Advancing Peptide Therapeutics: A Generative AI-Driven Approach

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

Master of Science
in
Computational Natural Sciences by Research

by

Vishva Saravanan Ramasubramanian
2019113019
vishva.saravanan@research.iiit.ac.in

Center for Computational Natural Sciences and Bioinformatics
International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **"Advancing Peptide Therapeutics: A Generative AI-Driven Approach"** by **Vishva Saravanan Ramasubramanian**, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Dr. Bhaswar Ghosh

To my family and friends.

Acknowledgments

I would like to express my deepest gratitude to my research advisor, Dr. Bhaswar Ghosh, for his guidance and support. His mentorship has been invaluable to me and his expertise, patience, and encouragement have been instrumental in the successful completion of this thesis.

I am eternally indebted to my family for their unconditional love, encouragement, and sacrifices throughout my academic journey. In particular, I would like to thank my parents, grandparents and my sister, who have always been there for me to provide emotional support and wise counsel.

I would also like to acknowledge the exceptional support I have received from my closest friends. Anir, Arvindh, Ashu, Chandan, Dinesh, Jai, Jerrin, Jose, Lokesh, Manas, Rohan, Sanjai and Vijay in particular have been a constant source of encouragement and laughter throughout my academic journey. Their unwavering friendship has made this journey all the more enjoyable. A special acknowledgement is to one of my first friends here, Harshit, for also being the best roommate one could ask for.

Last but certainly not least, I extend my heartfelt gratitude to IIIT Hyderabad for providing me with an exceptional platform for my academic and research endeavors. The institute has opened doors to incredible professional opportunities and, most importantly, allowed me to cross paths with so many remarkable people. As the saying goes, “the hardest part of ending is starting again”; but I am eager to embark on the next chapter of my journey with the invaluable experiences and connections I gained here.

Abstract

The field of therapeutic peptide design is ripe for transformation, fueled by the convergence of biotechnology and artificial intelligence. Peptides, short chains of amino acids, offer a promising avenue for targeted drug therapies due to their inherent advantages over small molecules, including specificity and reduced side effects. However, the development of peptide therapeutics has been hindered by their limited oral bioavailability and susceptibility to enzymatic degradation. Recent advancements in deep learning techniques have opened new possibilities for addressing these challenges through innovative peptide design strategies.

This thesis explores the development of a novel hybrid deep learning framework for de novo peptide design. By harnessing the power of diffusion models, known for their ability to learn complex data distributions, and integrating them with binding affinity maximization algorithms, we have created a system capable of generating peptide sequences optimized for specific target receptors. To demonstrate the applicability of this framework, we focus on designing therapeutic peptides targeting proteins expressed by Plasmodium falciparum Erythrocyte Membrane Protein 1 (PfEMP1) genes, key contributors to malaria pathogenesis.

Our results highlight the potential of this hybrid deep learning approach to revolutionize peptide drug discovery. By generating peptide candidates conditioned on the binding sites of target receptors, we offer a promising avenue for developing effective therapies for malaria and other diseases. This research underscores the transformative power of AI in peptide therapeutics, paving the way for a new era of precision medicine with enhanced efficacy and reduced toxicity.

Contents

Chapter Page

1 Introduction: The Evolving Landscape of Drug Discovery . . . 1
1.1 Computational Approaches to Drug Design . . . 1
1.2 The Role of Generative Artificial Intelligence . . . 1
1.3 Peptide Therapeutics: Potential and Challenges . . . 2
1.4 Malaria: A Persistent Global Health Issue . . . 3

2 HYDRA: A Novel Framework for Receptor-Aware Peptide Design . . . 4
2.1 Problem Formulation and Objectives . . . 4
2.2 Constructing a Hybrid Diffusion Model for Peptide Generation . . . 4
2.2.1 Prior-Based Residue Count Estimation . . . 4
2.2.2 Residue Generation Process . . . 6
2.2.3 Peptide Reconstruction Strategy . . . 8
2.3 *In Silico* Evaluation of Designed Peptides . . . 11
2.3.1 Assessment of Physicochemical Properties . . . 11
2.3.2 Binding Affinity Prediction . . . 13
2.3.3 Diversity Analysis . . . 14

3 Experiments: Design and Implementation . . . 15
3.1 Dataset Curation for Training and Evaluation . . . 15
3.2 Model Architecture and Specifications . . . 16
3.3 Training Methodology . . . 17
3.4 Optimization of Evaluation Algorithm Parameters . . . 17
3.5 Application to PfEMP1 Protein Targets . . . 20

4 Results: Analysis and Discussion . . . 25
4.1 Comparative Analysis with Baseline Methods . . . 25
4.2 Case Study 1: PepBDB Protein Targets . . . 25
4.3 Case Study 2: PfEMP1 Protein Targets . . . 26

5 Conclusion and Future Directions . . . 33

Bibliography . . . 36

List of Figures

Figure Page

2.1 **HYDRA's two-stage peptide generation workflow**. **(A)** The diffusion process gradually injects noise into the data, and the generative process learns to recover the data distribution from the noise distribution by means of an SE(3)-equivariant network. **(B)** Upon generating amino acid residues, inter-residue center-of-mass distances are then computed, generating a distance matrix that encodes spatial relationships between them. Subsequently, a threshold is applied to this matrix to eliminate unrealistic connections, defining the feasible solution space for peptide conformations. A non-differentiable optimization algorithm then navigates this solution space. The objective function minimizes the binding affinity computed by AutoDock Vina, which corresponds to the predicted peptide conformation resulting in the most stable complex with the receptor. . . 5

2.2 Distribution of inter-residue distances in peptides in the training data. 99.35% of distances lie below 8. 9

3.1 Evolution of individual training loss components throughout the training process. To improve visualization and reduce noise, a Time-Weighted Exponential Moving Average (EMA) algorithm has been applied to smooth the curves. 18

3.2 Evolution of individual validation loss components throughout the training process. . . 19

3.3 A consolidated flowchart illustrating the binding site detection and *de novo* peptide binder generation pipeline for PfEMP1 targets. 21

3.4 Gene expression profile of the PfEMP1 family. The heatmap suggests that the proteins PF3D7_1200600, PF3D7_1150400, and PF3D7_0712400 are highly expressed. 21

3.5 Using CAVITY to identify strong and medium binding sites. Visualization of identified binding sites of (*A*) PF3D7_0712400, (*B*) PF3D7_1200600, and (*C*) PF3D7_1150400. The red and blue regions highlight the binding sites. 22

3.6 3D structural representations of the binding sites on PF3D7_0712400 interacting with generated peptides (**A**) RLKAVIP, (**B**) FSVYGIVH, and (**C**) IVVGPAYG. 23

3.7 A heatmap visualizing the predicted binding affinities between the chosen target binding sites and peptides generated by HYDRA. The color intensity represents the predicted binding affinity, with lower values indicating stronger binding. 24

4.1 **Comparison of binding affinity and peptide diversity between HYDRA and RFDiffusion.** Figures **(A)** and **(B)** depict the mean binding affinities and FRODOCK correlation scores, respectively, for peptides generated by each model. Lower binding affinities and higher FRODOCK scores indicate stronger predicted protein-peptide complexes. Binding targets are sorted based on the mean score for HYDRA-generated peptides. **(C)** represents the pairwise Tanimoto diversity scores across the generated peptide sets for each binding target. While the average diversity scores are similar, the distribution of Tanimoto scores implies greater consistency in HYDRA's peptide diversity compared to RFDiffusion. 27
4.2 Comparing the binding affinities of peptides generated by HYDRA and RFDiffusion. . 28
4.3 Comparing the FRODOCK scores of peptides generated by HYDRA and RFDiffusion. 29
4.4 **Comparison of physicochemical properties of peptides generated by HYDRA and RFDiffusion.** Each histogram represents the distribution of peptides across different ranges for a specific property. The y-axis indicates the percentage of peptides within each range. **(A)** Molecular Weight in Daltons, **(B)** Instability Index, **(C)** Aliphatic Index, **(D)** Isoelectric Point, **(E)** Half-life in hours, **(F)** GRAVY Index. Peptides designed by HYDRA show a more favorable bias in their property distribution compared to RFDiffusion. 30

List of Tables

Table Page

Chapter 1

Introduction: The Evolving Landscape of Drug Discovery

1.1 Computational Approaches to Drug Design

The landscape of drug discovery is being reshaped by computational drug design, an interdisciplinary field that harnesses computational models, simulations, and data analysis to streamline the identification and development of novel therapeutics [47]. Computational Drug Design encompasses a spectrum of techniques, from structure-based methods like molecular docking [31] and virtual screening [48] to ligand-based approaches like quantitative structure-activity relationship (QSAR) modeling [6] and machine learning algorithms [16].

Recent advances in artificial intelligence, particularly deep learning, have propelled computational drug design into a new era. Deep learning models, trained on massive datasets of molecular structures and bioactivity data, can predict molecular properties, binding affinities, and even generate novel drug-like molecules [62] These AI-powered tools are not only accelerating the traditionally time-consuming and costly drug discovery process but also expanding the druggable chemical space and enabling the design of personalized medicines [16]

1.2 The Role of Generative Artificial Intelligence

Generative Artificial Intelligence (AI) has emerged as a transformative force across numerous disciplines, revolutionizing how we approach complex problems and creative tasks. Recent advances in this field have led to unprecedented capabilities in generating human-like text, realistic images, and even molecular structures. One of the most significant breakthroughs in generative AI has been the development of large language models. These models, such as GPT-3 and its successors, have demonstrated remarkable abilities in natural language processing tasks, including text generation, translation, and question-answering [TODO: cite]. In the medical field, these models have been applied to tasks such as generating clinical notes and assisting in medical diagnosis [TODO: cite].

Computer vision has also benefited greatly from generative AI. Generative Adversarial Networks (GANs) have been used to create highly realistic images and videos, with applications ranging from art creation to enhancing medical imaging [TODO: cite]. In the field of radiology, for instance, GANs have been employed to generate synthetic medical images for training purposes, addressing the issue of limited data in rare medical conditions [TODO: cite]. In the field of materials science, generative models have been used to design new materials with specific properties. For example, researchers have used AI to generate hypothetical zeolite structures, potentially leading to the discovery of new catalysts for industrial processes [TODO: cite].

In the realm of drug discovery, generative AI has shown immense potential. Deep generative models are reshaping drug discovery processes, enabling the rapid design of novel molecules with desired properties [TODO: cite]. These models can generate potential drug candidates much faster than traditional methods, significantly accelerating the early stages of drug development.

1.3 Peptide Therapeutics: Potential and Challenges

Peptide therapeutics have garnered significant attention due to their unique advantages over small-molecule drugs. Peptides, composed of short chains of amino acids, exhibit high target specificity, minimal off-target effects, and reduced immunogenicity [54]. Their inherent biocompatibility and the ease of chemical modifications make them versatile tools for therapeutic interventions. Peptides often contain a mix of hydrophilic and hydrophobic amino acids, which can affect their solubility and interaction with cell membranes [35]. This can impact their ability to pass through cell membranes and work effectively inside cells. Peptides with a slightly higher percentage of hydrophobic residues may have increased cell permeability [35] and be more effective at interacting with cell membranes, as hydrophobic residues can interact very well with the hydrophobic regions of lipid bi-layers, enhancing the transit of the peptide across cell membranes [35].

Peptide-based drugs have already made a substantial impact in the treatment of various diseases. For instance, insulin, a peptide hormone, has transformed diabetes management while peptide-based antibiotics have effectively treated bacterial infections [4]. The development of peptide-based vaccines against infectious diseases and cancer is another exciting frontier in peptide therapeutics [42].

Despite their promise, peptide therapeutics face challenges like limited oral bioavailability and susceptibility to proteolytic degradation [37]. However, recent advances in peptide engineering, such as cyclization [20], incorporation of non-natural amino acids [60], and the development of novel delivery systems, are addressing these limitations and expanding the therapeutic potential of peptides.

1.4 Malaria: A Persistent Global Health Issue

Malaria, predominantly transmitted through bites of infected female Anopheles mosquitoes, remains a significant global health burden, particularly in tropical and subtropical regions [10, 17]. Its severity is underscored by its lethality, especially in young children and pregnant women. In 2019, the World Health Organization (WHO) reported an estimated 229 million malaria cases and 409,000 deaths, highlighting the urgent need for effective treatment alternatives [24]. Around 6.3 million cases were reported from the Southeast Asia region, majority of cases were present in India [45].

The *Plasmodium falciparum* parasite, the most lethal species causing malaria, has developed resistance to multiple antimalarial drugs, highlighting the urgent need for novel therapeutic strategies [13]. This resistance hampers treatment efficacy, potentially leading to prolonged illness, increased healthcare costs, and elevated mortality risks. Beyond immediate mortality, malaria can have lasting detrimental effects on individuals, even in non-fatal cases. Recurrent infections can contribute to anemia, cognitive decline (particularly in children), and other complications, ultimately diminishing quality of life and economic productivity [49].

