Exploration of peptide structures and generation of therapeutic peptide candidates using Protein Language Model

Interdisciplinary Team Project

Mateusz Chojnacki, Younginn Park, Łukasz Milewski, Hubert Wąsiewicz

Team Warsaw 3

Contents

Abstract

The global COVID-19 pandemic caused by the SARS-CoV-2 virus has prompted extensive research to identify potential therapeutic targets for drug development. This project focuses on the nonstructural protein 13 (nsp13) helicase, a key player in the replication-transcription complex of the virus. With one of the largest viral genomes known, SARS-CoV-2 and its helicase has gathered attention for its potential vulnerabilities. Here, we explore the potential of therapeutic peptides targeting nsp13, leveraging existing knowledge from a limited pool of peptide drugs and candidates. We compile a dataset, conduct docking simulations, and identify binding sites, emphasizing the conserved regions crucial for the virus's replication process. Additionally, we employ ProtGPT2, a protein sequence generation model, to propose novel peptide sequences with desirable properties. Our findings lay the groundwork for further investigations into peptide-based therapies against SARS-CoV-2, providing insights into potential drug targets and expanding the spectrum of antiviral strategies.

1 Introduction

1.1 Coronavirus and the Nsp13 Helicase

SARS-CoV-2, the main culprit behind the global coronavirus (COVID-19) pandemic is an enveloped, positive sense single stranded RNA virus from genus *Betacoronavirus*, belonging to order *Nidovirales*. It has one of the largest viral genomes, which size is approximately 30kbp and is currently one among the best known viral genomes due to intensive studies aiming to find an effective drug. 2,131 of the 3,940 experimental structures of SARS-CoV-2 proteins that are currently available in the Protein Data Bank (PDB) are for the helicase nsp13. One of the most crucial proteins of Sars-Cov-2 is helicase nsp13. It forms the replication-transcription complex (RTC), which is necessary for the replication of all viral RNA molecules and permits the spread of the virus, by coupling with the RNA dependent RNA polymerase (RdRp) and binding the RNA strand. As a result, RTC, and hence nsp13, are promising targets for future therapeutics. There have been many previous studies on small molecules and antiviral agents, but the number of potential peptide drugs is limited, which is why we want to expand peptide drug discovery for the nsp13 helicase. Structure of nsp13 helicase is shown in Fig. 1

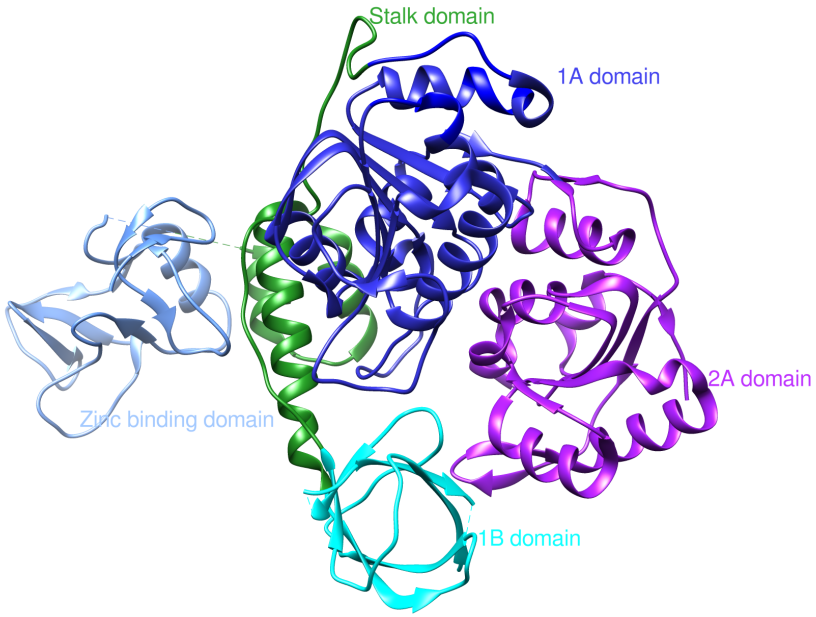

Figure 1: Structure of SARS-CoV-2 nsp13 helicase (PDB: 6ZSL).

1.2 Therapeutic Peptides

Therapeutic peptides represent a distinctive class of pharmacologically active compounds characterized by their smaller size compared to proteins. These peptides consist of short, organized aminoacids chains linked together by peptide bonds, typically ranging in size from 500 to 5000 Da[1]. Their compact structure enables them to interact with specific proteins, thereby exerting diverse therapeutic effects through various mechanisms, including modulation of protein-protein interactions, inhibition of enzymatic activity, and regulation of cellular signaling pathways By exploiting these mechanisms, therapeutic peptides hold immense potential for the treatment of a wide range of diseases, including cancer, metabolic disorders, and infectious diseases.

First peptides used as therapeutics were natural human hormones, such as insulin isolated in 1921 (Fig. 2). It became available to patients a year later, establishing itself as the first commercial peptide drug by 1923, replaced by recombinant insulin in the 1980s.

