002.

[265] T. Nagata, K. Shirakawa, N. Kobayashi, H. Shiheido, N. Tabata, Y. Sakuma-Yonemura, K. Horisawa, M. Katahira, N. Doi, and H. Yanagawa, "Structural basis for inhibition of the mdm2:p53 interaction by an optimized mdm2-binding peptide selected with mrna display," *PloS one*, vol. 9, no. 10, pp. e109163–e109163, 2014.

[266] L. Chang and A. Perez, "Ranking peptide binders by affinity with alphafold**," *Angewandte Chemie International Edition*, vol. 62, no. 7, p. e202213362, 2023.

[267] T. Tsaban, J. K. Varga, O. Avraham, Z. Ben-Aharon, A. Khramushin, and O. Schueler-Furman, "Harnessing protein folding neural networks for peptide–protein docking," *Nature Communications*, vol. 13, no. 1, p. 176, 2022.

[268] E. F. McDonald, T. Jones, L. Plate, J. Meiler, and A. Gulsevin, "Benchmarking alphafold2 on peptide structure prediction," *Structure*, vol. 31, no. 1, pp. 111–119.e2, 2023.

[269] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," 2022.

[270] D. W. Farrell, K. Speranskiy, and M. F. Thorpe, "Generating stereochemically acceptable protein pathways," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 14, pp. 2908–2921, 2010.

[271] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, p. e2016239118, 2021.

[272] P. Kunzmann and K. Hamacher, "Biotite: a unifying open source computational biology framework in python," *BMC Bioinformatics*, vol. 19, no. 1, p. 346, 2018.

[273] M. J. Macias, M. Hyvönen, E. Baraldi, J. Schultz, M. Sudol, M. Saraste, and H. Oschkinat, "Structure of the ww domain of a kinase-associated protein complexed with a proline-rich peptide," *Nature*, vol. 382, no. 6592, pp. 646–9, 1996. Macias, M J Hyvönen, M Baraldi, E Schultz, J Sudol, M Saraste, M Oschkinat, H Journal Article Research Support, Non-U.S. Gov't England 1996/08/15 Nature. 1996 Aug 15;382(6592):646-9. doi: 10.1038/382646a0.

[274] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, "Colabfold: making protein folding accessible to all," *Nature Methods*, vol. 19, no. 6, pp. 679–682, 2022.

[275] K. E. Roberts, P. R. Cushing, P. Boisguerin, D. R. Madden, and B. R. Donald, "Computational design of a pdz domain peptide inhibitor that rescues cftr activity," *PLOS Computational Biology*, vol. 8, no. 4, p. e1002477, 2012.

[276] C. Andrews, Y. Xu, M. Kirberger, and J. J. Yang, "Structural aspects and prediction of calmodulin-binding proteins," *Int J Mol Sci*, vol. 22, no. 1, 2020. 1422-0067 Andrews, Corey Orcid: 0000-0002-2310-7658 Xu, Yiting Kirberger, Michael Yang, Jenny J Journal Article Review Switzerland 2021/01/06 Int J Mol Sci. 2020 Dec 30;22(1):308. doi: 10.3390/ijms22010308.

[277] N. Datta and P. Kontomichalou, “Penicillinase synthesis controlled by infectious r factors in enterobacteriaceae,” *Nature*, vol. 208, no. 5007, pp. 239–41, 1965. Datta, N Kontomichalou, P Journal Article England 1965/10/16 Nature. 1965 Oct 16;208(5007):239-41. doi: 10.1038/208239a0.

[278] C. Avery, L. Baker, and D. J. Jacobs, “Functional dynamics of substrate recognition in tem beta-lactamase,” *Entropy (Basel)*, vol. 24, no. 5, 2022. 1099-4300 Avery, Chris Baker, Lonnie Jacobs, Donald J Orcid: 0000-0001-7711-1639 Journal Article Switzerland 2022/05/29 Entropy (Basel). 2022 May 20;24(5):729. doi: 10.3390/e24050729.

[279] S. M. Q. Chee, J. Wongsantichon, J. Siau, D. Thean, F. Ferrer, R. C. Robinson, D. P. Lane, C. J. Brown, and F. J. Ghadessy, “Structure-activity studies of mdm2/mdm4-binding stapled peptides comprising non-natural amino acids,” *PLoS One*, vol. 12, no. 12, p. e0189379, 2017. 1932-6203 Chee, Sharon Min Qi Wongsantichon, Jantana Siau, Jiawei Thean, Dawn Ferrer, Fernando Robinson, Robert C Lane, David P Brown, Christopher J Ghadessy, Farid J Orcid: 0000-0001-8559-903x Journal Article United States 2017/12/12 PLoS One. 2017 Dec 11;12(12):e0189379. doi: 10.1371/journal.pone.0189379. eCollection 2017.

[280] G. Hinton, “The forward-forward algorithm: Some preliminary investigations,” 2022.

[281] S. H. Schneider, J. Kozuch, and S. G. Boxer, “The interplay of electrostatics and chemical positioning in the evolution of antibiotic resistance in tem -lactamases,” *ACS Central Science*, vol. 7, no. 12, pp. 1996–2008, 2021. doi: 10.1021/acscentsci.1c00880.

[282] G. Ahdritz, N. Bouatta, C. Floristean, S. Kadyan, Q. Xia, W. Gerecke, T. J. O’Donnell, D. Berenberg, I. Fisk, N. Zanichelli, B. Zhang, A. Nowaczynski, B. Wang, M. M. Stepniewska-Dziubinska, S. Zhang, A. Ojewole, M. E. Guney, S. Biderman, A. M. Watkins, S. Ra, P. R. Lorenzo, L. Nivon, B. Weitzner, Y.-E. A. Ban, P. K. Sorger, E. Mostaque, Z. Zhang, R. Bonneau, and M. AlQuraishi, “Openfold: Retraining alphafold2 yields new insights into its learning mechanisms and capacity for generalization,” *bioRxiv*, p. 2022.11.20.517210, 2023.

[283] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives, “Evolutionary-scale prediction of atomic-level protein structure with a language model,” *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[284] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, “Language models enable zero-shot prediction of the effects of mutations on protein function,” *bioRxiv*, p. 2021.07.09.450648, 2021.

[285] V. Thumuluri, H.-M. Martiny, J. J. Almagro Armenteros, J. Salomon, H. Nielsen, and A. R. Johansen, “Netsolp: predicting protein solubility in escherichia coli using language models,” *Bioinformatics*, vol. 38, no. 4, pp. 941–946, 2021.