Peptide therapeutics offer a promising avenue for combating malaria. Peptides can target specific parasite proteins essential for survival and replication, potentially overcoming drug resistance mechanisms [34]. Additionally, the lower likelihood of resistance development against peptides compared to small molecules makes them attractive candidates for antimalarial drug development. Recent efforts have focused on designing peptide inhibitors against key parasite proteins like *Plasmodium falciparum* Erythrocyte Membrane Protein 1 (PfEMP1), which plays a crucial role in parasite sequestration and immune evasion [26]. Computational methods, particularly generative AI approaches, have proven invaluable in accelerating the discovery of novel peptide-based antimalarials by efficiently exploring the vast peptide sequence space and identifying promising candidates with high affinity and specificity for critical parasite targets [5].

Chapter 2

HYDRA: A Novel Framework for Receptor-Aware Peptide Design

2.1 Problem Formulation and Objectives

Our study aims to computationally design novel, stable peptide binders that form high-affinity complexes with specific receptor protein binding pockets. We represent the binding pocket as a set of atoms, $P = \{x_P^i, v_P^i\}$, where x_P^i denotes the three-dimensional Cartesian coordinates of each atom in the protein, and v_P^i signifies its corresponding feature vector. The chosen feature vector, v_P^i, specifically encodes the atom type and its associated amino acid residue. Following this, we aim to generate peptides represented as $M = \{x_M^i, v_M^i\}$. Here, x_M^i corresponds to the 3D Cartesian coordinates of each atom within the peptide, and v_M^i signifies its feature vector, similarly encoding atom type and the corresponding amino acid residue it belongs to.

2.2 Constructing a Hybrid Diffusion Model for Peptide Generation

To promote the generation of chemically stable peptides for the target receptor, HYDRA employs a two-stage process: (1) amino acid residue generation using a diffusion model and (2) peptide reconstruction with binding affinity optimization, as illustrated in Figure 2.1. This necessitates an intermediate representation for the generated residues following the first stage. We represent this intermediate state as $A = x_A^i, v_A^i$, where x_A^i denotes the center-of-mass for each amino acid in the putative peptide and v_A^i encodes the corresponding amino acid type. Subsequently, the second stage utilizes this intermediate representation as input to reconstruct the final peptide, aiming to maximize its predicted binding affinity towards the target receptor.

2.2.1 Prior-Based Residue Count Estimation

Since the generative diffusion process is non-autoregressive, unlike sequential generation models, the total number of residues within the protein pocket cannot be determined incrementally during the generation phase. To address this, we precompute a prior distribution for the number of residues based

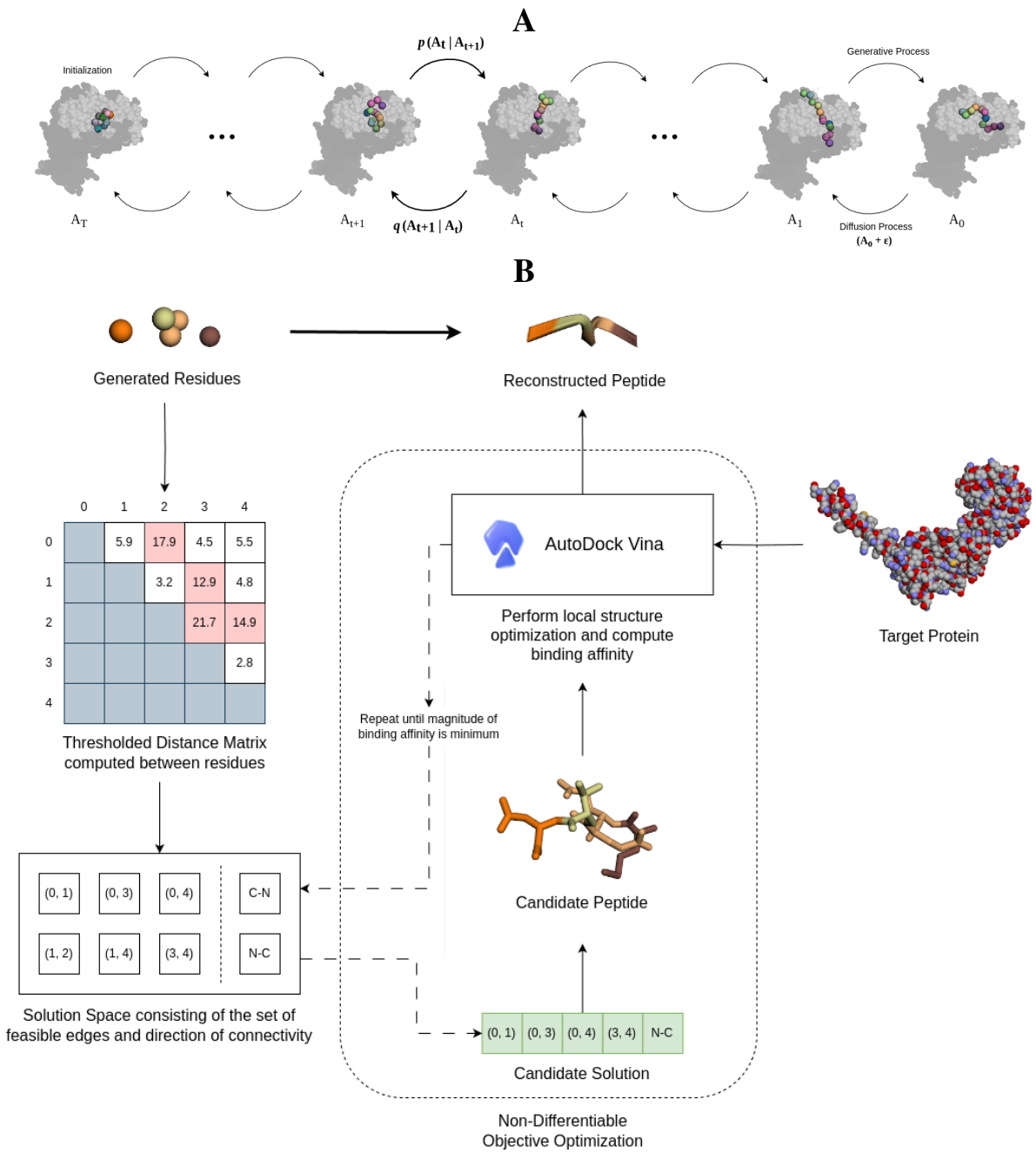


Figure 2.1 HYDRA's two-stage peptide generation workflow. **(A)** The diffusion process gradually injects noise into the data, and the generative process learns to recover the data distribution from the noise distribution by means of an SE(3)-equivariant network. **(B)** Upon generating amino acid residues, inter-residue center-of-mass distances are then computed, generating a distance matrix that encodes spatial relationships between them. Subsequently, a threshold is applied to this matrix to eliminate unrealistic connections, defining the feasible solution space for peptide conformations. A non-differentiable optimization algorithm then navigates this solution space. The objective function minimizes the binding affinity computed by AutoDock Vina, which corresponds to the predicted peptide conformation resulting in the most stable complex with the receptor.

on the training data.

The prior distribution leverages the concept of pocket size. First, we calculate the top 10 furthest pairwise distances between protein atoms within each pocket in the training set. This step captures the range of pocket sizes encountered in the training data. To mitigate the influence of potential outliers, we select the median value from these top 10 distances as a robust estimate of the pocket size for each training pocket. Next, we subdivide the range of observed pocket sizes into 10 quantiles. This discretization step creates 10 bins, each representing a specific pocket size range. For each bin, we leverage the histogram of the number of amino acids within training pockets falling within that specific size range. This histogram serves as the prior distribution for the number of residues to be generated within pockets of similar size during the diffusion process. During generation, the non-autoregressive diffusion model requires the number of residues to be determined upfront. We achieve this by randomly sampling from the precomputed prior distributions associated with the estimated pocket size for the protein being generated. This approach ensures that the generated protein sequences have a realistic number of residues based on the pocket size to which it is expected to bind.

2.2.2 Residue Generation Process

Deep generative modeling has recently surged in popularity due to its ability to generate novel, high-fidelity data across various domains, from drug discovery to artistic creation. Among deep generative models, diffusion models are gaining prominence due to their ability to effectively generate realistic data through a denoising process. Diffusion models are inspired by non-equilibrium thermodynamics [51, 22] and employ a sequential process of noise injection and denoising to learn the underlying distribution of data. During training, the model progressively adds Gaussian noise to real data points, gradually transforming them into isotropic Gaussian noise. This process is modeled as a Markov chain with T discrete steps. Leveraging the Markovian property, the model can efficiently compute the probability density at any given step t solely based on the probability density at the preceding step $t-1$.

Diffusion models have emerged as a promising approach in drug design due to their ability to generate diverse and high-quality 3D molecular structures in a non-autoregressive fashion [9] as an improvement over sequence-based molecule generation models that make use of Simplified Molecular-Input Line-Entry System (SMILES) [57] representations that lack detailed spatial information. Inspired by the success of diffusion models in creating 3D molecular structures, we explored their potential in generating peptide shapes specifically designed to fit into target binding pockets. For conciseness, we represent a peptide as a set of amino acid residues $A = [x, v]$, where $[\cdot, \cdot]$ is the concatenation operator, $x \in \mathbb{R}^{R\times 3}$ denotes the 3D Cartesian coordinates of the center-of-mass of each amino acid residue, and $v \in \mathbb{R}^{R\times K}$ denotes the one-hot encoded amino acid residue type.

We use a Gaussian distribution ($\mathcal{N}$) for the continuous coordinates and a categorical distribution ($\mathcal{C}$) for the discrete residue types represented as one-hot vectors. The residue distribution is modeled as a product of these individual distributions. A small Gaussian noise and a uniform noise across all categories are added to the coordinates and types, respectively, at each time step t:

$$q(A_t|A_{t-1}|P) = \mathcal{N}(x_t; \sqrt{1-\beta_t}; x_{t-1}, \beta_t I) \cdot \mathcal{C}(v_t|(1-\beta_t)v_{t-1} + \beta_t/K)$$

Here, β_i denotes fixed variance schedules, which may differ in practice but are denoted with the same symbol for maintaining conciseness. By employing the reparameterization trick and taking $\alpha_t = 1 - \beta_t$ and $\bar{\alpha_t} = \Pi_{s=1}^{t} \alpha_s$, the iterative noise injection process can be significantly accelerated. A_t can now be represented as:

$$A_t = \sqrt{\alpha_t} A_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1} = \sqrt{\alpha_t \alpha_{t-1}} A_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\epsilon}_{t-2} = \cdots = \sqrt{\bar{\alpha}_t} A_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

Here, $\epsilon_{t-1}, \epsilon_{t-2}, \ldots$ are noise from $\mathcal{N}(0, I)$ and ϵ combines the noise terms. Consequently

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1-\bar{\alpha}_t)I)$$

$$q(v_t|v_0) = \mathcal{C}(v_t|\bar{\alpha}_t v_0 + (1-\bar{\alpha}_t)/K)$$

In the denoising process, we aim to recover the initial peptide, A_0, from the final noisy state, A_T. However, directly calculating the exact reverse distribution added with the noise is intractable. To address this, we employ a neural network parameterized by θ to approximate this reverse distribution.

$$P_\theta(A_{t-1}|A_t, P) = \mathcal{N}(x_{t-1}; \mu_\theta([x_t, v_t], t, P), \sigma_t^2 I) \cdot \mathcal{C}(v_{t-1}|c_\theta([x_t, v_t], t, P)$$

Crucially, this generative process must be invariant to rotations and translations of the protein-peptide complex. This inductive bias ensures consistent likelihood predictions $P_\theta(A_0|P)$, essential for accurate 3D peptide structure generation. As informed by existing literature [32, 59], the Markov transition $P_\theta(A_{t-1}|A_t, P_b)$ must be SE(3)-equivariant and the initial density of our generative process, $P(A_T|P_b)$, is SE(3)-invariant. We need to take into account the amino acid residue coordinates because, during the generative process, atom types are always invariant to SE(3)-transformation. Here, we model $[x_0, v_0]$ to get $\mu_\theta([x_t, v_t], t, P)$ and $c_\theta([x_t, v_t], t, P)$.

At the l-th layer, the hidden embedding $\mathbf{h}$ and coordinates $\mathbf{x}$ of the amino acid residues are updated alternately as follows:

$$h_i^{l+1} = h_i^l + \Sigma_{j \in V, i \neq j} f_h(d_{ij}^i, h_i^l, h_j^l, e_{ij}; \theta_h)$$

$$x_i^{l+1} = x_i^l + \Sigma_{j \in V, i \neq j} (x_i^l - x_j^l) f_x(d_{ij}^i, h_i^{l+1}, h_j^{l+1}, e_{ij}; \theta_x). \mathbb{1}_{pep}$$

Here, d_{ij} is the Euclidean distance between residues i and j, and e_{ij} denotes a connection between these residues. We use a mask $\mathbb{1}_{pep}$ in order to refrain from updating protein atom coordinates. The initial residue embedding $\mathbf{h}^0$ is obtained from an embedding layer that encodes the amino acid information,

and the final residue embedding $\mathbf{h}^L$ is used to obtain the final peptide features through a Multi-Layer Perceptron and a Softmax function.