Throughout the 1990s, the global approval of 40 other natural peptide drugs underscored the growing significance of peptide therapeutics. Some of them are extracted from animal venomes, like exenatide (Fig. 2). Concurrently, the development of synthetic peptides, mimicking hormones like oxytocin and vasopressin alongside with new synthetic de novo designed peptides expanded the therapeutic repertoire, offering novel treatment options[1],

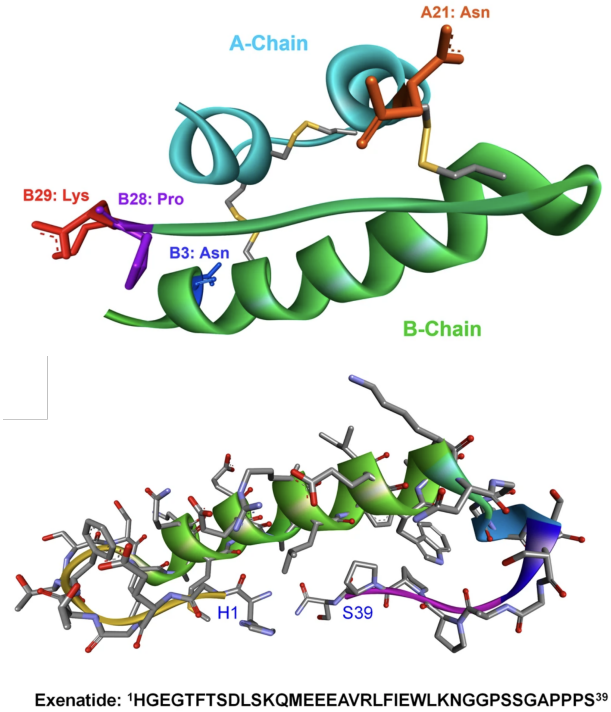

Figure 2: Left: Structure of human insulin (PDB: 1XDA) Right: exenatide optimized from Gila monster venom is a GLP-1 receptor agonist for chronic neuropathic pain treatment

Peptide–protein interactions are crucial in cellular functions, driving essential processes through short segments embedded within proteins. These interactions, accounting for 40% of cellular interactions, have broad applications in biotechnology and therapeutics, offering potential for targeted interventions.

Therapeutic peptides, acting as hormones, growth factors, neurotransmitters, or anti-infective agents, offer high specificity and affinity in binding to cell receptors, akin to biologics. Compared to biologics, peptides exhibit lower immunogenicity and production costs. However, they face challenges due to weak membrane permeability and poor in vivo stability, limiting their intracellular targeting potential. Peptide drug development requires addressing these limitations while capitalizing on their unique physiochemical properties for effective therapeutic interventions[1] (see Fig. 3)

2 Materials and Methods

2.1 Literature Review and Dataset Compilation

In our paper we decided to identify new peptide drugs using already known peptide drugs and candidates. To do so, we first had to construct a peptide base from peptide drugs found in previous studies. However, literature research made using the repository PubMed[2] found little to none papers containing studies upon the discussed subject. We found exactly one paper about peptide drugs targeting the SARS-CoV-2 helicase nps13[3], while there were few more targeting other coronavirus proteins. The paper contained 45 potential peptide drugs with lengths ranging from 5 to 13 amino acids, which we included in our peptide base and marked as peptides P1-P45. Nevertheless, 45 is a small number, therefore, we

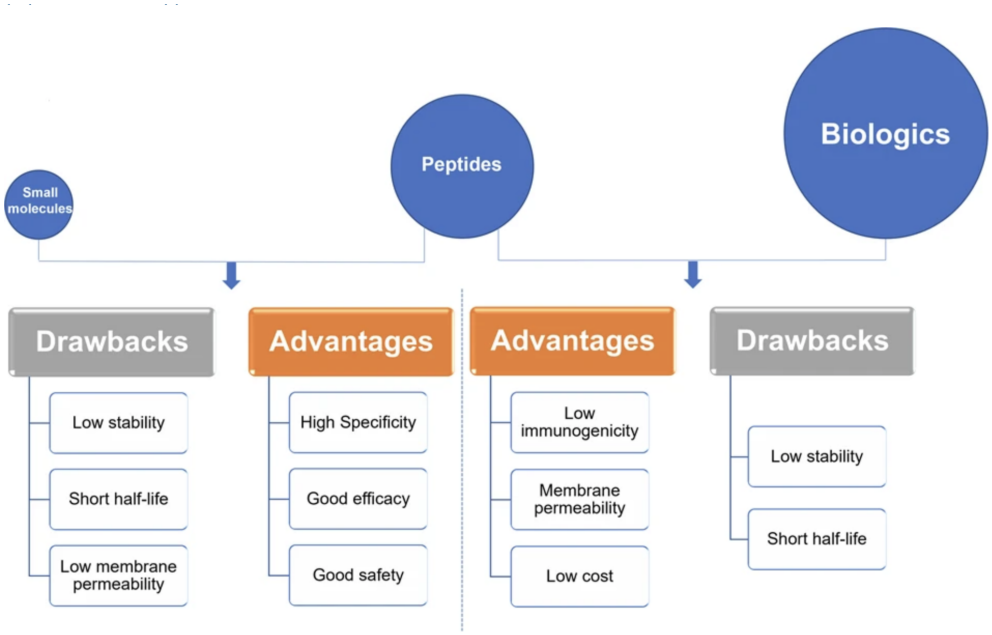

Figure 3: Peptides as therapeutic agents between small molecules and biologics in respect of molecular size. Below advantages and drawbacks of using peptides for therapy.

decided to expand our base with short peptides with high structural similarity to peptides P1-P45 found using advanced search options from the RCSB PDB database[4]. Using that method, we found another 27 unique peptides (which we marked K1-K27). Several had undetermined amino acids (X) on the C-end or N-end of their sequences, which were removed before docking procedure.