[286] B. L. Hie, V. R. Shanker, D. Xu, T. U. J. Bruun, P. A. Weidenbacher, S. Tang, W. Wu, J. E. Pak, and P. S. Kim, "Efficient evolution of human antibodies from general protein language models," *Nature Biotechnology*, 2023.

[287] S. K. Lüdemann, V. Lounnas, and R. C. Wade, "How do substrates enter and products exit the buried active site of cytochrome p450cam? 2. steered molecular dynamics and adiabatic mapping of substrate pathways11edited by j. thornton," *Journal of Molecular Biology*, vol. 303, no. 5, pp. 813–830, 2000.

[288] J.-M. Davière and P. Achard, "A pivotal role of dellas in regulating multiple hormone signals," *Molecular Plant*, vol. 9, no. 1, pp. 10–20, 2016.

[289] K. Murase, Y. Hirano, T.-p. Sun, and T. Hakoshima, "Gibberellin-induced della recognition by the gibberellin receptor gid1," *Nature*, vol. 456, no. 7221, pp. 459–463, 2008.

[290] J. Hernandez-GarcÌa, A. Briones-Moreno, and M. Bl·zquez, "Origin and evolution of gibberellin signaling and metabolism in plants," *Seminars in cell developmental biology*, 2020.

[291] J. Peng, D. E. Richards, N. M. Hartley, G. P. Murphy, K. M. Devos, J. E. Flintham, J. Beales, L. J. Fish, A. J. Worland, F. Pelica, D. Sudhakar, P. Christou, J. W. Snape, M. D. Gale, and N. P. Harberd, "'green revolution' genes encode mutant gibberellin response modulators," *Nature*, vol. 400, no. 6741, pp. 256–261, 1999.

[292] Y. Yamamoto, T. Hirai, E. Yamamoto, M. Kawamura, T. Sato, H. Kitano, M. Matsuoka, and M. Ueguchi-Tanaka, "A rice gid1 suppressor mutant reveals that gibberellin is not always required for interaction between its receptor, gid1, and della proteins," *The Plant Cell*, vol. 22, no. 11, pp. 3589–3602, 2010.

[293] G. F. Hao, S. G. Yang, G. F. Yang, and C. G. Zhan, "Computational gibberellin-binding channel discovery unraveling the unexpected perception mechanism of hormone signal by gibberellin receptor," *J Comput Chem*, vol. 34, no. 24, pp. 2055–64, 2013.

[294] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang, "The i-tasser suite: protein structure and function prediction," *Nat Methods*, vol. 12, no. 1, pp. 7–8, 2015. 1548-7105 Yang, Jianyi Yan, Renxiang Roy, Ambrish Xu, Dong Poisson, Jonathan Zhang, Yang R01 GM083107/GM/NIGMS NIH HHS/United States R01 GM084222/GM/NIGMS NIH HHS/United States GM083107/GM/NIGMS NIH HHS/United States GM084222/GM/NIGMS NIH HHS/United States Letter Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S. Nat Methods. 2015 Jan;12(1):7-8. doi: 10.1038/nmeth.3213.

[295] L. Camut, T. Regnault, M. Sirlin-Josserand, L. Sakvarelidze-Achard, E. Carrera, J. Zumsteg, D. Heintz, N. Leonhardt, M. J. P. Lange, T. Lange, J.-M. Davière, and P. Achard, "Root-derived ga12 contributes to temperature-induced shoot growth in arabidopsis," *Nature Plants*, vol. 5, no. 12, pp. 1216–1221, 2019.

[296] A. S. Ettayapuram Ramaprasad, S. Uddin, J. Casas-Finet, and D. J. Jacobs, "Decomposing dynamical couplings in mutated scfv antibody fragments into stabilizing and destabilizing effects," *Journal of the American Chemical Society*, vol. 139, no. 48, pp. 17508–17517, 2017.

[297] K. A. Dill, "Additivity principles in biochemistry," *J Biol Chem*, vol. 272, no. 2, pp. 701–4, 1997. Dill, K A Journal Article Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. Review United States 1997/01/10 J Biol Chem. 1997 Jan 10;272(2):701-4. doi: 10.1074/jbc.272.2.701.

[298] A. E. Mark and W. F. van Gunsteren, "Decomposition of the free energy of a system in terms of specific interactions. implications for theoretical and experimental studies," *J Mol Biol*, vol. 240, no. 2, pp. 167–76, 1994.

[299] D. J. Jacobs and M. F. Thorpe, "Generic rigidity percolation in two dimensions," *Physical Review E*, vol. 53, no. 4, pp. 3682–3693, 1996. PRE.

[300] T. Li, D. Verma, M. B. Tracka, J. Casas-Finet, D. R. Livesay, and D. J. Jacobs, "Thermodynamic stability and flexibility characteristics of antibody fragment complexes," *Protein Pept Lett*, vol. 21, no. 8, pp. 752–65, 2014.

[301] D. R. Livesay and D. J. Jacobs, "Conserved quantitative stability/flexibility relationships (qsfr) in an orthologous rnase h pair," *Proteins: Structure, Function, and Bioinformatics*, vol. 62, no. 1, pp. 130–143, 2006.

[302] D. Verma, D. J. Jacobs, and D. R. Livesay, "Changes in lysozyme flexibility upon mutation are frequent, large and long-ranged," *PLOS Computational Biology*, vol. 8, no. 3, p. e1002409, 2012.

[303] C. Jackel, P. Kast, and D. Hilvert, "Protein design by directed evolution," *Annu Rev Biophys*, vol. 37, pp. 153–73, 2008.

[304] L. C. James and D. S. Tawfik, "Conformational diversity and protein evolution–a 60-year-old hypothesis revisited," *Trends Biochem Sci*, vol. 28, no. 7, pp. 361–8, 2003.

[305] M. E. Glasner, J. A. Gerlt, and P. C. Babbitt, "Mechanisms of protein evolution and their application to protein engineering," *Adv Enzymol Relat Areas Mol Biol*, vol. 75, pp. 193–239, xii–xiii, 2007.

[306] V. A. Risso, J. A. Gavira, D. F. Mejia-Carmona, E. A. Gaucher, and J. M. Sanchez-Ruiz, "Hyperstability and substrate promiscuity in laboratory resurrections of precambrian beta-lactamases," *J Am Chem Soc*, vol. 135, no. 8, pp. 2899–902, 2013. 1520-5126 Risso, Valeria A Gavira, Jose A Mejia-Carmona, Diego F Gaucher, Eric A Sanchez-Ruiz, Jose M Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. United States J Am Chem Soc. 2013 Feb 27;135(8):2899-902. doi: 10.1021/ja311630a. Epub 2013 Feb 14.