We train the model by minimizing the variational bound on the negative log-likelihood. Due to the Gaussian nature of $q(x_{t-1}|x_t, x_0)$ and $P_\theta(x_{t-1}|x_t)$, the KL-divergence for the coordinate loss admits the closed-form expression:

$$L_{t-1}^x = \frac{1}{2\sigma_t^2}\|\mu_t(\tilde{x_t}, x_0) - \mu_\theta([x_t, v_t], t, P)\|^2 + C = \gamma_t\|x_0 - \hat{x_0}\|^2 + C$$

Here, $\gamma_t = \bar{\alpha_{t-1}}\beta_t^2/2\sigma_t^2(1-\bar{\alpha_t})^2$ and C is a constant. In practice, training the model with $\gamma_t = 1$ could achieve better performance [22]. We compute the residue type loss directly using the KL-divergence for categorical distributions, given by:

$$L_{t-1}^v = \Sigma_k c(v_t, v_0)_k \log c(v_t, v_0)_k / c(v_t, \hat{v_0})_k$$

The overall loss function is formulated as a weighted combination of the residue coordinate loss and the residue type loss:

$$L = L_{t-1}^x + \lambda L_{t-1}^v$$

2.2.3 Peptide Reconstruction Strategy

Following the generation of individual amino acid residues, the next step involves assembling them in the correct sequential order to form complete peptides. This reconstruction process effectively translates the intermediate state A into the final peptide sequences represented by M. To achieve chemically stable protein-peptide complexes, we prioritize the identification of optimal residue connectivity patterns within the generated peptides. This optimization process maximizes the binding affinity of the connected peptide with the target receptor, thereby promoting complex stability. To achieve this, we first compute the distance matrix $\mathbf{D}_{n\times n}$ such that

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

Here, n denotes the number of amino acid residues generated that form the peptide, and d_{ij} denotes each element of the distance matrix that represents Euclidean distance between the centers of mass of residue i and residue j in 3D space. Due to the symmetry of the distance matrix, the lower triangle can be ignored, leaving $\frac{n\times(n-1)}{2}$ possible edges in the solution space. Following the initial edge identification, a filtering step is applied to eliminate connections exceeding a biologically relevant distance threshold. This step prioritizes the generation of peptides with realistic conformations, as the chemical nature of peptide bonds restricts the maximum distance attainable between adjacent amino acids.
The reconstruction process generates a distance matrix encompassing all potential inter-residue edges. However, due to the inherent limitations of peptide bond lengths, connections between residues positioned far apart in 3D space are physically implausible. To address this, we employ a distance threshold

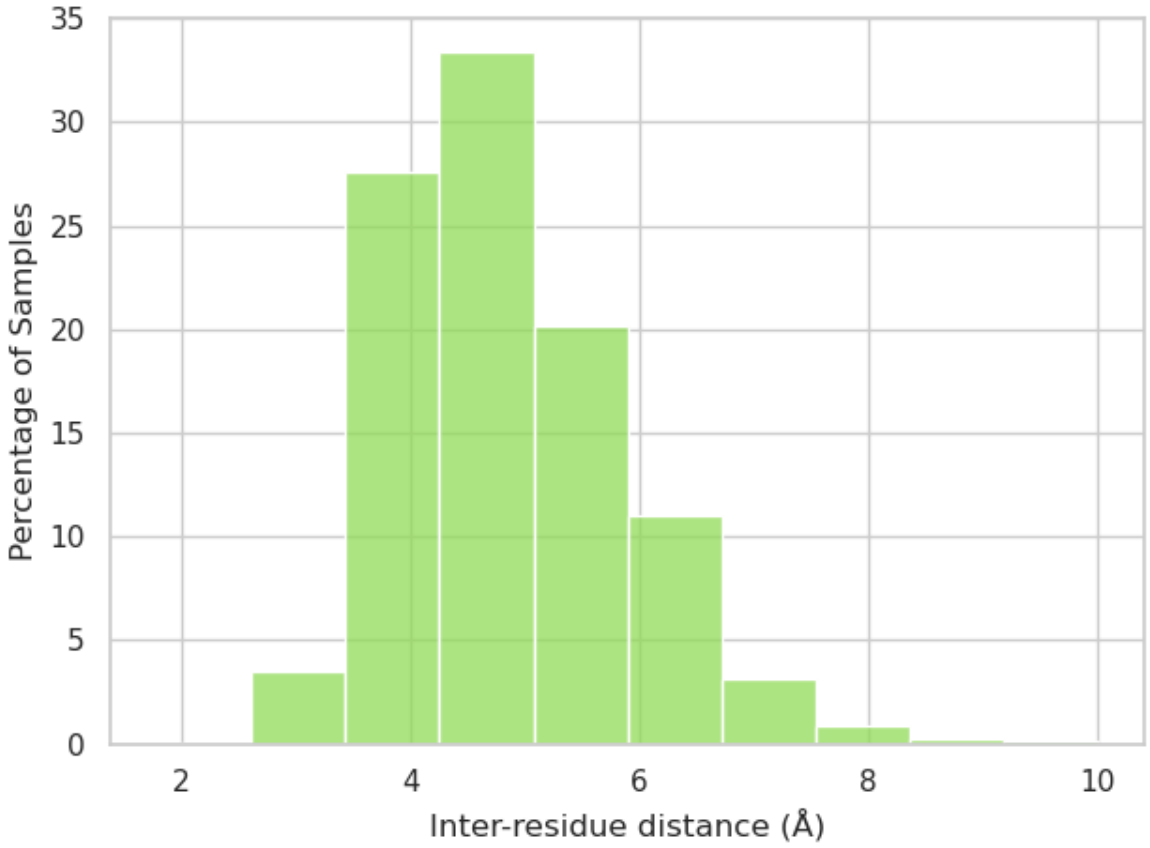


Figure 2.2 Distribution of inter-residue distances in peptides in the training data. 99.35% of distances lie below 8.

to eliminate unrealistic edges.

The threshold value is meticulously chosen based on the training data. We calculated the distances between the center of mass for each residue pair within the training complexes. Figure 2.2 illustrates the distribution of inter-residue distances of peptides in the training data. This analysis revealed that over 99% of inter-residue distances fell below 8. Consequently, a threshold of 8 was established to effectively filter out improbable connections during reconstruction.

The solution space is reduced to M possible edges following the distance-based thresholding. However, selecting $(n-1)$ edges from this set to define the final peptide structure remains challenging. Each edge specifies a connection between residues i and j, but its directionality determines which residue harbors the C-terminus and, consequently, which residue holds the N-terminus upon bonding.

Construction of a candidate peptide from a chosen set of edges and their directionalities involves a two-stage process. First, the amino acid residues are virtually placed at their predicted center-of-mass positions, and peptide bonds are formed between residues based on the chosen edges, establishing the initial peptide structure. This initial structure then undergoes energy minimization using a Merck Molecular Force Field (MMFF) [19] to optimize its geometry and achieve a more relaxed, lower-energy conformation. The resulting structure represents the fully reconstructed candidate peptide. Subsequently, the binding affinity between this peptide and the target receptor is assessed using AutoDock Vina [53]. This software performs a local structure optimization of the peptide within the target binding pocket. The optimization employs the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [21], followed by the calculation of binding affinity using the Vina Scoring Function [53]:

$$\Delta G = w_{vina}(E_{vd} + E_{elec} + E_{hbond}) + w_g E_g + w_h E_h$$

The Vina Scoring Function incorporates the following terms: E_{vd}, representing the van der Waals interaction energy; E_{elec}, denoting the electrostatic interaction energy; E_{hbond}, accounting for the hydrogen bonding interaction energy; E_g, the torsional free energy; and E_h, a reference state correction term. These terms are weighted by coefficients w_{vina}, w_g, and w_h, respectively, which were optimized and assigned during the training of the scoring function [53].

To identify the peptide with the best binding affinity, an exhaustive evaluation of $2 \cdot \binom{s}{n-1}$ possible amino acid configurations would be required. This becomes computationally intractable as the peptide length (n) and, consequently, the solution space size ($s \geq (n-1)$) increases due to the high cost associated with construction, local structure optimization, and binding affinity calculations. This extensive search space necessitates a heuristic-based approach to guide the search towards promising regions. The peptide reconstruction process inherently lacks a differentiable objective function due to discrete steps in establishing the initial peptide structure, energy minimization using a force field, and local structure optimization through MMFF. Consequently, gradient-based optimization methods become inapplicable. We opted for a heuristic optimization approach, which treats the objective function as a "black box" and iteratively refines candidate solutions based on their performance within this function. Several no-gradient stochastic optimization algorithms were evaluated, including Genetic Algorithms [23], Simulated Annealing [30], and Particle Swarm Optimization (PSO) [11, 28]. Binary Particle Swarm Optimization (BPSO) emerged as our preferred choice as we observed faster convergence while achieving comparable results to other methods.

PSO utilizes a population of candidate solutions (particles) representing potential peptide bonds. Each particle has a position in the search space and a velocity guiding its movement. The objective function, which consists of constructing the peptide from the candidate bonds and computing its binding affinity, acts as a fitness function, evaluating each particle's suitability. Particles track their personal best (p_{best}) position, and the swarm maintains a global best (g_{best}) position found by any particle. Particle velocities are updated iteratively based on their current state, attraction to their p_{best}, and attraction to the

g_{best}. This guides them toward promising areas of the search space. Positions are updated based on the new velocity, followed by fitness evaluation. If a particle finds a better position than its p_{best}, it updates its p_{best}. The g_{best} is updated if a particle discovers a superior position. This cycle repeats for several iterations, allowing the swarm to converge towards optimal solutions iteratively. Binary Particle Swarm Optimization (BPSO) is an implementation of the PSO algorithm where the particles are restricted to the binary domain. A detailed example of the reconstruction workflow for a peptide consisting of 5 amino acid residues is illustrated in Figure 2.1.

2.3 *In Silico* Evaluation of Designed Peptides

Upon generation, we conducted a comprehensive computational evaluation of the generated peptides. This evaluation aimed to assess their potential as functional drug molecules in various applications. We employed various computational tools to analyze crucial physicochemical properties and binding affinities relevant to their potential therapeutic efficacy.

2.3.1 Assessment of Physicochemical Properties

Physicochemical properties are critical determinants of a peptide's suitability as a therapeutic agent. For example, an important factor that determines their effectiveness as drugs is their half-life [38], which refers to the time it takes for half of the peptide to be broken down or eliminated from the body. Peptides with longer half-lives can exhibit their therapeutic effects over a longer time, which can be especially important for medications that require prolonged activity to be effective. Peptides often contain a mix of hydrophilic and hydrophobic amino acids, which can affect their solubility and interaction with cell membranes [36]. This can impact their ability to pass through cell membranes and work effectively inside cells. Peptides with a slightly higher percentage of hydrophobic residues may have increased cell permeability [36] and be more effective at interacting with cell membranes, as hydrophobic residues can interact very well with the hydrophobic regions of lipid bi-layers, enhancing the transit of the peptide across cell membranes [36]. Given the importance of physicochemical properties for therapeutic applications, we evaluated those of the generated peptides. The following properties were assessed:

I. Molecular Weight (MW)

The molecular weight of each peptide was calculated and compared to the established range (500-5000 Da) characteristic of drug-like peptides [12]. This parameter significantly influences a peptide's solubility, membrane permeability, and potential for toxicity. Peptides falling outside this range might exhibit undesirable pharmacological properties, hindering their potential as viable drug candidates.

II. Isoelectric Point (pI)

The pI of each peptide was determined, reflecting the specific pH at which the molecule possesses a net neutral charge. This property is crucial in determining a peptide's solubility, stability, and interactions with biological targets. Peptides with pI values outside the physiological pH range (7.4) might exhibit reduced solubility and stability in biological environments, compromising their therapeutic efficacy [14].

III. Half-life ($t_{1/2}$)

The half-life of a drug is the duration required for a drug's concentration in the bloodstream (or any other pertinent compartment) to drop by half. The half-life of each peptide was calculated with respect to in-vitro conditions in mammalian reticulocytes. This parameter is critical for determining the efficacy and duration of action of a potential drug. Peptides with shorter half-lives might require more frequent administration to maintain their therapeutic effect, whereas those with longer half-lives could potentially offer sustained drug action, reducing dosing frequency and improving patient compliance. Most peptides have *in vivo* half-lives of 230 minutes due to protease enzymatic breakdown and quick renal clearance (molecules smaller than 30 kDa are quickly eliminated by glomerular filtration) [41]. Consequently, increasing the *in vivo* half-life of peptides to fulfill their therapeutic potential without requiring high dosages or frequent administration is desirable [41].

IV. Instability Index (II)

The instability index, based solely on the amino acid sequence of each peptide, was calculated using the formula $\frac{10}{L}\sum_{i=1}^{L}\left[\frac{1}{log(f_i)}\right]$, where L is the length of the protein sequence and f_i is the dipeptide frequency of occurrence for each dipeptide in the protein sequence. A peptide's instability index is a numerical value that indicates how stable a protein or peptide will be [27]. Proteins or peptides are generally classified as stable if their instability index is lower than 40 and unstable if it is 40 or above [27]. Stable peptides are generally preferred for drug development due to their longer shelf life and potential for sustained activity *in vivo*. Peptides with high instability indices might undergo rapid degradation, limiting their therapeutic potential.