To confirm, that peptides included in our base actually bind with helicase nsp13, we used following docking algorithms: HPepDock[5] and CABS-dock[6]. Results showed that each of them is binding to at least one of the three potential binding sides, with almost everyone targeting first or/and second one. Detailed informations about nsp13 binding sites are described in the subsection 2.2.

2.2 Binding Sites of nsp13

The SARS-CoV-2 helicase nps13 is a multidomain protein from superfamily 1 helicase[7,8,9], which participates in unwinding of RNA/DNA strands in 5' to 3' direction, containing 5 domains. These domains starting from N-terminal are: ZBD - zinc-binding domain, S - stalk domain, 1B - β-domain, 1A - catalytic "RecA1 like" helicase domain and 2A - catalytic "RecA2 like" helicase domain (see Fig. 1. According to the previous studies[7,8] helicase nsp13 contains two binding sides important for replication/transcription process, so binding a therapeutic peptide there ought to result in termination of these processes. The first one, binding ATP, is located between 1A and 2A domains, the second one binding the 5 $'$-end of the substrate RNA is situated in the pocket between 1A, 2A and 1B domains. Amino acid sequences of these binding sites are strongly conserved through the coronavirus family[7], therefore they are good potential target sides for peptides tested in this paper. Binding sites are presented in Fig. 5.

There is also a third potential target pocket between ZBD, Stark and 1B domains, to which peptides from our dataset were often docked, and which was marked as a potential allosteric site, but due to lack of solid evidence from prior studies, we decided to focus more on the two sites mentioned earlier. We excluded the Zinc binding pocket as a potential target site, because it is too small to contain the whole peptide. Our test docking using HpepDock and CABS-dock programs confirmed this by showing almost no attachment of

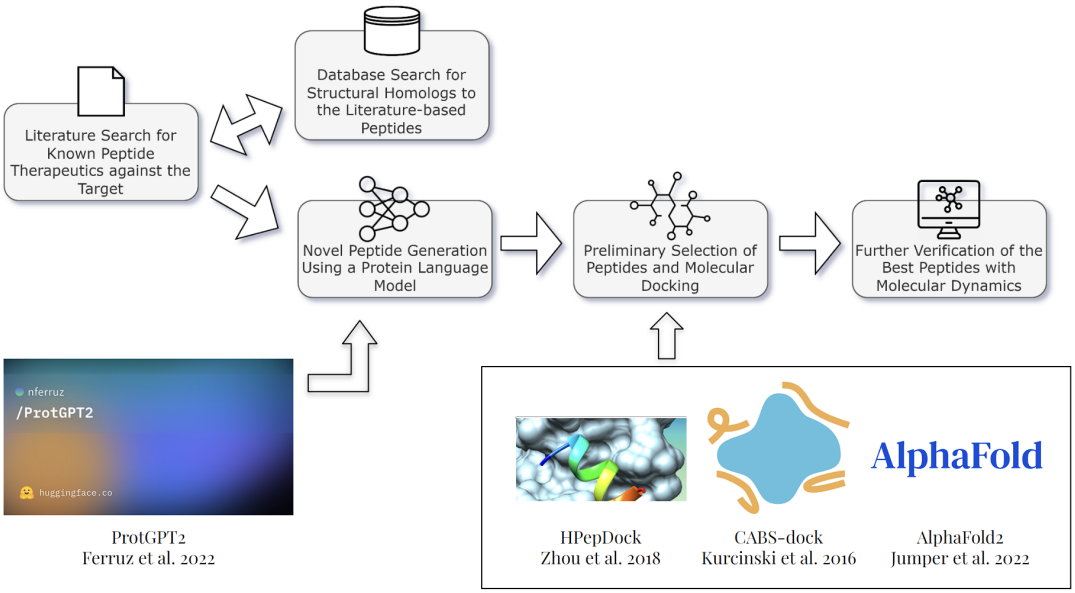

Figure 4: Outline of the project pipeline.

the peptides in that region among all of the top 10 docking results. These possible binding sites are shown in Fig. 5. Table 1 contains a list of residues in three binding sites described above.

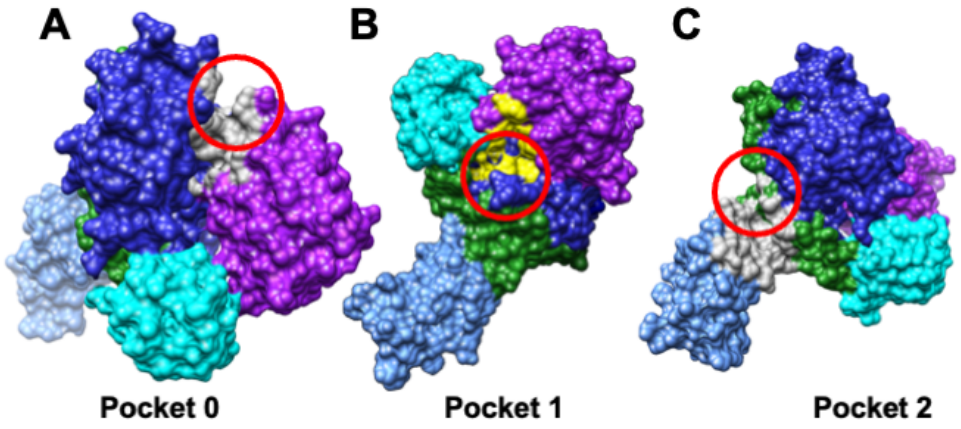

Figure 5: Possible peptide binding sites of nsp13 helicase. A) Pocket '0' involved in ATP binding, B) pocket '1' involved in binding 5 ′-end of the substrate RNA, C) pocket '1" which can be possible allosteric binding site