[307] D. Verma, D. J. Jacobs, and D. R. Livesay, “Variations within class-a -lactamase physiochemical properties reflect evolutionary and environmental patterns, but not antibiotic specificity,” *PLOS Computational Biology*, vol. 9, no. 7, p. e1003155, 2013.

[308] T. Grear, C. Avery, J. Patterson, and D. J. Jacobs, “Molecular function recognition by supervised projection pursuit machine learning,” *Sci Rep*, vol. 11, no. 1, p. 4247, 2021.

[309] H. Jeng-Neng, L. Shyh-Rong, M. Maechler, R. D. Martin, and J. Schimert, “Regression modeling in back-propagation and projection pursuit learning,” *IEEE Transactions on Neural Networks*, vol. 5, no. 3, pp. 342–353, 1994.

[310] Z. Ying and C. G. Atkeson, “Implementing projection pursuit learning,” *IEEE Transactions on Neural Networks*, vol. 7, no. 2, pp. 362–373, 1996.

[311] J. B. Kruskal, *Toward a Practical Method Which Helps Uncover the Structure of a Set Of Multivariate Observations By Finding the Linear Transformation Which Optimizes a New “Index Of Condensation”*, pp. 427–440. Academic Press, 1969.

[312] J. H. Friedman and J. W. Tukey, “A projection pursuit algorithm for exploratory data analysis,” *IEEE Transactions on Computers*, vol. C-23, no. 9, pp. 881–890, 1974.

[313] P. J. Huber, “Projection pursuit,” *The Annals of Statistics*, vol. 13, no. 2, pp. 435–475, 1985.

[314] N. d. Silva, D. Cook, and E.-K. Lee, “A projection pursuit forest algorithm for supervised classification,” *Journal of Computational and Graphical Statistics*, pp. 1–13, 2021.

[315] J. Patterson, C. Avery, T. Grear, and D. J. Jacobs, “Biased hypothesis formation from projection pursuit,” *Adv. Artif. Intell. Mach. Learn.*, vol. 1, no. 3, pp. 221–233, 2021.

[316] J. Farmer, F. Kanwal, N. Nikulsin, M. C. Tsilimigras, and D. J. Jacobs, “Statistical measures to quantify similarity between molecular dynamics simulation trajectories,” *Entropy*, vol. 19, no. 12, pp. 653–646, 2017.

[317] A. Egorov, M. Rubtsova, V. Grigorenko, I. Uporov, and A. Veselovsky, “The role of the ω-loop in regulation of the catalytic activity of tem-type β-lactamases,” *Biomolecules*, vol. 9, no. 12, pp. 843–854, 2019.

APPENDIX A: Additional Computational Biology Contributions

A.1 Molecular Dynamics on GID1A to Investigate Ligand Binding

Using randomly accelerated molecular dynamics (RAMD) method simulations of the dynamics and characteristics of small molecules binding to a protein dimeric complex were investigated [43, 287].

Gibberellin (GA) is a prominent growth hormone in plants with over a hundred variants. This study reexamines the traditional classification of GA into active and inactive forms by probing the GA-GAI-GID1A complex and its binding mechanisms [288]. Here, GID1A stands for the Gibberellin Insensitive Dwarf 1 receptor, and GAI represents the DELLA family disordered signaling protein, highlighted by a highly conserved residue set that contributes to its Molecular Recognition Feature (MoRF) [289]. The study aimed to discern how the GID1A-GA-DELLA system reacts to varied GA variants and the role DELLA plays in these variants' binding. The GAI protein, one of five in Arabidopsis, functions as a growth-inhibiting regulatory protein within the gibberellin response pathway [290]. The renowned 'Green Revolution' in dwarfed high-yield varieties is attributed to DELLA family mutations, which disrupt the interaction between the GA receptor GA-Insensitive Dwarf1 (GID1) and a DELLA protein, making them "master regulators of growth and development" [291, 292]. A pivotal advance by Hao et al. previously revealed a channel in the GAI-GID1A complex, enabling GA to bind without opening the GID1A lid [293]. Using molecular modeling, Hao et al. revised the suggested capping mechanism by the GID1A N-terminal lid, uncovering a channel allowing GA's ingress and egress from the binding pocket. Examination of 12 GA variants and probes the impact of DELLA on the binding process. MD simulations combined with τRAMD, the dissociation pathways of these GA variants are analyzed, synthesizing a complete model of the GID1A-GA-DELLA and GID1A-DELLA complexes from the Arabidopsis Thaliana crystal structure. With over 1.6 terabytes of molecular dynamics trajectory data, a dual binding pathway narrative surfaces, suggesting a regulatory molecular mechanism akin to a transistor. The τRAMD method, implemented in the NAMD MD simulation software, calculates a ligand's residence time within a binding pocket and, in this context, explores the egression pathways GA adopts across various DELLA-GID1 con formations.

A.1.1 Methods

Starting from a crystal structure of A. thaliana GA receptor GID1A, together with a partially resolved N-terminal fragment of DELLA GAI (PDB Accession no: 2ZSH), the absence of large segments of the disordered N-terminal fragment prevents realistic simulation of this hormone receptor system. All missing segments were constructed using I-TASSER (Iterative Threading ASSEmbly) ab-initio and threading method software [294]. A predicted model was choosen and the DELLA GAI N-terminus was placed by a structural alignment to the original fragments of the DELLA subunit, the gDG system. Each GA(x) ligand was aligned to the GA3 ligand in the 2ZSH structure, before energy minimization for the GA(x)-GAI-GID1A and GA(x)-GID1A systems was performed. A diverse set of GA(x) that includes GA variants widely referred to as bioactive and inactive, respectively denoted as GA(a) and GA(i). For bioactive GA variants, we consider GA1, GA4, GA3, GA7, and the recently reported GA12-16ox [295]. For putative inactive GA variants, we consider GA8, GA9, GA12, GA34, GA4MeO and GA4-16/17ox. Among the inactive set, GA4MeO, GA4-16/17ox, and GA34 are all oxidation products of GA(a) and are of interest to understand the signaling control in this system.