V. Aliphatic Index (AI)

The aliphatic index, reflecting a peptide's overall hydrophobicity, is determined by calculating the proportionate volume that aliphatic side chains (Leucine, Isoleucine, Valine, and Alanine) occupy in a protein or peptide [39]. Due to the presence of hydrophobic interactions, peptides with higher AI may display improved structural stability, contributing to their overall stability, which is crucial for a therapeutic peptide to remain active. Peptides with higher AI values might also have higher membrane permeability, enabling them to reach intracellular targets effectively [39].

VI. Grand Average of Hydropathy (GRAVY)

GRAVY is a numeric representation of a protein or peptide sequence's total hydrophobicity or hydrophilicity. This value is determined by the sum of the hydropathy of all amino acids divided by the total number of amino acids [33]. This value can be positive, negative, or zero. Positive GRAVY values indicate a hydrophobic sequence, whereas negative values show a hydrophilic sequence. A GRAVY near zero indicates a sequence with a balance of hydrophobic and hydrophilic residues. Positive GRAVY values indicate the sequence is dominated by hydrophobic residues, which may reduce the peptide's solubility in aqueous solutions. Particularly at high concentrations, hydrophobic residues tend to group to minimize interaction with water molecules, which can cause protein aggregation. Hydrophobic residues can interact very well with the hydrophobic part of lipid bilayers, enhancing the transit of the peptides across cell membranes, which may have increased cell permeability. Extremely high hydrophobicity, however, may result in non-specific interactions with cell membranes that compromise the integrity of the membrane and impair cellular processes [33]. Negative GRAVY values indicate higher solubility in aqueous solutions because of the greater interactions with water molecules.

2.3.2 Binding Affinity Prediction

Peptides exhibiting high binding affinity towards their target are more likely to disrupt crucial disease-associated processes and achieve therapeutic outcomes successfully. We assessed the binding affinity of each peptide towards its intended target using AutoDock Vina [53] and FRODOCK [43].

I. AutoDock Vina

AutoDock Vina is a widely used program for molecular docking [53]. It employs an empirical scoring function to estimate binding affinity, considering factors like hydrophobic interactions, hydrogen bonding, and electrostatics. Lower scores (in kcal/mol) indicate stronger predicted binding. Vina then performs local structure optimization using the Steepest Descent algorithm to refine the ligand's pose within the binding pocket, seeking the pose with the lowest binding energy. While these scores are estimates, Vina provides a valuable tool for exploring potential ligand-receptor interactions [53]. While AutoDock Vina excels at docking small molecules, its exhaustive search for optimal conformations becomes computationally inefficient for longer peptides due to their increased conformational space [44]. However, in this work, we circumvent this limitation by leveraging Vina's efficient scoring function for local optimization. We bypass the initial docking step, assuming pre-defined peptide poses within the protein-peptide complex. This allows us to exploit Vina's established capabilities for protein-peptide interaction energy calculations without incurring the full computational cost associated with peptide conformational sampling.

II. FRODOCK

FRODOCK, or Fast Rotational DOCKing, is an approach for protein-protein docking simulations [43]. Unlike AutoDock Vina, which uses an empirical scoring function, FRODOCK employs a correlation function for protein-protein docking. This function assesses the overlap of interaction patterns (e.g., hydrophobicity, electrostatics) between protein surfaces, with higher scores indicating greater potential for favorable interactions but not directly reflecting binding affinity. This distinction highlights the importance of selecting the appropriate docking tool based on the specific type of molecular interaction under investigation [43].

Utilizing two distinct scoring functions helps mitigate potential biases inherent to any single method and provides a more robust evaluation of predicted binding affinity. A comparison between different aggregations of these metrics computed for relevant datasets can be found below. (TODO: attach table)

2.3.3 Diversity Analysis

We quantify the diversity of the generated peptide sets using the average pairwise Tanimoto distance [2, 52]. This metric assesses the similarity between peptide pairs by considering the presence or absence of specific amino acids at each sequence position. Higher average Tanimoto distances indicate a more diverse peptide set capable of exploring a wider range of potential binding interactions.

Chapter 3

Experiments: Design and Implementation

3.1 Dataset Curation for Training and Evaluation

To construct training and evaluation sets, we curated peptide structures from the PepBDB database [58]. PepBDB (Peptide Binding DataBase) is a curated structural database specializing in biological peptide-protein interactions [58]. It provides clean data for structure-based peptide drug design, particularly for docking and scoring studies. Compiled from the Protein Data Bank (PDB), PepBDB focuses on structures of interacting peptide-protein complexes, with peptides limited to 50 amino acid residues in length. Regular monthly updates ensure the database reflects the latest data released in the PDB.

We curated 9225 protein-peptide complexes for the training dataset, 200 complexes for the validation dataset as well as 193 complexes for the test dataset. The complete dataset comprising 9618 protein-peptide complexes contained 1082 complexes that shared receptors with a subset of 504 unique receptors. To minimize overfitting and training set bias, the test set was carefully selected to have minimal overlap with the receptor proteins present in the train set. This resulted in the test set of 193 complexes with only 7 complexes sharing receptor proteins found in the training set. This minimal overlap between the training and test set data assures that our evaluation metric values are not a product of overfitting or train set bias. Table TODO summarizes the analysis of various physicochemical properties of peptides found in the dataset.

To ensure efficient computational processing and focus on the key interaction regions, a data preprocessing pipeline was implemented. The first step involved outlier removal based on protein size. Complexes containing exceptionally large proteins were excluded by imposing a maximum atom count threshold of 200 per protein. This filtering step mitigated potential memory limitations during subsequent deep learning model training. Next, to isolate the crucial binding pockets within the protein structures for analysis, the full protein structures were strategically sectioned. This was achieved by extracting a circumscribed spherical region with a radius of 10 centered around the bound peptide. This approach streamlined the data and made the process more computationally efficient by focusing solely

on the most relevant interaction interfaces critical for peptide binding.

The preprocessed dataset was then partitioned into distinct sets to prevent overfitting and enable robust evaluation of the trained HYDRA model. A hold-out test set comprising 2% (193 complexes) of the entire dataset was established. This unseen test set served as a completely novel benchmark for unbiased model performance assessment. Furthermore, a validation set of 200 complexes (approximately 2% of the remaining data) was designated for continuous monitoring of the training process and hyperparameter optimization. Finally, the remaining 9225 complexes constituted the primary training set.

To assess HYDRA's ability to generalize to unseen receptors, we focused on *de novo* peptide binder generation (novel peptides not encountered during training) for each of the 193 binding pockets within the independent test set. For each pocket, we generated 30 unique peptides, resulting in a total of 5790 novel peptides to evaluate HYDRA's binding prediction for unseen receptors.

3.2 Model Architecture and Specifications

The target-aware in-place residue generation step is implemented through a SE(3)-equivariant Graph Neural Network to model the interaction between the peptide residuals and the target protein atoms. The key/value embedding and attention scores are generated through a 2-layer Multi Layer Perceptron with LayerNorm and ReLU activation. The SE(3)-equivariant network contains 9 equivariant layers where f_h and f_x are implemented as graph attention layers for features and coordinates with 16 attention heads and 128 hidden features. We initiated our architecture and hyperparameter search using values reported in related literature [18] as a starting point. Subsequently, we conducted a series of experiments with multiple configurations, systematically adjusting and refining these parameters. Through this iterative process of testing and optimization, we ultimately arrived at the current set of model parameters that yielded the best performance. Figures pertaining to the neural network architecture iteration are given below. (TODO: figures from wandb)

For the peptide reconstruction process, the Binary Particle Swarm Optimization (Binary PSO) algorithm is used. This variant of the standard PSO algorithm restricts particles to binary values (0 or 1). This directly aligns with our problem of selecting a subset of edges from the complete edge space. Each edge can be represented by a binary value: 1 signifying the edge is included in the final peptide and 0 indicating exclusion. For Binary PSO, 50 particles are instantiated with cognitive (c1) and social (c2) acceleration coefficients set to 2.5 and 0.5, respectively. Additionally, the inertia weight is set to 0.9. The optimal number of particles in PSO can vary based on the complexity of the problem. We opted for a common starting point of 50 particles, striking a balance between exploration of the search space and exploitation of promising regions. The cognitive (c1) and social (c2) coefficients were determined empirically to further balance exploration and exploitation within the algorithm. We found that c1 = 2.5

and c2 = 0.5 achieved this balance effectively. The cognitive coefficient emphasizes individual exploration based on a particle's best prior position (p_{best}), while the social coefficient promotes convergence towards the best position found by any particle in its neighborhood (g_{best}). Finally, the inertia weight (w) was also tuned empirically to a value of 0.9. This parameter controls the momentum of particles, allowing them to maintain some exploration history while gradually decreasing over time to encourage convergence towards promising solutions in later stages of the optimization process.

3.3 Training Methodology

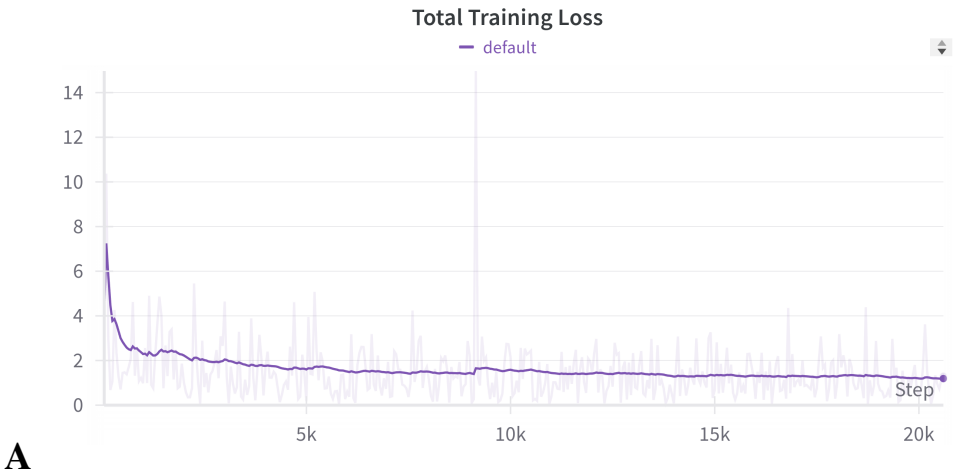
The model was trained using the Adam optimizer [29], with an initial learning rate of 10^{-3}. To prevent overfitting and improve generalization, a data augmentation strategy was employed during training. This involved adding a small Gaussian noise with a standard deviation of 0.1 to the protein atom coordinates. The forward and reverse diffusion processes took place through 1000 steps. Additionally, a learning rate decay schedule was implemented to decay the learning rate exponentially with a factor of 0.6 towards a minimum value of 10^{-6} if there is no improvement in the validation loss for 10 consecutive steps. The model was trained using a batch size of 2, and to balance the contributions of different loss terms within the overall loss function, a factor of $\alpha = 100$ was multiplied onto the residue type loss. Figures concerning the training progression are provided below.

To gain deeper insights into the training process, we tracked not only the total training and validation losses but also their individual components: position loss and feature loss. Position loss quantifies the error between the predicted continuous Cartesian coordinates (x, y, z) and the ground truth. Feature loss, on the other hand, measures the discrepancy between the predicted categorical feature distribution and the actual distribution. Figures 3.1 and 3.2 illustrate the behavior of these loss components throughout training.

The deep diffusion model was trained on the Ada HPC Cluster with 4x NVIDIA GeForce RTX 2080 Ti GPUs using the Distributed Data Parallel (DDP) Strategy. All inference and reconstruction experiments were carried out on multiple nodes, each with 40x Intel Xeon E5-2640 v4 CPUs, 80 GB of RAM, and 1x NVIDIA GeForce RTX 2080 Ti GPU.