2.3 Sequence Generation with ProtGPT2

Protein sequence generation is the task of creating novel protein sequences that have desirable properties, such as folding stability, biochemical activity, or compatibility with a given structure. This task is challenging due to the vastness and complexity of protein sequence space, and the difficulty of evaluating the quality of generated sequences[10]. In recent years, there has been remarkable progress in natural language processing (NLP), largely driven by the emergence of large pre-trained language models. These models have not only transformed our interaction with everyday tools like chatbots and translation machines but have also inspired new applications in scientific domains. Drawing an analogy between protein sequences and human languages, amino acids form a chemically defined alphabet that assemble into structural elements that resemble "words" and functional domains comparable

to "sentences." Despite the nuanced differences, the information-completeness of protein sequences parallels natural languages, storing both structure and function with remarkable efficiency[11].

One noteworthy contribution to this area is the introduction of ProtGPT2, an autoregressive Transformer model with 738 million parameters, designed to generate *de novo* protein sequences at a high throughput. The model was trained on approximately 50 million non-annotated sequences spanning the entire known protein space[12]. ProtGPT2 sequences go into 'dark' areas of the protein space, expanding beyond natural superfamilies. The model's accessibility on standard workstations and its adaptability through fine-tuning on user-selected sequence sets makes it a valuable asset in the task of efficient protein engineering across biomedical and environmental sciences. The model, along with its datasets, is available on the HuggingFace repository[13]. Unfortunately, there was not enough data to fine-tune the model and give it a direction during the generation process without the risk of overfitting. Instead, an alternative method was implemented, where the pre-trained model was fed the known literature peptides, which then were treated as a 'context' for the model to base its generation. The model would then append these known peptide with new amino acid that are 'appropriate' given this 'context'.

The process of protein sequence generation using ProtGPT2 involved setting various parameters to tailor the output. The input served as the context, guiding the model, while max_length = 30 controlled the sequence length, counted in tokens, which are 4 amino acids long on average. The do_sample = True indicated random generation based on the model's probability distribution, and top_k = 950 determined the number of highest probability tokens considered during sampling. Repetition_penalty = 1.2 discouraged the model from repeating amino acids excessively. The number of generated sequences was controlled by num_return_sequences = 50, and eos_token_id = 0 indicated the end of the sequence.

In evaluating the generated sequences, some key metrics were applied (Figure 6). These included hydrophobicity measurements, which were calculated using the grand average of hydropathy (GRAVY)[14], assessing the balance between hydrophobic and hydrophilic properties of the amino acids in the chain. Metrics like instability index[15] and isoelectric point (pI) also provided crucial insights for drug design. For instance, any value of instability index above 40 is said to imply instability in a test tube, while the isoelectric point informs about the pH of a solution at which the net charge of a peptide becomes zero[16].

Table 1: Nsp13 residues involved in three binding sites.

Nr	**Name**	**Nsp13 residues**
1	ATP	E261, S264, N265, P284, G285, T286, G287, K288, S289, H290, K320, E375, Q404, L438, R442, R443, G538, E540, R567
2	nt	N177, R178, N179, Y180, H230, M233, H311, P335, A362, N361, L363, M378, R390, L405, P406, P408, R409, T410, L412, L417, H482, S485, S486, P514, Y515, N516, T532, D534, S535, Q537, H554, R560
3	Zn/stalk	A1, V2, G3, A4, C5, V6, N9, R15, I20, R21, R22, P23, F24, R129, F133, E136, P234, L235, S236, P238