The GA ligands were optimized in AMBER using semi-empirical bond charge corrections available inside the Ambertools software and acepype, then the ligands were aligned with GA3 of the 2ZSH crystal structure before minimizing energy of the system in NAMD (3.0). The LEaP Amber tool was used for building a rectangular simulation box, with a minimum gap of 10 Å. To neutralize the system, Cl and Na ions were added at the concentration of (NaCl 0.015 M). The protein was protonated in pymol using pH of 7.2 and dissolved into TIP3P water. In each system, solvent within 5 nm of the 2ZSH GA3 was kept and placed back into the structure before energy minimization, as defined in Kokh et al. using ambertools. Each gDG system was minimized with 500 steps of steepest descent, followed by conjugate gradient minimization using 1000 steps. Force constants were used to generate starting replicas (each 1 ns with force constants 50, 10, 5, and none using the SHAKE algorithm) all kept at 300K using a Langevin thermostat and constant pressure of 1 atm using a Berendsen barostate. A cutoff of 10 Å for nonbonded Columbic and van der Waals interactions is applied, while using periodic boundary conditions with particle mesh Ewald (PME) method for the long-range interactions. Each replica system (4 per ligand) was heated in NAMD-3 over 1 ns in steps of 10 K without restraints. The production run was performed using Langevin thermostat and Nose-Hoover Langevin pressure controls to maintain the system at 1 atm and 300 K. The final frame for each 40 ns was taken for starting points of a set of 10 RAMD simulations in NAMD-2.

Each set of RAMD was evaluated for residence time statistics using bootstrapping at 2000 sets. Using a 50% threshold for the standard deviations for the residence time of a system, systems not meeting this requirement were extended with 5 more RAMD simulations until this variance criteria was met. Systematic evaluation of force constants was performed to determine an optimal force of 20 kcal mol-1 Å-1 to allow egression to a threshold of 30 Å from the docked position in a computationally tractable time span. Similar force evaluation was performed for the egressions of the entire DELLA protein, finding a significantly higher force of 80 kcal mol-1 Å-1 was needed to reach an egression distance of 30 Å from the original position of DELLA chain, to break most intermolecular interactions between the proteins.

A.1.2 Results

The residence times from τRAMD were obtained using a previously reported protocol adapted to MATLAB 2020a using PDFEstimator. Visual Molecular Dynamics (VMD) analysis of trajectories was performed using in-house designed scripts in TCL. A cut-off distance of 3.5 Angstrom and an angle of 30°for hydrogen bond counting. Hydrophobic contacts were defined using known nonpolar residues and a contact cut–off distance of 3.5 Å for conservative estimates of contacts. These were tabulated into a text file for further processing. Raw counts of interactions to residues from a defined reference set were used as samples in statistical analysis. Significance of differences was selected from a rational p-value threshold of 0.05, Figure A.2. Recurrent residue sets were identified using the union of logical comparisons between lists of interaction residues across simulation sets. For recurrent residues in various comparisons, the active GA subsets were used to create a residue subset, which was then compared to inactive GA interaction lists.

Regardless of the GA(i) and GA(a) systems monitored, the previously identified H-bonding residues from Hao et al. are observed to be recurrent. We found that residue TYR247 interacts with GA(x) C3-OH, where SER116 and GLY115 wrap around the C7 carboxylate. The TYR31 side chain has interactions with OH groups in GA(x), which have this moiety around C13, and also interacts with all GA vicinal C16-C17 bonds, supporting pi-pi interactions that help align GA within the GID1A pocket for most GA ligands. It is notable that most hydrophobic contacts are adjacent to these residues, but they are additionally highly conserved across the phylogeny of GID1. The only recurrent residue with large discrepancy between the GA(a) and GA(i) subsets is ASN32. Unlike the pocket residue interactions, the GA(i) and GA(a) interacted

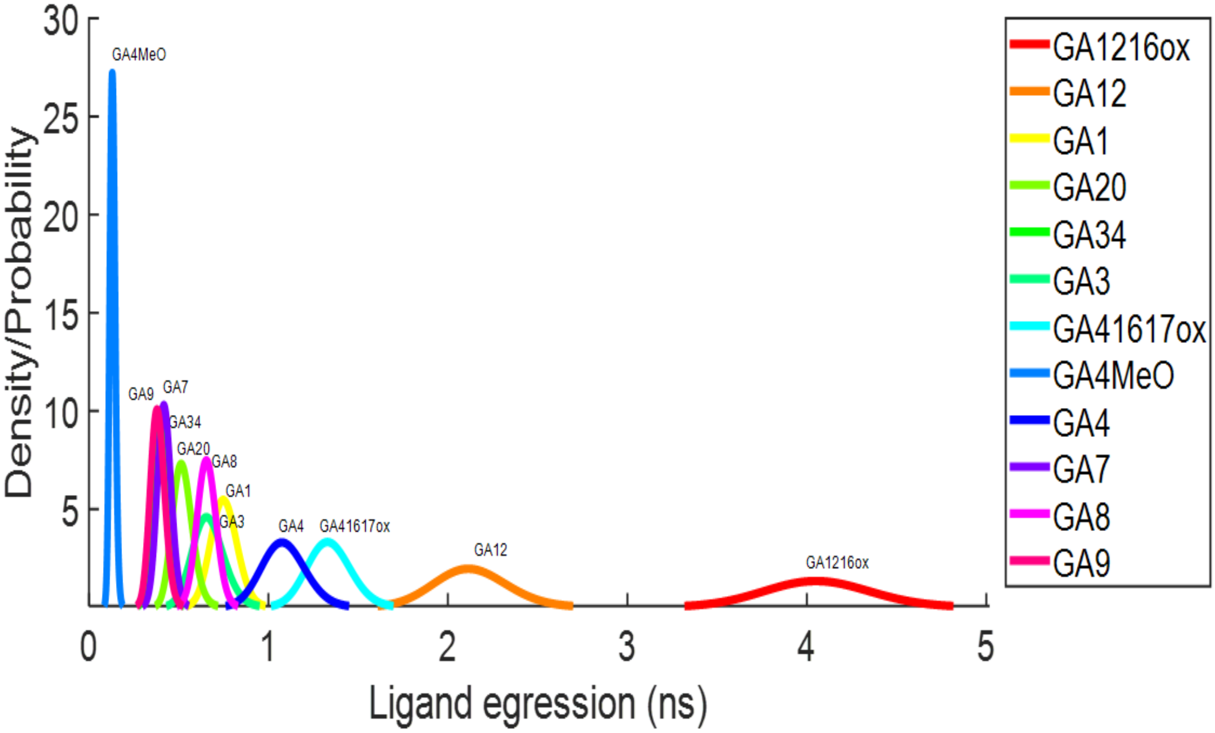


Figure A.1: Average residence time for each ligand in gDG. Values obtained from Bootstrapping the 40 egression times in each system then fitting the probability distribution.