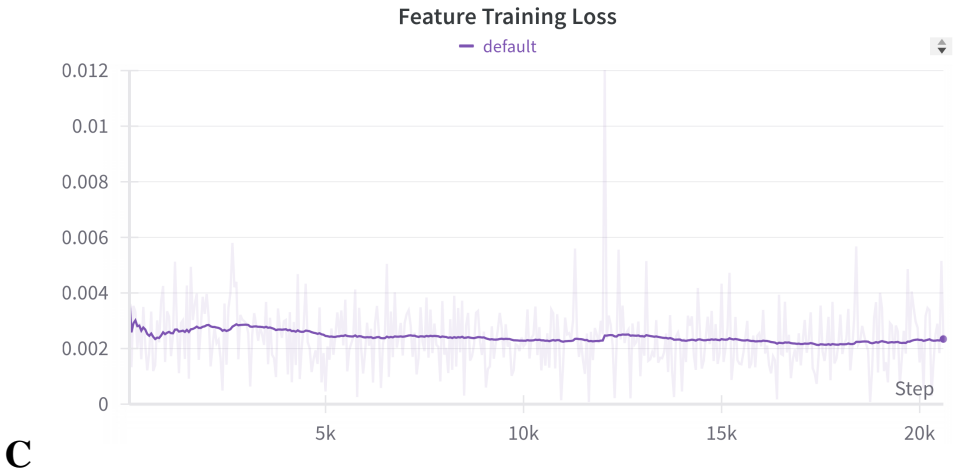
3.4 Optimization of Evaluation Algorithm Parameters

Autodock Vina was employed to compute the binding affinity between the reconstructed peptide and the target protein. Although docking was not performed, Vina's built-in scoring function served as the metric for evaluating both the reconstruction process and the final generated peptides. To facilitate Vina docking calculations, the reconstructed peptide's size (in Angstroms) and center coordinates were precomputed and provided as input parameters. We opted to utilize the default settings within Vina for the



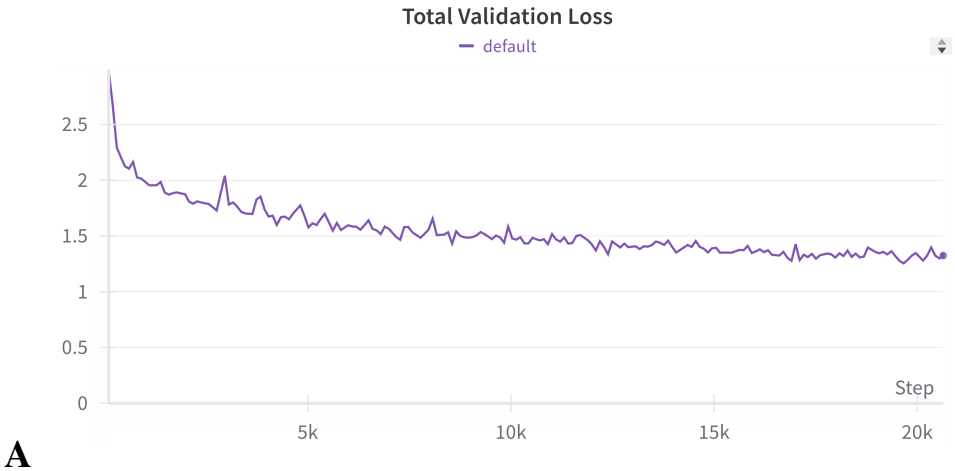
A

B

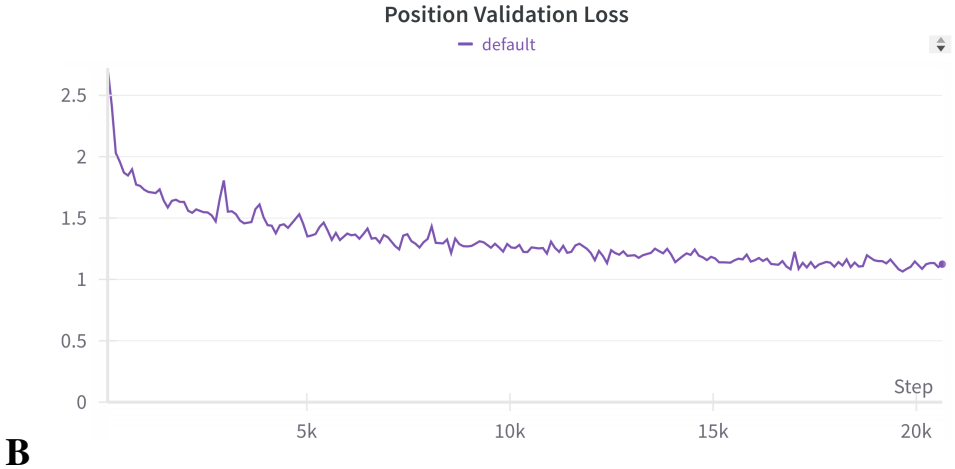


C

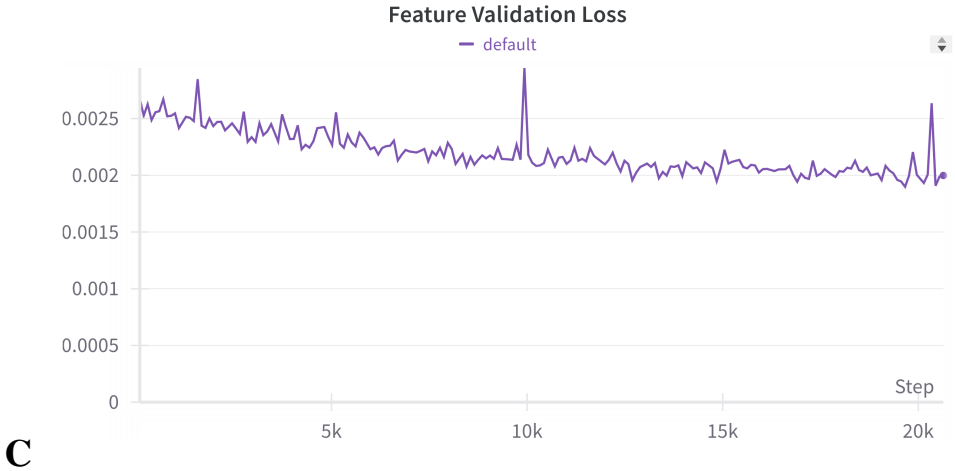
Figure 3.1 Evolution of individual training loss components throughout the training process. To improve visualization and reduce noise, a Time-Weighted Exponential Moving Average (EMA) algorithm has been applied to smooth the curves.



A



B



C

Figure 3.2 Evolution of individual validation loss components throughout the training process.

exhaustiveness of the global search (set to 8) and the maximum number of binding modes to be generated (limited to 9). These defaults ensure a balance between computational efficiency and exploration of the search space.

For protein-peptide docking simulations, we utilized FRODOCK. However, prior to docking, the reconstructed peptides underwent preprocessing using the PDB2PQR tool. This preprocessing step ensured the peptides were properly prepared for docking by reconstructing missing atoms, adding hydrogens, and assigning atomic charges and radii based on the CHARMM force field. Following this preparation step, FRODOCK simulations were conducted using the Rosetta force field and the default SOAP potentials. Docking simulations can be computationally expensive, so to improve efficiency, the process was parallelized across 8 threads using OpenMP. This parallelization strategy significantly reduced the overall computational time required for the docking simulations.

3.5 Application to PfEMP1 Protein Targets

Plasmodium falciparum, the causative agent of the most severe form of malaria, expresses a diverse family of PfEMP1 proteins. These highly polymorphic surface antigens, comprising approximately 60 known variants, play a critical role in severe malaria pathogenesis by mediating the cytoadherence of infected erythrocytes to endothelial receptors. This adherence leads to sequestration and contributes to tissue damage [25]. PfEMP1 is a pivotal virulence factor secreted by the malaria parasite. It binds to the erythrocyte membrane, triggering the binding of red blood cells (RBCs) to blood vessels [40]. By obstructing tiny blood arteries, they exacerbate malaria infections and increase the risk of cerebral malaria, placental malaria, and severe anemia [25]. The parasite can avoid the host immune system because of this antigenic diversity of PfEMP1 family genes, since new infection can express different PfEMP1 variants that are not recognized by preexisting immune responses [46]. We leveraged single-cell RNA-seq data to identify the five most highly expressed PfEMP1 genes. Using CAVITY [61], we predicted strong and medium druggable binding sites within the proteins encoded by these genes (Figures 3.3, 3.4, 3.5). Subsequently, HYDRA was used to design potential peptide molecules specifically targeted to these binding sites (Figure 3.6). We also checked the binding affinities of generated peptides of one with other proteins (Figure 3.7). During analysis of the protein structures, we focused on extracellular and intracellular domains for peptide generation, leveraging cavities as potential binding pockets. Finally, we evaluated the binding affinities between the designed peptides and their target proteins on these PfEMP1 variants.

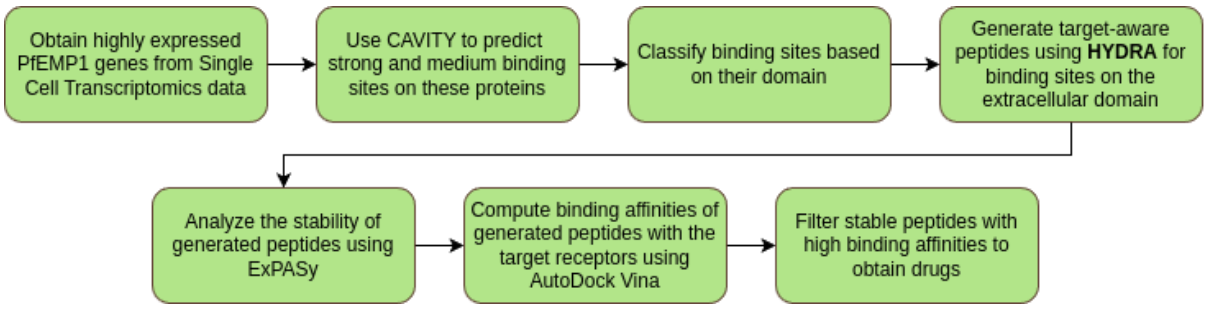


Figure 3.3 A consolidated flowchart illustrating the binding site detection and *de novo* peptide binder generation pipeline for PfEMP1 targets.

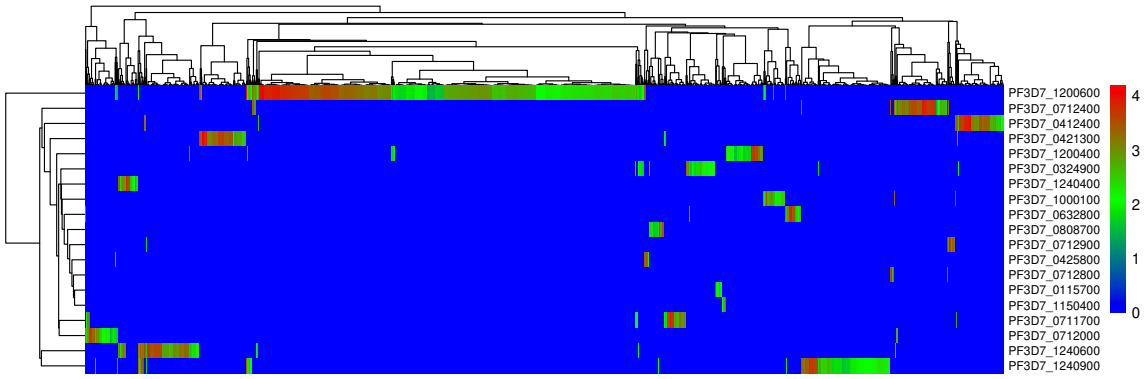


Figure 3.4 Gene expression profile of the PfEMP1 family. The heatmap suggests that the proteins PF3D7_1200600, PF3D7_1150400, and PF3D7_0712400 are highly expressed.

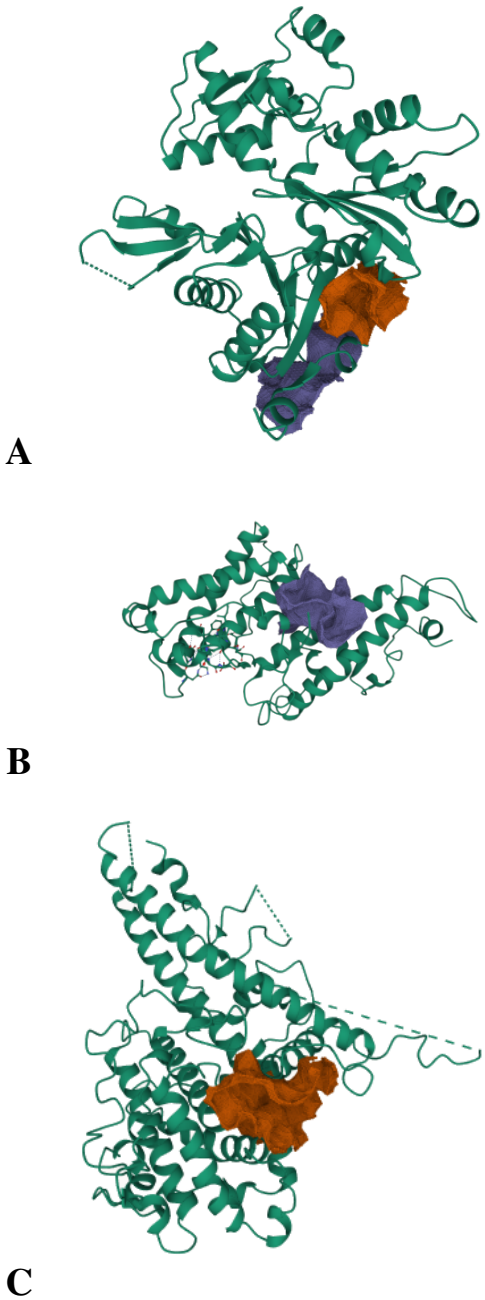
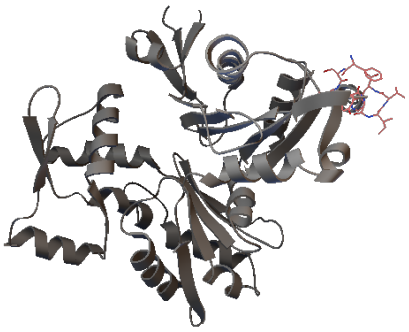


Figure 3.5 Using CAVITY to identify strong and medium binding sites. Visualization of identified binding sites of (*A*) PF3D7_0712400, (*B*) PF3D7_1200600, and (*C*) PF3D7_1150400. The red and blue regions highlight the binding sites.