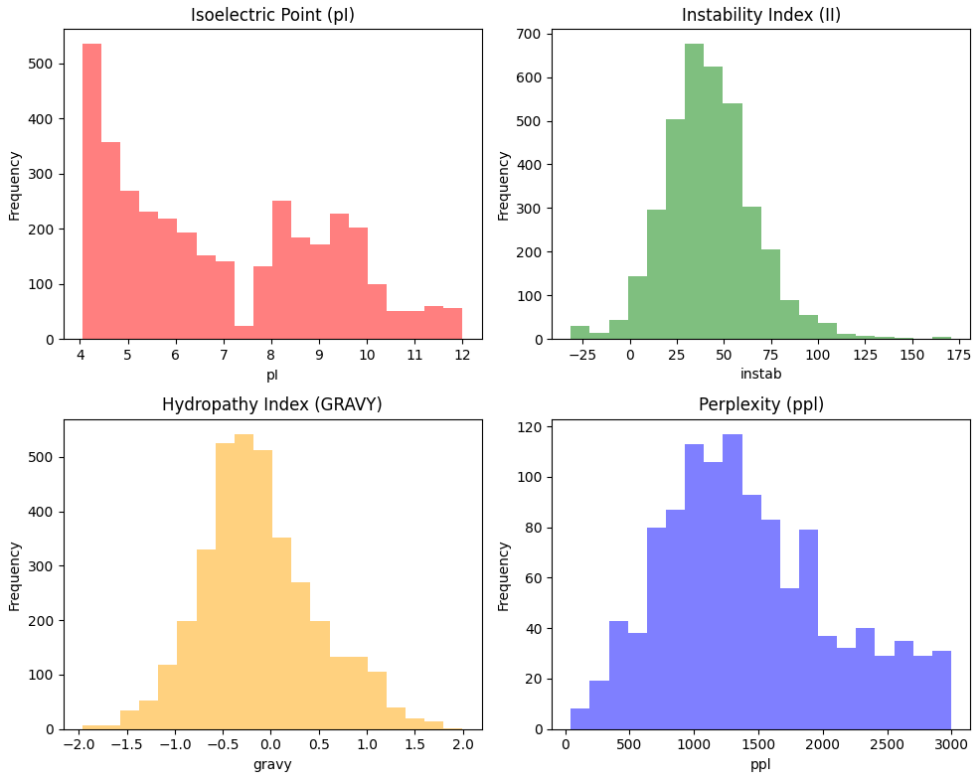

Figure 6: Histograms of preliminary metrics done on the generated sequences.

2.4 Docking Simulation with AlphaFold 2 Multimer

Peptide-protein interactions present significant challenges in computational modeling due to their fragility, transient nature, and contextual dependence. Understanding the binding processes and identifying hotspot residues necessitate knowledge of the three-dimensional structures of these complexes. Moreover, these structures serve as templates for designing stable peptidomimetics. However, the lack of known peptide structures poses a formidable obstacle, particularly in blind docking simulations, unlike traditional domain-domain docking, where specific domain structures are often available. Therefore, gaining insights into peptide conformational preferences is crucial for successful peptide-protein interaction research and design[17]. Recent advancements in protein structure prediction, especially employing deep learning neural network architectures, show promising outcomes. Notably, the remarkable accuracy demonstrated in tests such as CASP14, particularly with AlphaFold2 by Google DeepMind[18], offers potential solutions to the peptide-protein docking problem.

Our proposed approach views peptide-protein binding as the final step in protein folding, where functional proteins can be experimentally reconstituted from short fragments, highlighting the importance of non-covalent linkages in monomer folding. Successful modeling of peptide-protein interactions entails identifying complementary fragments in monomer structures and protein-protein interfaces, integrating information from folded monomer structures into the peptide docking search space.In neural network based peptide-protein docking simulations, adopting a global approach that mimics protein folding dynamics and utilizes neural networks trained to predict monomeric protein structures is essential. By connecting the peptide to the receptor, neural networks for monomer folding can generate accurate peptide-protein complex structures. This capability is facilitated by AlphaFold2's accurate identification and modeling of unstructured regions, treating them as extended linkers, and predicting peptide-receptor complexes without requiring a multiple sequence alignment for the peptide partner.

The integration of AlphaFold2 Multimer into our docking simulation framework represents a significant advancement in elucidating the structural intricacies of peptide-protein interactions and holds promise for accelerating drug discovery and design efforts. Leveraging the predictive power of AlphaFold2 Multimer allows us to overcome the challenges associated with blind docking simulations and provides valuable insights into the complex interplay between peptides and proteins. This approach not only enhances our understanding of biological systems but also offers new opportunities for rational drug design and therapeutic development.

2.5 Docking simulations with HPepDock and CABS-dock

In this work we compare Alphafold docking protocole and results with the another two software suites, namely HPepDock[5] and CABS-dock[6], which were also used to test our database.

To address peptide flexibility in docking, a hierarchical algorithm named HPEPDOCK has been developed. This algorithm efficiently models peptide conformations and globally samples binding orientations. Peptide flexibility is considered by generating an ensemble of conformations. These conformations are then docked against the protein using rigid docking protocol. HPEPDOCK generates up to 1000 conformations, allowing for better flexibility consideration. It is also effective for local peptide docking when binding site information is known. Recent participation in CAPRI experiments has demonstrated HPEPDOCK's strong performance in peptide-docking challenges[5].

On the other hand, CABS-dock method offers a unified approach to address conformational changes during binding. Utilizing a coarse-grained simulation of protein dynamics, CABS-dock employs a reduced representation of the protein chain and a knowledge-based force-field derived from statistical potentials based on known protein structures. The protein is represented by pseudoatoms, including alpha and beta carbons, side chains, and a geometric center for hydrogen bond definition. The docking process involves flexible docking, initial filtering, structural clustering, and all-atom model reconstruction, resulting in top-ranked models suitable for further analysis. Notably, structural clustering identifies dominant conformations, facilitating the selection of representative models. During the process, thousands of models are generated, which are subsequently filtered to retain a subset for further analysis, typically around 1000 models. This subset undergoes structural clustering to identify distinct conformations, ultimately yielding a final selection of around 10 representative models. The CABS-dock server provides an interface for protein-peptide docking, offering up-to-date documentation and benchmark examples, making it a valuable tool for understanding complex biological processes through comprehensive exploration of binding interactions.[6].