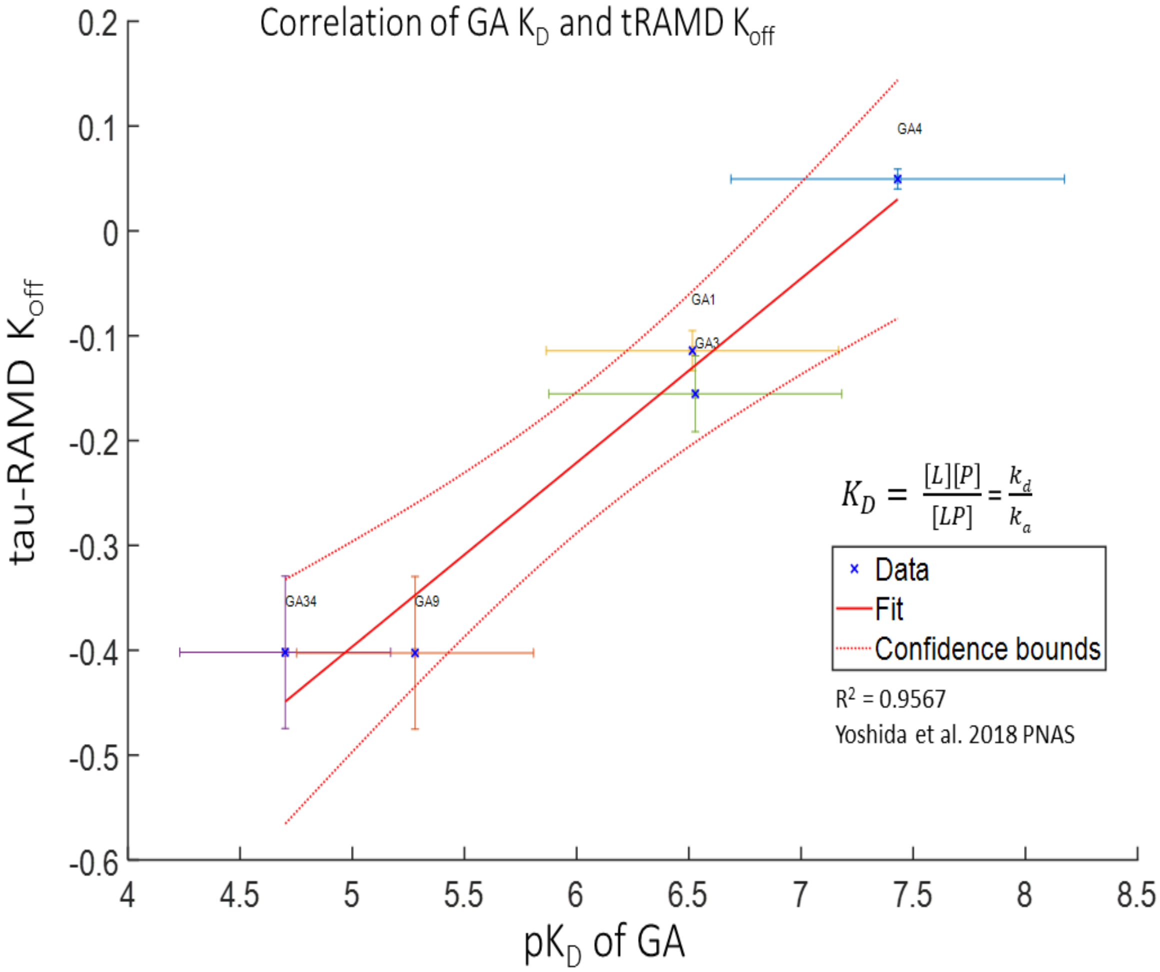


Figure A.2: K_D *(from Yoshida et. al 2018 PNAS)* plotted against the calculated k_{off} from τRAMD simulations of the various GA bound to GID1A.

similarly with channel residues along the egressions. In simulations without the GAI N-terminus (gG systems), GID1A was found to be flexible enough to allow egression from this same pathway without a large portion of the interactions reported previously. The second observed cleft pathway is shown in Figure A.3 in blue. Matching shape and electrostatic field maps give an indication for a binding pathway, which is a persistent signature in all simulations. Without the DELLA protein, this pathway is slightly shorter than the channel, even with dynamical fluctuations considered (6-8 Å for the cleft versus 9-12 Å for the channel). When the DELLA protein is bound to GID1A, the cleft opening is blocked. Nevertheless, observed in simulation the GA(x) can egress from the cleft pocket opening even in the presence of the DELLA protein, albeit in a heavily perturbed pathway.