A



B

C

Figure 3.6 3D structural representations of the binding sites on PF3D7_0712400 interacting with generated peptides (**A**) RLKAVIP, (**B**) FSVYGIVH, and (**C**) IVVGPAYG.

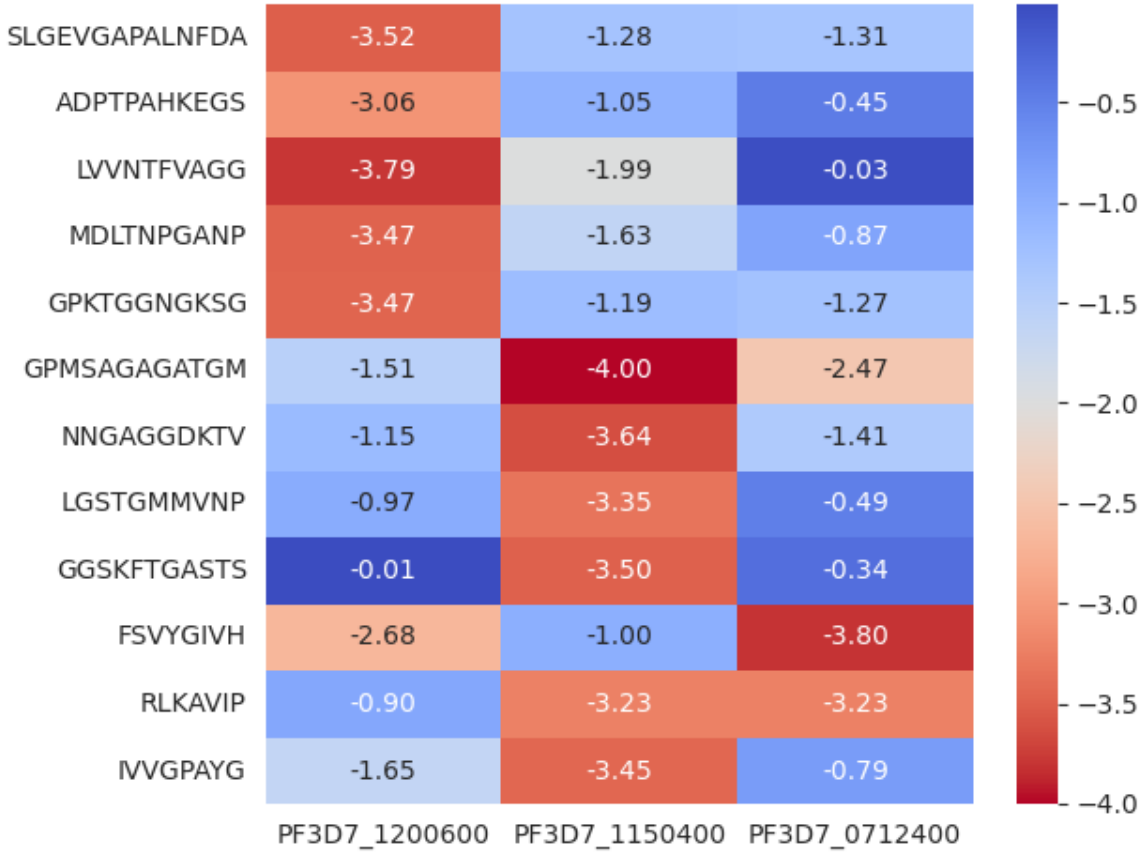


Figure 3.7 A heatmap visualizing the predicted binding affinities between the chosen target binding sites and peptides generated by HYDRA. The color intensity represents the predicted binding affinity, with lower values indicating stronger binding.

Chapter 4

Results: Analysis and Discussion

4.1 Comparative Analysis with Baseline Methods

To evaluate HYDRA's performance relative to the current state-of-the-art, we compared it against RFDiffusion [56]. RFDiffusion leverages a diffusion model framework to iteratively refine candidate peptides for target receptor binding. It begins with a pool of random sequences and progressively adjusts them based on predicted binding scores, aiming to converge toward high-affinity binders over multiple iterations [56]. While RFDiffusion stopped at predicting just the peptide backbone, we went a step further by utilizing ProteinMPNN [8] to generate full peptide sequences from these backbones for analysis of physicochemical properties.

4.2 Case Study 1: PepBDB Protein Targets

We generated 10 peptide binders for each binding pocket using RFDiffusion, mirroring the sampling strategy used with HYDRA. Since RFDiffusion solely predicts backbones, we used ProteinMPNN to translate each backbone into 3 candidate peptide sequences. The residue lengths for both RFDiffusion and ProteinMPNN were determined programmatically using the same algorithm employed by HYDRA. Finally, all generated peptides (from both HYDRA and RFDiffusion) underwent identical analyses, encompassing statistical properties, physicochemical properties, and binding affinity predictions.

Our evaluation demonstrates that HYDRA-designed peptides exhibit significantly better binding affinities to receptor pockets compared to those generated by RFDiffusion. As shown in Figure 4.1A, for approximately 65% of the target receptors, HYDRA-generated peptides achieved higher mean FRODOCK scores. This is further corroborated by Figure 4.1B, where HYDRA peptides display lower mean binding affinity scores (indicating stronger binding) compared to RFDiffusion for roughly 78% of the targets. This suggests that complexes formed by target receptors and HYDRA-designed peptides possess significantly greater stability. Additionally, we computed evaluation metrics using median and Top1

aggregations. Due to the nature of binding affinity predictions, when evaluating a set of peptide metrics, the median can provide a more reliable representation of the central tendency of the data, especially if there are a few peptides with exceptionally high or low values. The Top1 metric focuses on the single peptide with the best evaluation score and highlights the model's capability to identify the absolute best candidate peptide within a set. Figures 4.2 and 4.3 illustrate the complete distributions of different aggregations of binding affinity metrics.

In terms of diversity, both HYDRA and RFDiffusion-generated peptides exhibit similar mean Tanimoto scores around 0.6. However, Figure 4.1C reveals key differences in the distribution of scores across targets. HYDRA displays greater consistency in generating peptides with diversity scores close to the mean, suggesting a more uniform level of diversity. However, RFDiffusion exhibits a much wider variance in scores, indicating a less predictable level of diversity across generated peptides.

Analysis of physicochemical properties revealed that both models produced peptides within the established range for drug-like molecules in terms of molecular weight. However, HYDRA-designed peptides had isoelectric points closer to 7, indicating potentially greater solubility and stability compared to RFDiffusion-generated ones. Interestingly, the half-life data for both models showed a right-skewed distribution, meaning a few outliers skewed the mean values. While RFDiffusion peptides had a higher average half-life, HYDRA peptides boasted a higher median half-life, suggesting a more consistent level of stability across the generated set. Furthermore, HYDRA demonstrated a clear advantage in peptide stability. The percentage of stable peptides generated by HYDRA was significantly higher (58.31%) compared to RFDiffusion (31.02%). This trend was further confirmed by the lower instability index distribution for HYDRA peptides. Aliphatic Index values were comparable for both sets, with HYDRA having a slight edge in mean AI and RFDiffusion edging out on the median value. Finally, the GRAVY scores (generally negative for peptides) indicated higher aqueous solubility for RFDiffusion peptides compared to those generated by HYDRA.

The performance gap between HYDRA and RFDiffusion likely stems from their underlying design principles. HYDRA incorporates an explicit binding affinity optimization step, directing the generated peptides toward forming more stable complexes with target proteins. Conversely, RFDiffusion operates in a purely data-driven manner, resulting in peptides directly reflecting the distribution of the training data. A detaild comparison of both methods across various metrics is presented in Tables TODO and TODO. Figures 4.1C and 4.4C depict the distribution plots for the aforementioned properties.

4.3 Case Study 2: PfEMP1 Protein Targets

Our single-cell RNA sequencing (scRNA-seq) data revealed that certain PfEMP1 variants are highly expressed. We have utilized these highly expressed PfEMP1 proteins as targets, as they can potentially

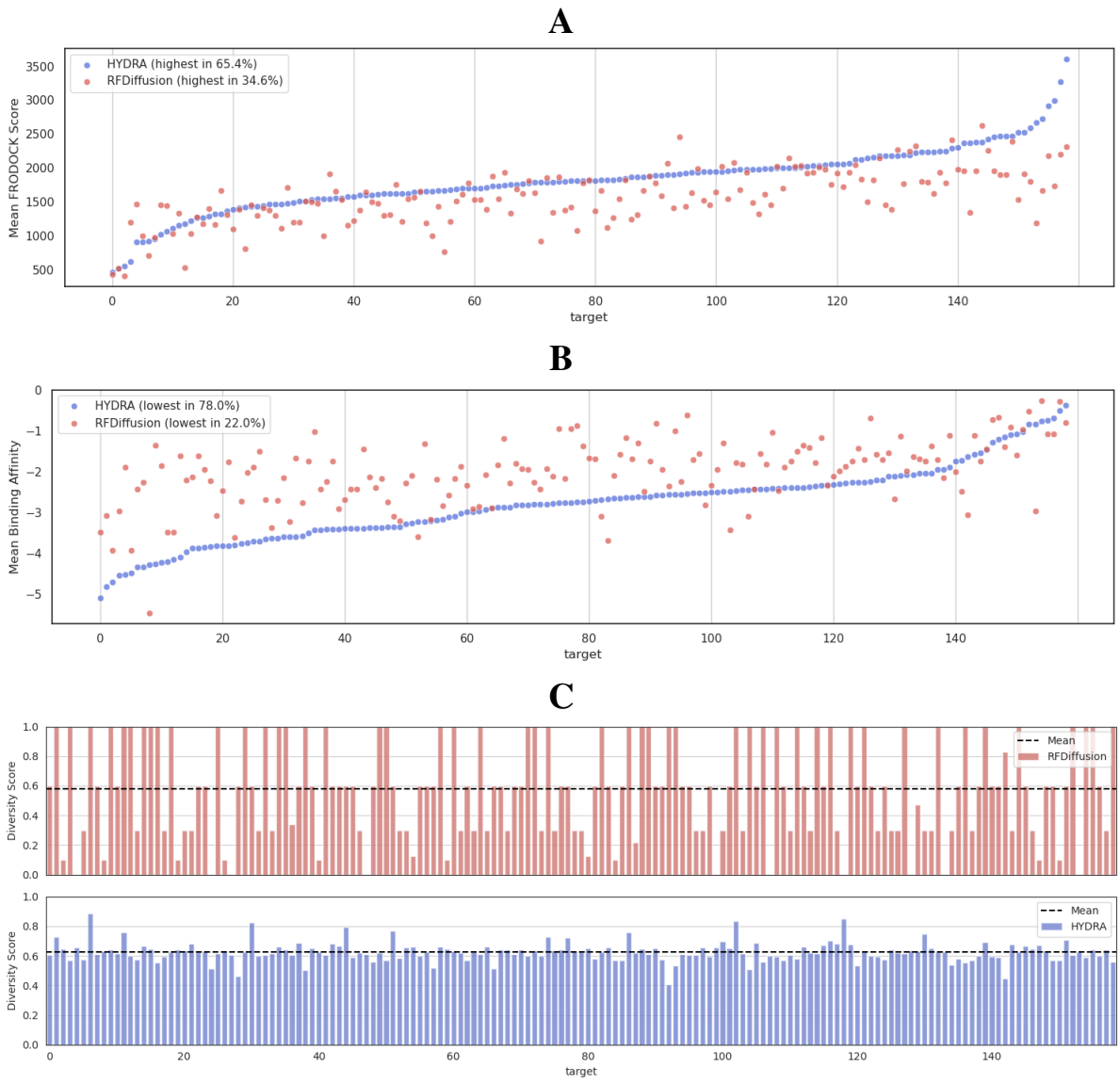


Figure 4.1 Comparison of binding affinity and peptide diversity between HYDRA and RFDiffusion. Figures **(A)** and **(B)** depict the mean binding affinities and FRODOCK correlation scores, respectively, for peptides generated by each model. Lower binding affinities and higher FRODOCK scores indicate stronger predicted protein-peptide complexes. Binding targets are sorted based on the mean score for HYDRA-generated peptides. **(C)** represents the pairwise Tanimoto diversity scores across the generated peptide sets for each binding target. While the average diversity scores are similar, the distribution of Tanimoto scores implies greater consistency in HYDRA's peptide diversity compared to RFDiffusion.