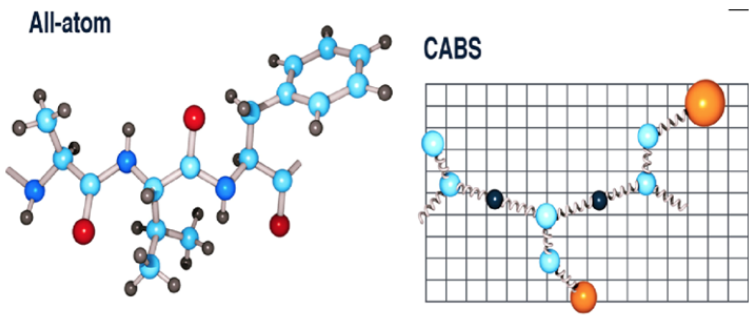

Figure 7: In the CABS model, a single residue is represented by two atoms (alpha and beta carbon, colored in black) and two pseudo-atoms (side chain, colored in orange, and center of the peptide bond, colored in green)

3 Results

3.1 Novel Peptide Sequence Generation

During the sequence filtering process, specific criteria were implemented to ensure the selection of high-quality sequences. The GRAVY (Grand Average of Hydropathicity) values were capped within the range of -1 and 1, to ensure good pharmacokinetic properties by maintaining a balance between hydrophobicity, preventing substance accumulation in fatty tissues and being toxic to humans, and hydrophilicity, which might cause easy dissolution in blood and excretion. Moreover, to emulate physiological conditions, the pH values of the sequences were closed to the pH 6-8 range, mirroring the typical pH range of blood.

Finally the perplexity metric (*ppl*) was used as a measure of the quality of generated sequences. In the context of protein generation, perplexity measures the model's ability to generate coherent amino acid sequences similar to those found in natural proteins given an input context. Although there is no standard threshold for what perplexity value yields a 'good' or 'bad' sequence, the approach here involves sampling numerous sequences (50 for each of the 72 input peptides in our case), ordering them by perplexity, and selecting those with lower values, as lower perplexity is generally preferred for higher quality sequences correlating later with AlphaFold's confidence level called pLDDT. At the end of this task, 20 sequences were selected for further analyses (Table 2).

3.2 Peptide-Protein Docking

In comparing various docking methods, including HpepDock, CABS-dock, and Alphafold, it becomes evident that the choice of method significantly influences the outcome of peptide docking experiments. Building upon the previous analysis of peptide docking, particularly focusing on peptides from group G, we delve into the specifics of each method's performance. Both HpepDock and CABS-dock exhibited successful docking of G peptides to at least two binding sites (sites 1 and 2). For every G peptide, number of models docked to specific binding pockets is shown in Table 1. In Fig. 8 is shown example of CABS-dock docking result for G13 peptide, which part is in the binding pocket '0' (ATP binding pocket, as it is described in section 2.2).
This outcome underscores the reliability and versatility of these methods in accommodating different peptide conformations and binding scenarios. However, the results took a different turn when employing Alphafold 2 Multimer for docking.

Despite its renowned capabilities in protein structure prediction, Alphafold encountered

Table 2: Sequences generated with ProtGPT2 with selected preliminary metrics (pI - isoelectric point, II - instability index, gravy - grand average of hydropathicity index, ppl - perplexity)

Alias	**pI**	**II**	**gravy**	**ppl**	**Sequence**
G1	6.46	21.70	-0.48	844.69	SLPYPFIWGNQMWMLTWPDHR
G2	6.74	32.63	-0.56	868.12	HMWPGDIKPAAVSRDLSQ
G3	6.92	27.30	0.21	904.70	IIVTQTMKSGDVSVILHQIHYKAD
G4	6.06	19.66	-0.38	1007.95	WNPADYGGIKPLLTETNIVGKY
G5	7.84	25.80	-0.41	1020.43	GCCSDPLCAWRCHAGRCGRD
G6	7.94	35.50	0.74	1063.46	CKFFWATYTSCCLSGGNLGIFVPS
G7	6.22	18.45	-0.37	1089.33	LSITENGEFKPLGFQFSQKSIEKV
G8	6.77	29.40	0.18	1100.54	LVGPTIWRAALLESAPRHAAE
G9	7.82	11.31	-0.03	1200.32	GCCSDPRCAWRCYGCLS
G10	6.80	35.61	0.29	1287.75	ALKIPISKIYIDSHSVLSPE
G11	6.75	35.87	0.02	1371.14	LHTPLPLTRRDKALLDDALSLFG
G12	6.21	39.74	-0.47	1400.99	GWLEPLLARPWLIVGRDQRGVMTRPYDEG
G13	6.91	14.87	-0.71	1567.13	HEGFTSDFRNPQHAFGSLMCRFNT
G14	7.02	27.67	-0.31	1689.99	LTFQHNFQTHRGHEVGSAQGFTAILW
G15	6.05	34.96	0.60	1731.80	YCKFEWATFAKSCAFPVDGLSFPFFGI
G16	6.00	33.07	-0.33	1800.79	QIPTVNNLKVSEPFTP
G17	6.12	6.10	-0.03	1831.41	GLDIQKVKDMEQLLTQVRLSI
G18	6.74	27.94	-0.04	1927.21	VLEKYKDVIMNSSSLLEHIATGIKKFE
G19	6.40	3.73	-0.26	1964.08	TLPFHSVIYVDSATGQTWTGNR
G20	6.21	37.61	-0.89	2220.56	GYDPETGTWGRRMTLFTPDSRAEVAAR