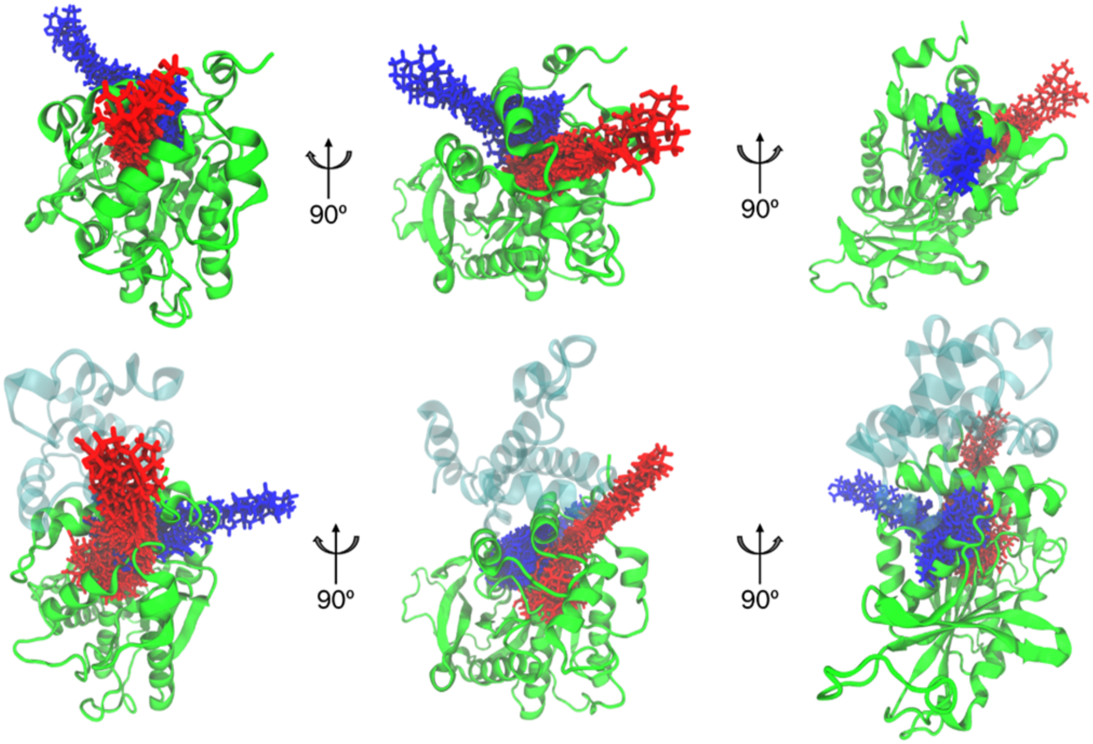


Figure A.3: Representations of pathways in GID1A system from simulated trajectories. Bottom DELLA in cyan and GID1A in green, top GID1A. The blue pathway is a newly discovered GID1 cleft, red is known channel pathway. Middle perspective looks down alpha helix B of GID1A with N-terminal to the right.

A dual-pathway binding mechanism in GID1A rectifies bioactive gibberellin forms, harmonizing with the myriad gibberellin variants observed in plants. These gibberellins, ranging from highly bioactive to inactive, facilitate the DELLA binding to the GA–GID1 complex. Upon DELLA binding, a selective channel pathway further promotes major bioactive GA forms within the binding pocket, stabilizing the entire complex. This three-step mechanism, underscored by the disordered DELLA signaling hub, maintains cellular homeostasis amidst various GA variants. Data-driven kinetic models illustrate how gibberellin fosters plant growth by amplifying the DELLA-GID1 binding, regulating DELLA degradation through two pathways: an open cleft pathway, accessible without DELLA, and a channel in GID1. Once GA binds GID1, both bioactive and less active GA forms bolster DELLA-GID1 binding. The distinctiveness of a bioactive GA is shaped by GID1 pathways and modulated by DELLA-GID1 interactions in the gDG complex, facilitated by conserved MoRFs and DELLA's flexible traits. If an inactive GA binds, a rectification process swaps it for a bioactive form, with rectification being essential if bioactive forms are in the minority. Given the binding time-scale differences between GA and DELLA, multiple GA exchanges can occur before complex dissociation, allowing minor shifts in bioactive GA concentrations to dictate plant growth, obviating the need for significant concentration differentials between active and inactive forms.

A.2 Other Physics Based Models of Proteins

As discussed, the physical mechanisms underlying protein function can be simulated computationally. An example of a direct application are mutation studies, where monitoring changes in dynamics of proteins as certain residues are mutated, usually on timescales of 1 μs or less, afford screening or comparative analysis [114]. Native state dynamics also stimulated interest in research dealing with dynamic allostery models [296]. These capabilities of simulation naturally lead to the development of models and algorithms for computational protein design and protein stability prediction.

While MD can simulate molecular physics at short time scales, it is not appropriate for calculating protein stability. Force–field approximated simulations are unable to properly generate ensembles large enough to calculate thermodynamic properties, without incredible resources or meta-dynamics techniques.

To avoid brute force MD simulation, thermodynamic models that assumed additivity in free-energy components were proposed to rapidly calculate thermodynamic quantities. Unfortunately, simple additive models failed miserably [297] due to the non-additivity of conformational entropy [298]. Free energy decomposition methods classically assume there is additivity in both enthalpy, like Hess's law, and entropy components. However, this assumption is only true for systems that can be divided into independent subsystems [297, 298]. The additivity assumption yields good approximate estimates for changes in entropy and free energy in small molecular systems with sterically limited conformational changes, but this breaks down for large polymer systems like proteins which have long range interactions.

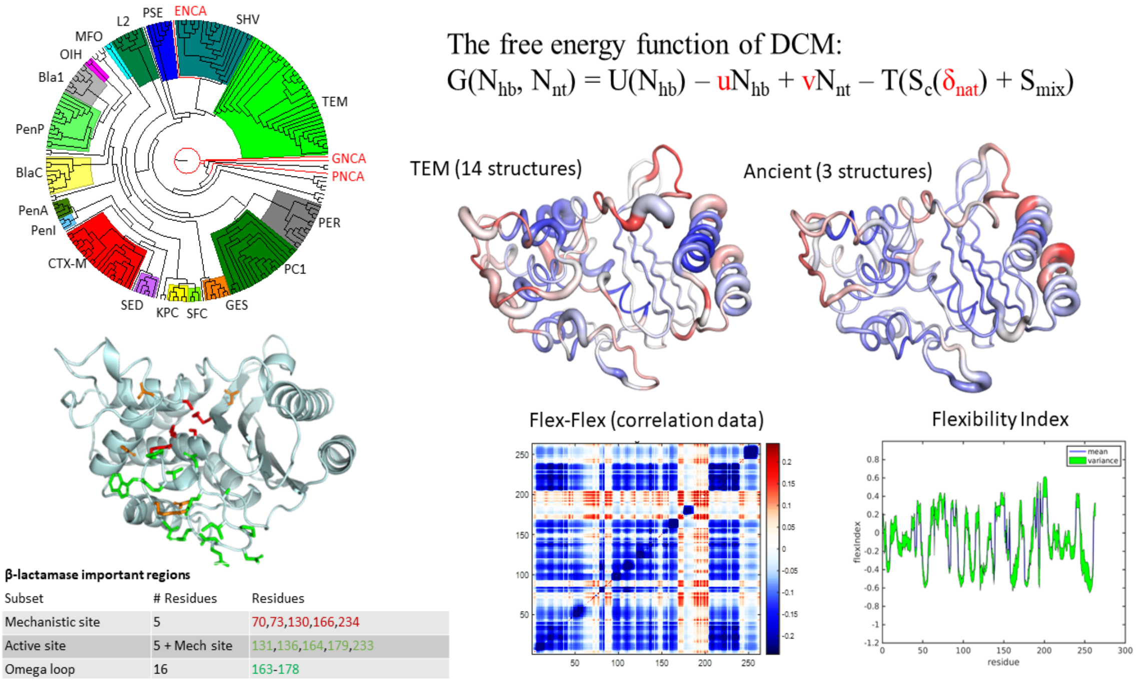
Using rigidity theory, the non-additivity found in conformational entropy can be accurately incorporated using a distance constraint model (DCM) [244, 299]. A protein conformation can be cast into a semi-empirical formula to determine the allowable space of flexible and inflexible components that allows the given conformation to unfold and refold on a simplistic free energy surface fitting to a landau function, Figure A.4. Previous work with DCM accurately described the heat capacity of proteins, including cold denaturation, because the model accounts for solvent effects [300–302]. A refined version called minimal DCM (mDCM) has been applied to elucidate the nuanced context-dependent flexibility in antibody structures and the significant changes in protein backbone flexibility and cooperativity due to point mutations, particularly evident in the comparison between human lysozyme and its hen egg white lysozyme ortholog [300, 302]. This method provides quantitative stability and flexibility relationships for protein [301], which has been useful for understanding protein evolution and helps with protein design [303–305].