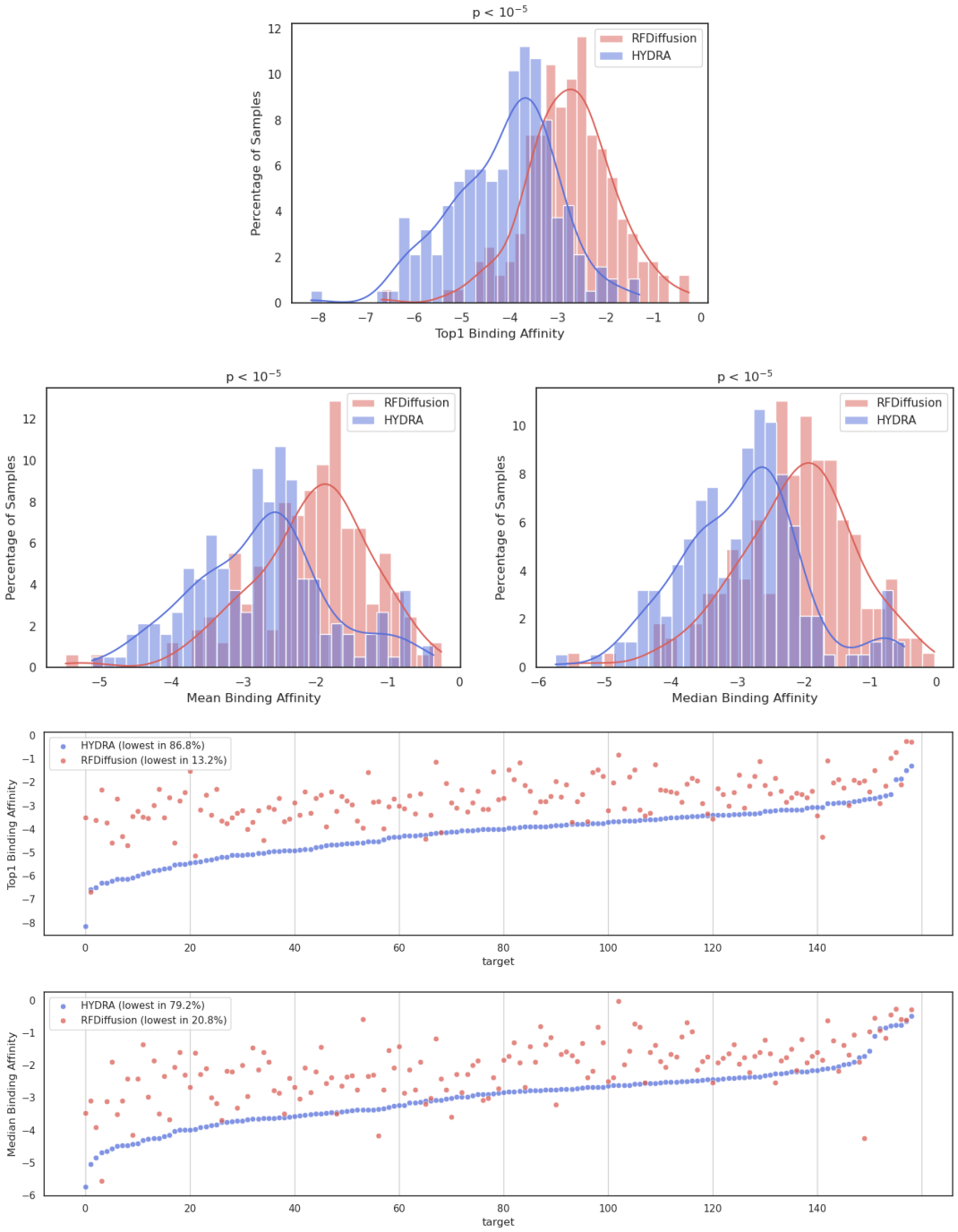


Figure 4.2 Comparing the binding affinities of peptides generated by HYDRA and RFDiffusion.

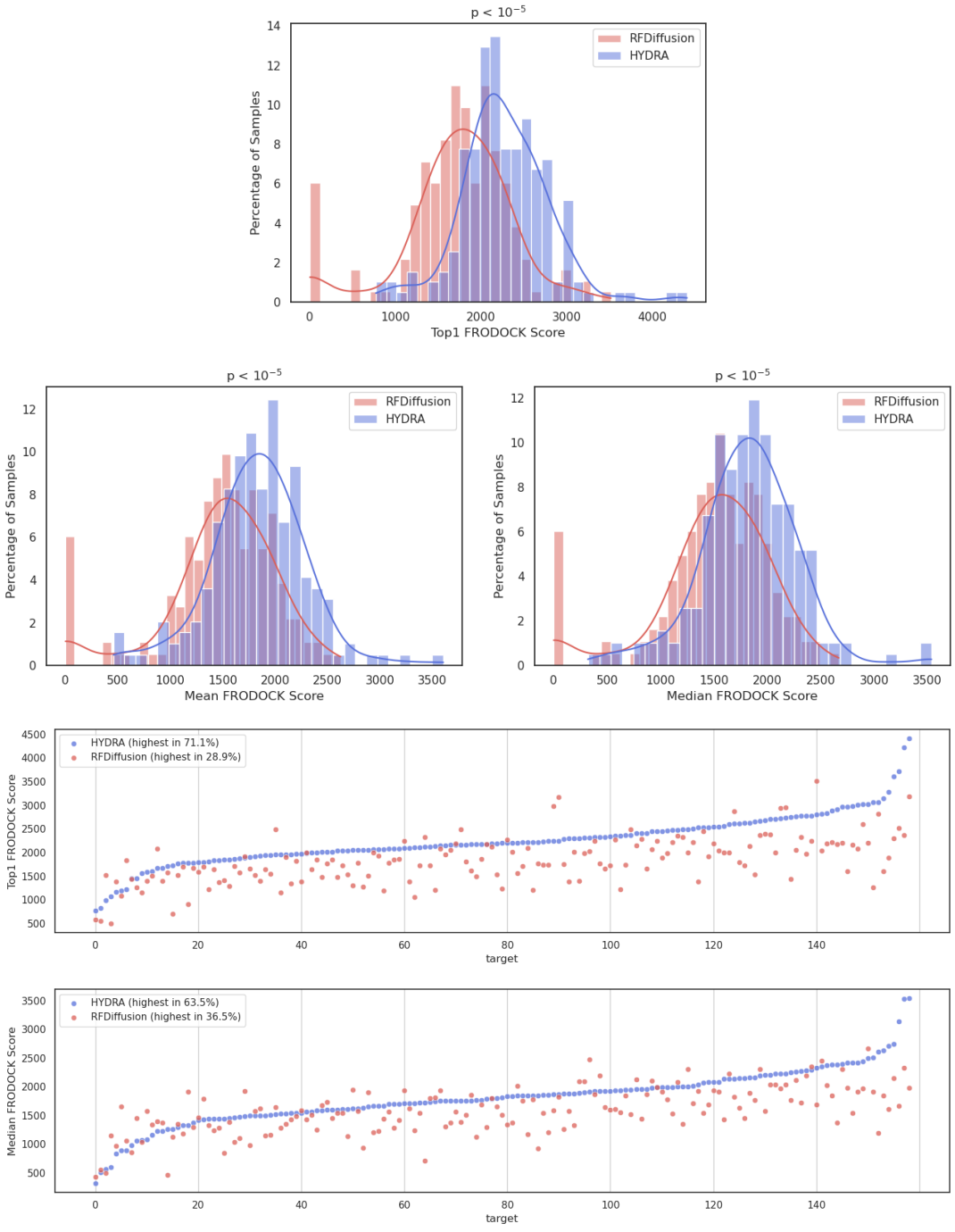


Figure 4.3 Comparing the FRODOCK scores of peptides generated by HYDRA and RFDiffusion.

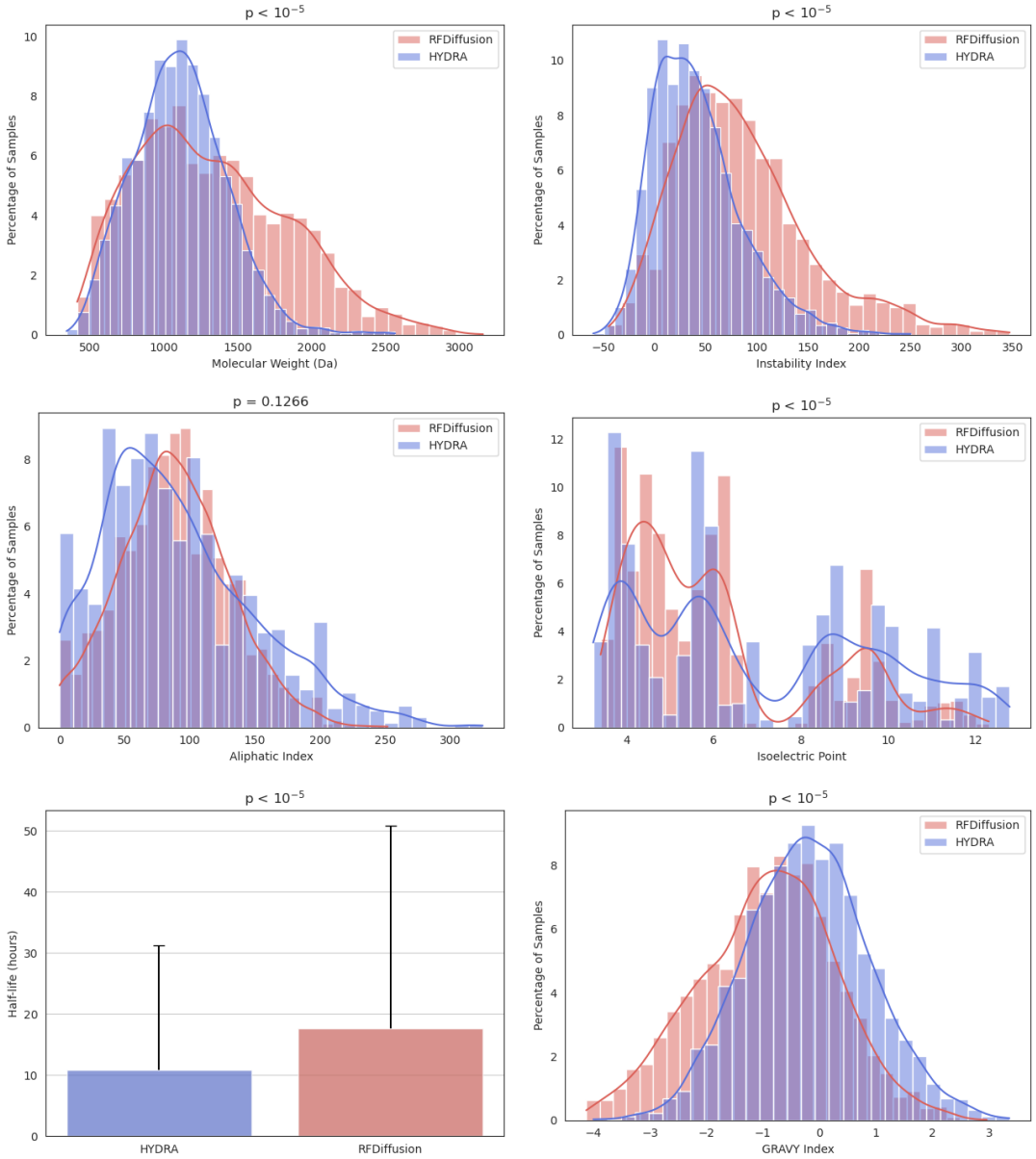


Figure 4.4 Comparison of physicochemical properties of peptides generated by HYDRA and RFDiffusion. Each histogram represents the distribution of peptides across different ranges for a specific property. The y-axis indicates the percentage of peptides within each range. **(A)** Molecular Weight in Daltons, **(B)** Instability Index, **(C)** Aliphatic Index, **(D)** Isoelectric Point, **(E)** Half-life in hours, **(F)** GRAVY Index. Peptides designed by HYDRA show a more favorable bias in their property distribution compared to RFDiffusion.

inhibit the binding of RBCs to blood vessels.

The emergence of parasite resistance to established antimalarial drugs like chloroquine and sulfadoxine-pyrimethamine poses a critical challenge. This resistance hampers treatment efficacy, potentially leading to prolonged illness, increased healthcare costs, and elevated mortality risks. Beyond immediate mortality, malaria can have lasting detrimental effects on individuals, even in non-fatal cases. Recurrent infections can contribute to anemia, cognitive decline (particularly in children), and other complications, ultimately diminishing quality of life and economic productivity [50]. Drug resistance presents a central obstacle in malaria control. The effectiveness of new drugs wanes as the Plasmodium parasite develops resistance mechanisms.

In an effort to design *de novo* therapeutic peptide binders against these parasites, we analyzed single-cell transcriptomics data of *P. falciparum* clone 3D7 to identify highly expressed PfEMP1s (PF3D7_1200600, PF3D7_1150400, PF3D7_0712400) based on cellular heterogeneity, cell-to-cell variability, and cellular states [7]. Using CAVITY [61], we predicted strong and medium binding sites based on their druggability score. Combining data from CAVITY and Protein Data Bank (PDB) [3], we identified extracellular domain binding sites crucial for potential drug interaction. After identifying binding sites, we used HYDRA to generate a set of candidate therapeutic peptides. ExPASy ProtParam [15] then assessed their properties (molecular weight, isoelectric point, half-life, stability, and hydrophobicity). For batch processing of peptides, an open-source Python command-line program wrapping around the ExPASy ProtParam Webserver was built and is available for free use at `https://github.com/v15hv4/easyprotparam`. We prioritized stable peptides with high predicted binding affinities and favorable biophysical properties. Finally, to gain a complete understanding of their potential as therapeutic agents, we selected a few peptides per protein for further analysis.