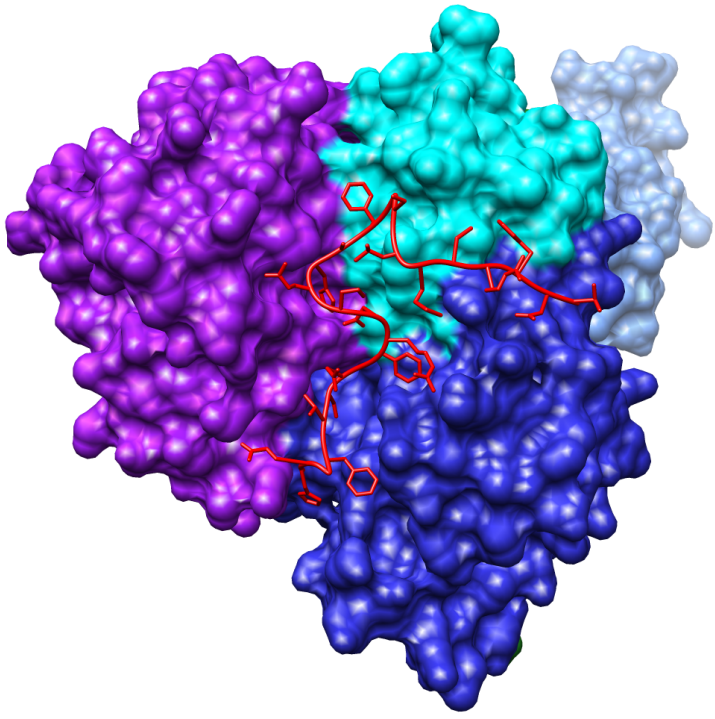

Figure 8: Visulization of G13 peptide docked to ATP binding pocket '0' in CABS-dock software.

Table 3: Number of G peptide models that docked to specific binding pockets (described in section 2.2) from docking runs from HPepDock and CABS-dock to helicase 6zsl with 10 results with the best binding score.

Alias	HpepDock			CABS-dock		
	0	1	2	0	1	2
G1		1	6	3	5	1
G2	1	3	3	4	2	1
G3	1	1	8	7	1	
G4	1	1	4	1	3	4
G5	2	3	4	1	6	1
G6	2	1	7	2	3	5
G7		2	6	5	3	
G8		1	7	4	4	
G9	2	1	7	6	1	
G10		3	3	4	4	1
G11		1	8	4	4	1
G12		3	5	6	4	
G13	1	3	6	8		
G14	2	3	3	3	2	1
G15		2	6	3	5	1
G16	2	1	7	4	2	
G17	1	1	7	3	2	2
G18	1		4	4	3	1
G19		1	8	5	5	
G20	1	2	7	8	2	

challenges in docking G peptides effectively. None of the peptides were docked to any of the three pockets, and the predicted plDDT values ranged disappointingly between 10-20%. This discrepancy in performance raises intriguing questions about the suitability of Alphafold for peptide docking tasks, especially when dealing with shorter peptide sequences.

The discrepancy in performance among the methods prompts further investigation into the underlying factors influencing their efficacy. While HpepDock and CABS-dock demonstrated proficiency in handling the peptides' structural complexities, Alphafold's limitations in accurately predicting peptide folding could be attributed to several factors. One plausible explanation is the inherent difficulty in simulating the folding dynamics of short peptides within the constraints of Alphafold's algorithms. Thus, while the outcomes of docking experiments may vary depending on the selected method, it's essential to consider the method's strengths and limitations in the context of the specific peptide sequences and binding scenarios under investigation. This comparative analysis highlights the importance of employing a diverse array of computational tools and methodologies to gain a comprehensive understanding of peptide-protein interactions.

4 Discussion

Our designed peptides were evaluated based on their ability to dock onto the SARS-CoV-2 nsp13 helicase protein using both Alphafold and CABS-dock. Despite our efforts to design peptides with potential therapeutic properties, we observed varying degrees of success in their docking capabilities.

When assessed using Alphafold, our designed peptides exhibited challenges in effectively docking onto the target protein. Alphafold's predictions suggested that the designed pep-

tides struggled to bind to the intended binding sites on the nsp13 helicase protein, indicating potential limitations in their structural compatibility.

However, when subjected to docking simulations with CABS-dock, our designed peptides showed more promising results. CABS-dock's ability to generate multiple docking clusters allowed for a more comprehensive exploration of peptide-protein interactions. This led to the identification of docking clusters indicating successful binding of our designed peptides to specific pockets on the nsp13 helicase protein. The dominance of clusters docking into a single pocket typically indicates that >600-700 out of 1000 simulations successfully matched a peptide there.

Furthermore, when considering Hpepdock, it's essential to acknowledge its approach of rigidly attaching to the protein, a methodology that introduces a single inaccuracy. This rigidity tends to favor regions that are both easier to access and more exposed on the protein surface. Consequently, it often results in a bias towards docking in these prominent areas, potentially overlooking binding sites that may be less accessible or buried within the protein structure. This preference for exposed sites can lead to certain portions of the peptide chain extending beyond the boundaries of the helicase. It's noteworthy that Hpepdock's rigid docking approach may cause a portion of the peptide chain to protrude beyond the helicase structure. This occurrence can have implications for the accuracy of the docking predictions, as it may result in the neglect of potential binding sites that are deeply embedded within the protein structure.