It was hypothesized that this DCM model can be used in a protein classification problem. Beta-lactamase is a protein family responsible for antibiotic resistance in bacteria. The family is widely studied, understood, and still an applicable problem today. With an reasonable data and known classifications, this seemed like a reasonable family to investigate using a refined version of DCM, minimal DCM (mDCM).

A.2.1 Using mDCM to Classify Beta Lactamase Structures

In a comprehensive analysis of the structural nuances among class A beta-lactamase proteins, a key cause of bacterial multidrug antibiotic resistance, we have amassed a dataset of over 100 structures, encompassing not just present-day variants but also three ancestral reconstructions reflective of pre-hominid evolution untouched by contemporary antibiotic-driven selective pressures [306]. The intrigue lies in the paradox: why do members of this protein class, with their shared functional sites, global dynamics, and overarching structural motifs, exhibit such divergent patterns of antibiotic resistance? Our exploration factors in extant pharmacological data, shedding light on the binding intricacies between the enzymes and their diverse substrates. This diachronic collection provides an unparalleled window into the evolution of thermodynamic stability motifs and cooperativity utilizing the mDCM for these enzymes [307], we quantitatively charted stability/flexibility relationships across the β-lactamase lineage.

To performed the large calculations entire sub-spaces of variable combinations were samples in large embarrassingly parallelized computing by individually calculating unique variable combinations for u, v, and D_{nat}. A graphical example of the structures' QSFR is encapsulated in Figure A.4. This investigation, at its core, seeks to distill functional insights from homological diversity, offering a unique perspective the antibiotic resistance problem. Grouping these proteins by their local families, QSFR differences can be observed by mapping the quantitative rigidity values onto the three dimensional structures, as seem in Figure A.4, where blue shows rigidity and red shows flexibility. Notably the rigidity in the core of the Ancient structures while extant species, like TEM family, have a lack of rigidity. These findings suggest a dynamics based mechanism behind the ancestral reconstructs promiscuity towards all generations of antibiotics, making them extended spectrum beta-lactamses.



The free energy function of DCM:

$$G(N_{hb}, N_{nt}) = U(N_{hb}) - uN_{hb} + vN_{nt} - T(S_c(\delta_{nat}) + S_{mix})$$

β-lactamase important regions

Subset	# Residues	Residues
Mechanistic site	5	70,73,130,166,234
Active site	5 + Mech site	131,136,164,179,233
Omega loop	16	163-178

Figure A.4: Beta Lactamase proteins are a family of hydrolases responsible for a considerable percentage of multi-antibiotic resistance in bacteria (including penicillin). Flexibility index describes the back bone QSFR character while Flex-Flex represents probabilities of residue-residue couplings in a flexible or rigid manner. Our dataset includes ancestral reconstructions of pre-hominid β-lactamase structures2, and about 100 extant Class A β-lactamase structures.

The collected data was also used to distinguish between Bush-Jacobi functional Class 2b-Type and other variants (namely, 2a, 2e, 2f types) within our dataset Both Support Vector Machine (SVM), linear hyperplane separation maximizing class separation, and logistic regression models were experimented with. This classification was executed by juxtaposing the computed flex-index of the structures against a sequence alignment spanning the entire dataset. The extremely small size of these models and the limited data offer extremely fast training times. With K-Fold cross validation multiple replicates can be measured rapidly. One hundred replicated runs of the SVM model yielded a commendable mean accuracy of approximately 80%, accompanied by a 19% false positive rate, inferred from the receiver operating characteristic curves. On the other hand, logistic regression trailed slightly with a mean accuracy of 77% and a higher false positive rate of around 23%.

Despite the preparatory requirements for the mDCM calculation, it's salient to highlight that this approach potentially accelerates analysis relative to the time-intensive task of generating and scrutinizing multiple MD trajectories across all beta-lactamase structures. The data produced from mDCM gives atomistic level information via cooperativity and flexibility of the backbone, which could take 100 to 1000 times longer to collected using MD on modern compute. Encouragingly, our preliminary findings indicate a moderate level of success in classifying resistance profiles using QSFR data.

A.3 SPLOC

Supervised Projective Learning with Orthogonal Completeness (SPLOC) was initially developed for molecular function recognition as a projection pursuit (PP) optimized by an NN [308]. However, SPLOC transcends applications as a general purpose recurrent NN (RNN). a recurrent NN, SPLOC performs a data-driven process for binary discriminant analysis of data streams. which is a self-directed learning process that involves the formation and refinement of a working hypothesis as new data is presented to the machine.

The SPLOC-RNN depicted in Figure A.5 provides discriminant analysis and creates perception. The PP-based NN was shown to be an effective model [309, 310] in the 1990s. The process of collapsing high-dimensional data onto a line for a mode projection represents a tremendous loss of information in exchange for an immense gain in specificity. Although no information is lost when using a complete orthonormal basis, how information is distributed is not unique. Each mode is mapped to a perceptron which is governed by a rectifying adaptive nonlinear unit (RANU). For the m-th mode, when $S(m) > S_o$ this forms a data-driven hypothesis that there is a difference between functional and nonfunctional data streams. Conditional upon a statistically significant consensus, the proposed hypothesis is confirmed when the quality of clustering within the MFSP for a d-mode is positive

Here, several ML strategies are integrated with PP operating on data packets.