Through this rigorous selection process, we identified a final set of 12 candidate peptides as presented in Table TODO. These peptides are promising therapeutic agents targeting 3 distinct PfEMP1 variants. Furthermore, to assess the potential for broad-spectrum activity, we evaluated the binding affinity of these peptides with additional target receptors beyond those initially considered during their design, the results of which are illustrated in Figure 3.7. This analysis aimed to determine if any individual peptide could potentially bind to and interact with binding sites on multiple proteins, offering a broader therapeutic scope. We observe that two of the generated peptides, GPMSAGAGATGM and RLKAVIP, exhibit comparable binding affinities towards pockets 1150400 and 0712400, signaling a potential interaction with both sites.

Moving forward, the next crucial step in this research involves sending these selected peptides for wet lab experiments. This practical evaluation will be instrumental in validating our *in silico* findings

and assessing the efficacy and safety of these peptides in a biological context.

Chapter 5

Conclusion and Future Directions

Our novel hybrid deep learning approach, HYDRA, presents a significant advancement in target-aware peptide binder design. It goes beyond existing data-driven peptide design approaches by introducing a novel hybrid model that leverages a denoising diffusion model in conjunction with a binding affinity maximization algorithm. This approach utilizes a 3D-structure-specific representation of the protein binding site and the peptide, allowing for the generation of diverse, high-quality peptides tailored to specific binding locations. Due to the model's optimization of the binding affinity of protein-peptide complexes, it aims to generate peptides with enhanced stability and longer half-lives, potentially prolonging their therapeutic effect. HYDRA is also observed to generate peptides with a balanced ratio of hydrophilic and hydrophobic residues, enhancing their membrane contact and facilitating efficient cellular uptake.

To showcase the model's capabilities, we successfully used it to design *de novo* peptide binders for proteins expressed by PfEMP1 genes, key contributors to antigenic variation in the malaria parasite. By analyzing *Plasmodium falciparum* parasite single-cell transcriptome data, we identified highly expressed PfEMP1 genes and subsequently pinpointed strong and medium binding sites within their structures using CAVITY [61]. Following peptide generation using HYDRA, we selected a subset of peptides exhibiting superior characteristics for each protein, demonstrating the model's ability to generate highly stable, cell-permeable, and potentially effective drug candidates.

Overall, our findings demonstrate the remarkable potential of hybrid deep learning approaches like HYDRA in revolutionizing peptide drug discovery. This work paves the way for the development of novel therapeutic peptides with improved stability, efficacy, and targeted delivery, ultimately contributing to advancements in healthcare. In the future, it might be worth exploring the possibility of using fully differentiable local peptide structure optimization and binding affinity computation algorithms [55] in order to construct a faster, end-to-end deep learning framework for *de novo* design of stable, therapeutic peptides. It is also suggested that the inclusion of some unnatural (non-proteinogenic) amino acids in the peptide would benefit its stability [1]. This aspect can also be explored in future studies. From a

practical point of view, we acknowledge the limitation concerning the efficacy of the designed peptides in the real experimental and, more importantly, in clinical setup. However, the study would definitely facilitate potential peptide design for further experimental investigations. Recent studies have already demonstrated that machine learning-based peptide design protocols have successfully synthesized effective peptides. Current peptide design paradigms are still in their infancy, and we believe many more such investigations would accelerate its growth towards maturity in the near future.

Related Publications

Vishva Saravanan R, Soham Choudhuri, and Bhaswar Ghosh. **Hybrid Diffusion Model for Stable, Affinity-Driven, Receptor-Aware Peptide Generation**. Journal of Chemical Information and Modeling, DOI: 10.1021/acs.jcim.4c01020; August 28, 2024.

Bibliography

[1] A. Adhikari, B. R. Bhattarai, A. Aryal, N. Thapa, K. Puja, A. Adhikari, S. Maharjan, P. B. Chanda, B. P. Regmi, and N. Parajuli. Reprogramming natural proteins using unnatural amino acids. *RSC advances*, 11(60):38126–38145, 2021.

[2] D. Bajusz, A. Rcz, and K. Hberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):1–13, 2015.

[3] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[4] A. Cesaro, S. Lin, N. Pardi, and C. de la Fuente-Nunez. Advanced delivery systems for peptide antibiotics. *Advanced Drug Delivery Reviews*, 196:114733, 2023.

[5] L. Chang, A. Mondal, and A. Perez. Towards rational computational peptide design. *Frontiers in Bioinformatics*, 2, 2022.

[6] Cherkasov and A. et al. Qsar modeling: where have you been? where are you going to? *Journal of medicinal chemistry*, 57:4977–5010, 2014.

[7] S. Choudhuri and B. Ghosh. Computational approach for decoding malaria drug targets from single-cell transcriptomics and finding potential drug molecule. *bioRxiv*, pages 2024–01, 2024.

[8] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. Wicky, A. Courbet, R. J. de Haas, N. Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[9] Y. Du, T. Fu, J. Sun, and S. Liu. Molgensurvey: A systematic survey in machine learning models for molecule design. *arXiv preprint arXiv:2203.14500*, 2022.

[10] Duguma and T. et al. Prevalence of malaria and associated risk factors among the community of mizan-aman town and its catchment area in southwest ethiopia. *Journal of parasitology research*, 2022, 2022.

[11] R. Eberhart and J. Kennedy. A new optimizer using particle swarm theory. In *MHS'95. Proceedings of the sixth international symposium on micro machine and human science*, pages 39–43, 1995.

[12] V. Erckes and C. Steuer. A story of peptides, lipophilicity and chromatography–back and forth in time. *RSC Medicinal Chemistry*, 13(6):676–687, 2022.

[13] Fairhurst and R. M. et al. Artemisinin-resistant malaria: research challenges, opportunities, and public health implications. *The American journal of tropical medicine and hygiene*, 87, 2012.

[14] A. I. Frolov, S. V. Chankeshwara, Z. Abdulkarim, and G. M. Ghiandoni. pichemist free tool for the calculation of isoelectric points of modified peptides. *Journal of Chemical Information and Modeling*, 63(1):187–196, 2022.

[15] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel, and A. Bairoch. Expasy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic acids research*, 31(13):3784–3788, 2003.

[16] Gawehn and E. et al. Deep learning in drug discovery. *Molecular informatics*, 35:3–14, 2016.

[17] B. Ghosh and S. Choudhuri. Drug design for malaria with artificial intelligence (ai). *IntechOpen*, 2021, 2021.

[18] J. Guan, W. W. Qian, X. Peng, Y. Su, J. Peng, and J. Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. 2023.

[19] T. A. Halgren. Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of computational chemistry*, 17(5-6):490–519, 1996.

[20] H. C. Hayes, L. Y. P. Luk, and Y.-H. Tsai. Approaches for peptide and protein cyclisation. *Org Biomol Chem*, 19:3983–4001, 2021.

[21] J. D. Head and M. C. Zerner. A broydenfletchergoldfarbshanno optimization procedure for molecular geometries. *Chemical physics letters*, 122(3):264–270, 1985.

[22] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[23] J. H. Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. 1992.

[24] Jain and K. et al. Prevalence of pregnancy associated malaria in india. *Frontiers in global women's health*, 3, 2022.

[25] A. R. Jensen, Y. Adams, and L. Hviid. Cerebral plasmodium falciparum malaria: The role of pfemp1 in its pathogenesis and immunity, and pfemp1-based vaccines to prevent it. *Immunological reviews*, 293(1):230–252, 2020.

[26] A. R. Jensen, Y. Adams, and L. Hviid. Cerebral plasmodium falciparum malaria: The role of pfemp1 in its pathogenesis and immunity, and pfemp1based vaccines to prevent it. *Immunol Rev*, 293:230–252, 2020.

[27] G. K, R. BV, and P. MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng.*, 4:155–61, 1990.

[28] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pages 1942–1948, 1995.

[29] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[30] S. Kirkpatrick, C. D. Gelatt Jr, and M. P. Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[31] D. Kitchen and B. J. Decornez H, Furr JR. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov*, 3:935–49, 2004.

[32] J. Köhler, L. Klein, and F. Noé. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International conference on machine learning*, pages 5361–5370, 2020.

[33] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, 1982.

[34] X. Li, H. Chen, N. Bahamontes-Rosa, J. F. Kun, B. Traore, P. D. Crompton, and A. H. Chishti. Plasmodium falciparum signal peptide peptidase is a promising drug target against blood stage malaria. *Biochemical and Biophysical Research Communications*, 380(3):454–459, 2009.

[35] Madani and F. et al. Mechanisms of cellular uptake of cell-penetrating peptides. *Journal of biophysics*, 2011, 2011.

[36] F. Madani, S. Lindberg, Ü. Langel, S. Futaki, and A. Gräslund. Mechanisms of cellular uptake of cell-penetrating peptides. *Journal of biophysics*, 2011, 2011.

[37] Muttenthaler and M. et al. Trends in peptide drug discovery. *Nat Rev Drug Discov*, 20:309–325, 2021.

[38] P. P. Nugrahadi, W. L. Hinrichs, H. W. Frijlink, C. Schöneich, and C. Avanti. Designing formulation strategies for enhanced stability of therapeutic peptides in aqueous solutions: A review. *Pharmaceutics*, 15(3):935, 2023.

[39] S. Panda and G. Chandra. Physicochemical characterization and functional analysis of some snake venom toxin proteins and related non-toxin proteins of other chordates. *Bioinformation*, 8(18):891, 2012.

[40] N. D. Pasternak and R. Dzikowski. Pfemp1: an antigen that plays a key role in the pathogenicity and immune evasion of the malaria parasite plasmodium falciparum. *The international journal of biochemistry & cell biology*, 41(7):1463–1466, 2009.

[41] S. C. Penchala, M. R. Miller, A. Pal, J. Dong, N. R. Madadi, J. Xie, H. Joo, J. Tsai, P. Batoon, V. Samoshin, et al. A biomimetic approach for enhancing the in vivo half-life of peptides. *Nature chemical biology*, 11(10):793–798, 2015.

[42] Purcell and A. W. et al. More than one reason to rethink the use of peptides in vaccine design. *Nat Rev Drug Discov*, 6:404–14, 2007.

[43] E. Ramrez-Aportela, J. R. Lpez-Blanco, and P. Chacn. Frodock 2.0: fast proteinprotein docking server. *Bioinformatics*, 32(15):2386–2388, 2016.

[44] R. Rentzsch and B. Y. Renard. Docking small peptides remains a great challenge: an assessment using autodock vina. *Briefings in bioinformatics*, 16:1045–56, 2015.

[45] D. C. Restrepo-Posada, J. Carmona-Fonseca, and J. A. Cardona-Arias. Systematic review of microeconomic analysis of pregnancy-associated malaria. *Heliyon*, 6, 2020.

[46] A. Scherf, J. J. Lopez-Rubio, and L. Riviere. Antigenic variation in plasmodium falciparum. *Annu. Rev. Microbiol.*, 62:445–470, 2008.

[47] P. Schneider, W. Walters, and A. e. a. Plowright. Rethinking drug design in the artificial intelligence era. *Nat Rev Drug Discov*, 19:353–364, 2020.

[48] Shoichet and B. K. Virtual screening of chemical libraries. *Nature*, 432:862–5, 2004.

[49] S. Shukla, S. Choudhuri, G. P. Iragavarapu, , and B. Ghosh. Supervised learning of plasmodium falciparum life cycle stages using single-cell transcriptomes identifies crucial proteins. *Journal of Bioinformatics and Systems Biology*, 6, 2023.

[50] S. Shukla, S. Choudhuri, G. P. Iragavarapu, and B. Ghosh. Supervised learning of plasmodium falciparum life cycle stages using single-cell transcriptomes identifies crucial proteins. *Journal of Bioinformatics and Systems Biology*, 6:31–46, 2023.

[51] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[52] T. Tanimoto. *An Elementary Mathematical Theory of Classification and Prediction*. 1958.

[53] O. Trott and A. J. Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

[54] L. Wang, N. Wang, and W. Zhang. Therapeutic peptides: current applications and future directions. *Sig Transduct Target Ther*, 7, 2022.

[55] Z. Wang, L. Zheng, S. Wang, M. Lin, Z. Wang, A. W.-K. Kong, Y. Mu, Y. Wei, and W. Li. A fully differentiable ligand pose optimization framework guided by deep learning and a traditional scoring function. *Briefings in Bioinformatics*, 24(1):bbac520, 2023.

[56] 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, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[57] D. Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[58] Z. Wen, J. He, H. Tao, and S. Huang. Pepbdb: a comprehensive structural database of biological peptide-protein interactions. *Bioinformatics*, 35:175–177, 2019.

[59] M. Xu, L. Yu, Y. Song, C. Shi, S. Ermon, and J. Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

[60] L. Yiwen and D. P. A. Engineering of enzymes using non-natural amino acids. *Biosci Rep*, 42, 2022.

[61] Y. Yuan, J. Pei, and L. Lai. Binding site detection and druggability prediction of protein targets for structure-based drug design. *Curr Pharm Des*, 19:2326–33, 2013.

[62] X. Zeng, F. Wang, Y. Luo, S. gu Kang, J. Tang, F. C. Lightstone, E. F. Fang, W. Cornell, R. Nussinov, and F. Cheng. Deep generative molecular design reshapes drug discovery. *Cell Reports Medicine*, 3:100794, 2022.