Overall, although our designed peptides encountered difficulties in docking onto the target protein when assessed using Alphafold, the results obtained from CABS-dock were notably more promising. Furthermore, considering the inherent limitations of Hpepdock, which favors docking onto exposed regions of rigid proteins, it became evident that relying solely on a single docking tool may lead to biased results. These findings underscore the critical importance of employing a diverse range of computational tools for peptide design and evaluation. Specifically, CABS-dock emerges as a particularly valuable resource for peptide docking studies, offering robust and reliable predictions that complement the insights garnered from other tools such as Alphafold and Hpepdock. This integrative approach enhances the accuracy and comprehensiveness of peptide-protein interaction analyses, thereby facilitating advancements in peptide-based drug discovery and design.

Bibliography

[1] L. Wang, N. Wang, W. Zhang, X. Cheng, Z. Yan, G. Shao, X. Wang, R. Wang, and C. Fu, "Therapeutic peptides: current applications and future directions," *Signal Transduction and Targeted Therapy*, vol. 7, no. 1, 2 2022. [Online]. Available: http://dx.doi.org/10.1038/s41392-022-00904-4

[2] "PubMed." [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/

[3] H. Pradeep, U. Najma, and H. S. Aparna, "Milk Peptides as Novel Multi-Targeted Therapeutic Candidates for SARS-CoV2," *Protein J*, vol. 40, no. 3, pp. 310–327, Jun. 2021. [Online]. Available: https://doi.org/10.1007/s10930-021-09983-8

[4] "RCSB PDB." [Online]. Available: https://www.rcsb.org/

[5] P. Zhou, B. Jin, H. Li, and S.-Y. Huang, "HPEPDOCK: a web server for blind peptide–protein docking based on a hierarchical algorithm," *Nucleic Acids Research*, vol. 46, no. W1, pp. W443–W450, Jul. 2018. [Online]. Available: https://doi.org/10.1093/nar/gky357

[6] M. Blaszczyk, M. P. Ciemny, A. Kolinski, M. Kurcinski, and S. Kmiecik, "Protein–peptide docking using CABS-dock and contact information," *Briefings in Bioinformatics*, vol. 20, no. 6, pp. 2299–2305, Nov. 2019. [Online]. Available: https://doi.org/10.1093/bib/bby080

[7] J. A. Newman, A. Douangamath, S. Yadzani, Y. Yosaatmadja, A. Aimon, J. Brandão-Neto, L. Dunnett, T. Gorrie-stone, R. Skyner, D. Fearon, M. Schapira, F. von Delft, and O. Gileadi, "Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase," *Nat Commun*, vol. 12, no. 1, p. 4848, Aug. 2021, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-021-25166-6

[8] J. Chen, Q. Wang, B. Malone, E. Llewellyn, Y. Pechersky, K. Maruthi, E. T. Eng, J. K. Perry, E. A. Campbell, D. E. Shaw, and S. A. Darst, "Ensemble cryo-EM reveals conformational states of the nsp13 helicase in the SARS-CoV-2 helicase replication–transcription complex," *Nat Struct Mol Biol*, vol. 29, no. 3, pp. 250–260, Mar. 2022, number: 3 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41594-022-00734-6

[9] K. J. Mickolajczyk, P. M. M. Shelton, M. Grasso, X. Cao, S. E. Warrington, A. Aher, S. Liu, and T. M. Kapoor, "Force-dependent stimulation of RNA unwinding by SARS-CoV-2 nsp13 helicase," *Biophys J*, vol. 120, no. 6, pp. 1020–1030, Mar. 2021.

[10] N. Ferruz and B. Höcker, "Controllable protein design with language models," *Nat Mach Intell*, vol. 4, no. 6, pp. 521–532, Jun. 2022, number: 6 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s42256-022-00499-z

[11] N. Ferruz, S. Schmidt, and B. Höcker, "ProtGPT2 is a deep unsupervised language model for protein design," *Nat Commun*, vol. 13, no. 1, p. 4348, Jul. 2022, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-022-32007-7

[12] "nferruz/UR50_2021_04." [Online]. Available: https://huggingface.co/datasets/nferruz/UR50_2021_04

[13] "nferruz/ProtGPT2." [Online]. Available: https://huggingface.co/nferruz/ProtGPT2

[14] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105–132, May 1982. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022283682905150

[15] K. Guruprasad, B. Reddy, and M. W. Pandit, "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 Engineering, Design and Selection*, vol. 4, no. 2, pp. 155–161, Dec. 1990. [Online]. Available: https://doi.org/10.1093/protein/4.2.155

[16] C.-H. Shen, "Chapter 8 - Extraction and purification of proteins," in *Diagnostic Molecular Biology (Second Edition)*, C.-H. Shen, Ed. Academic Press, Jan. 2023, pp. 209–229. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323917889000077

[17] 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, 2022. [Online]. Available: http://dx.doi.org/10.1038/s41467-021-27838-9

[18] 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, 7 2021. [Online]. Available: http://dx.doi.org/10.1038/s41586-021-03819-2