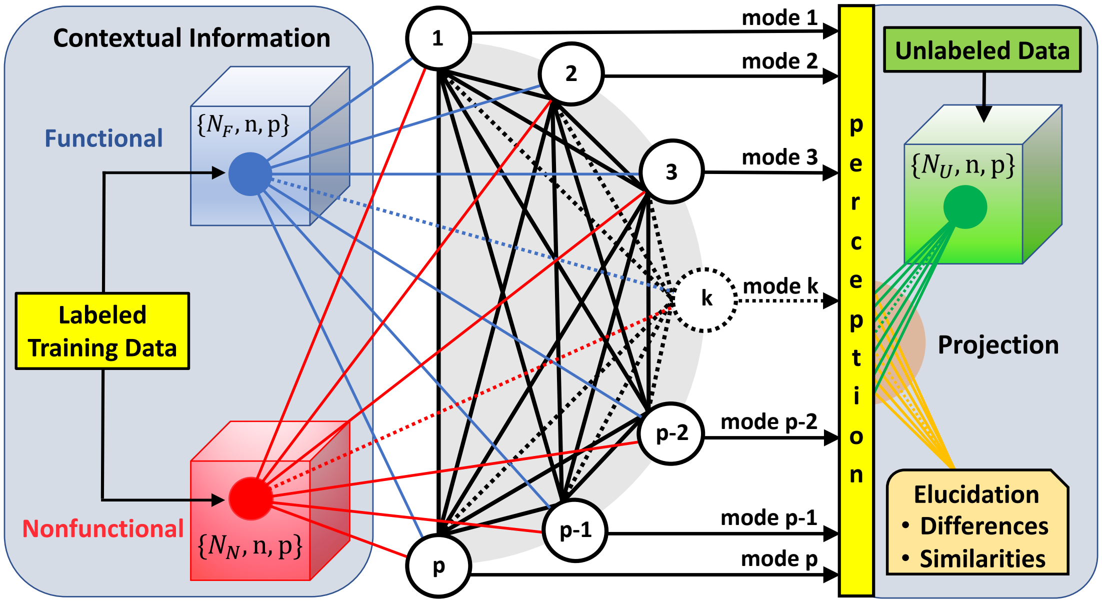


Figure A.5: Schematic of SPLOC as a recurrent neural network and data flow. For p variables there are p perceptrons, labeled from 1 to p, comprising the input layer that receives N_F functional and N_N nonfunctional data packets of n samples. The basis set is rotated in a search heuristic as the neural network evolves to maximize efficacy. Upon convergence, the output layer is composed of an orthonormal complete basis set.

Without required preprocessing of input data and void of hyperparameters, SPLOC-RNN performs derivative-free optimization within a nonparametric model on high dimensional data without limit on sample size. Furthermore, mitigation of overfitting to training data is an automated process that improves with greater observations per variable (OPV). For efficient hypothesis refinement, a discovery-likelihood is introduced using Bayesian inference for candidate ranking.

A.3.0.1 Architecture and Application

Unfortunately, results from a NN are often difficult to interpret; by extension, the underlying biases are difficult to characterize. In contrast, biases can be effectively controlled with an objective function within projection pursuit (PP) during the exploration of high-dimensional data [308, 311–314]. Moreover, PP is robust against statistical estimation errors [163], and perception is interpretable by linear projection operators that govern dimension reduction.

The objective function is net efficacy (E), which is the sum over the efficacy of each mode, $E(m)$. For all pairs of data streams, each mode is evaluated for (1) *selection* power, $S(m)$, that quantifies signal-to-noise; (2) *consensus* power, $C(m)$, that quantifies statistical significance, and (3) *quality* of clustering within a MFSP. The conditional selection power, $S(m|\alpha,\beta)$, associated with two data packets, α and β respectively representing functional and nonfunctional classes, is calculated for mode m using the formula:

$$S(m|\alpha,\beta) = \begin{cases} \sqrt{snr(m,\alpha,\beta)^2 + rex(m,\alpha,\beta)^2} + 1 & \text{if} < S_i \\ \sqrt{sbr(m,\alpha,\beta)^2 + rex(m,\alpha,\beta)^2} + 1 & \text{if} > S_d \\ S_o & \text{otherwise} \end{cases} \tag{A.1}$$

Where snr is the signal-to-noise-ratio given by:

$$snr(m,\alpha,\beta) = |\mu(m|\alpha) - \mu(m|\beta)| / \sqrt{\sigma(m|\alpha)^2 + \sigma(m|\beta)^2} \tag{A.2}$$

sbr is the signal-beyond-noise defined by:

$$sbr(m,\alpha,\beta) = \max\left[0, snr(m,\alpha,\beta) - 1\right] \tag{A.3}$$

rex is the excess ratio of standard deviations defined as:

$$rex(m, \alpha, \beta) = \max\left[\frac{\sigma(m|\alpha)}{\sigma(m|\beta)}, \frac{\sigma(m|\beta)}{\sigma(m|\alpha)}\right] - 1 \tag{A.4}$$

The values of S_i and S_d are respectively 1.3 and 2.0 for the upper threshold for indifference and the lower threshold for discrimination, and $S_o = 1.6125$ as their geometrical mean given as $S_o = \sqrt{S_i S_d}$. The collection of results for emergent properties of the data streams obtained for a specific mode, m, must be statistically consistent across all data stream pairs. The consensus measure $C(m)$ quantifies this consistency level. If different data stream pairs between functional and nonfunctional classes cannot reach a statistically significant consensus on whether there exists a difference or similarity between the two classes, then the information associated with that mode is set as undetermined.

SPLOC identifies the major differences between functional and nonfunctional molecules, using dynamics molecular toy datasets, and found that the shared conserved properties across all labeled molecules are markedly high. When strongly biased views either dismiss potential differences or similarities, SPLOC cannot validate or refute them as opposing data is overlooked due to the inherent bias. However, this latent information is retained in the u-modes that can be extracted to refine the working hypothesis once supporting data for the contrary view is taken into account [315].

A.3.1 SPLOC Applied to Beta Lactamase

Application using long MD simulation of TEM-1, TEM-2 and TEM-52 where the dynamics were analyzed to provide alignment over 263 residues involving 789 df. SPLOC was trained on TEM-1 as “functional” and TEM-52 as “nonfunctional”. The antibiotic extended spectrum resistant (ESR) TEM-52 and non-ESR TEM-1 represent the core difficulty with antibiotics that persist in human medicine. The utility of SPLOC-RNN is established by its ability to differentiate two closely related enzymes by functional dynamics, while classifying TEM-2 accurately. The problems with comparative analyses reported previously [316] entailing laborious effort are removed; replaced by an automated procedure.

The results suggest that the H10-H11 loop (residues 214-221) are a secondary anchor for larger extended spectrum ligands, while the H9-H10 loop (residues 194-202) is distal from the active site and stabilizes the protein against structural changes. These secondary non-catalytically-active loops offer attractive targets for novel noncompetitive inhibitors of TEM beta-lactamase. [278] This study also found residues known to play an important role in catalytic activity [317]. These results provide guidance in designing novel antibiotics to withstand mutation pathways in beta-lactamase that cause antibiotic resistance.