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. To date, Protein Data Bank (PDB) contains 3,940 experimental structures of SARS-CoV-2 proteins, including 2,131 experimental structures of helicase nsp13. Helicase nsp13 is one of the most important proteins of Sars-Cov-2. It couples with the RNA dependent RNA polymerase (RdRp) and binds the RNA strand to make the replication-transcription complex (RTC), which is essential in replicating all viral RNA molecules, thus allowing its propagation. Due to that RTC and consequently nsp13 is a good target for potential drugs. There are many previous papers about small molecules and antiviral agents, but the number of potential peptide drugs is limited.

1.2 Therapeutic Peptides

2 Materials and Methods

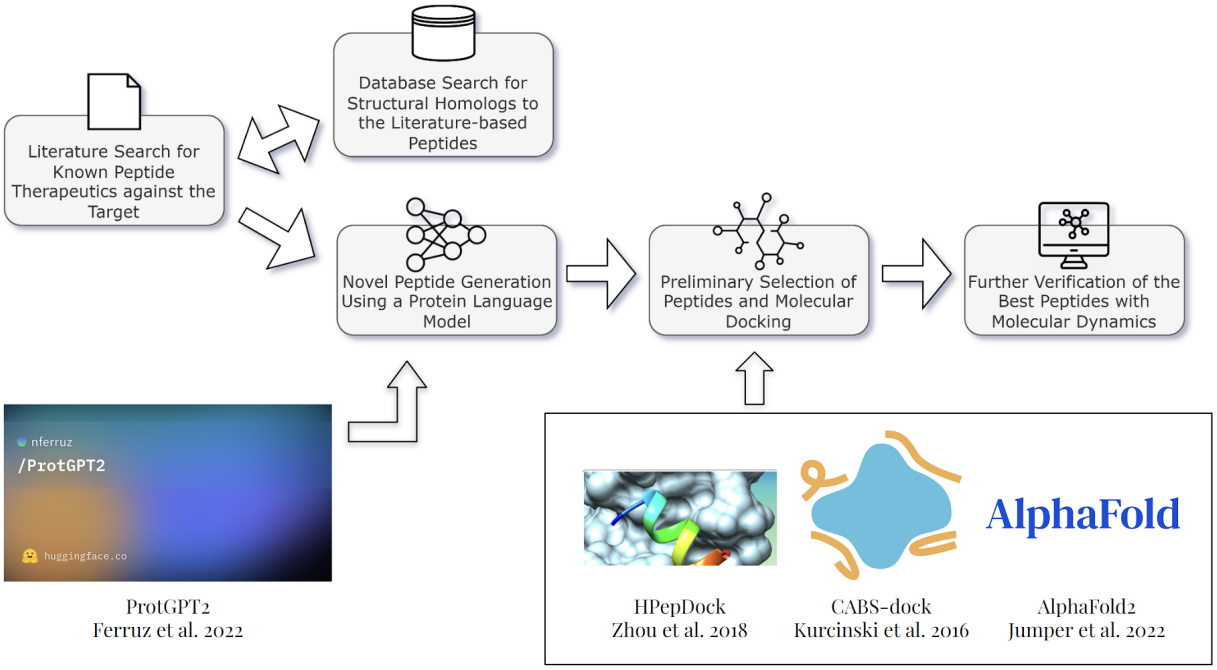

Figure 1: Outline of the project pipeline.

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[1] 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[2], 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 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[3]. Using that method, we found another 27 unique peptides (which we marked K1-K27). Several had

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

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[4] and CABS-dock[5]. 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[6],[7],[8], 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. According to the previous studies[6],[7] 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[6], therefore they are good potential target sides for peptides tested in this paper.

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

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[9]. 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[10].

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[11]. 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[12]. 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 2). These included hydrophobicity measurements, which were calculated using the grand average of hydropathy (GRAVY)[13], assessing the balance between hydrophobic and hydrophilic properties of the amino acids in the chain. Metrics like instability index[14] 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[15].

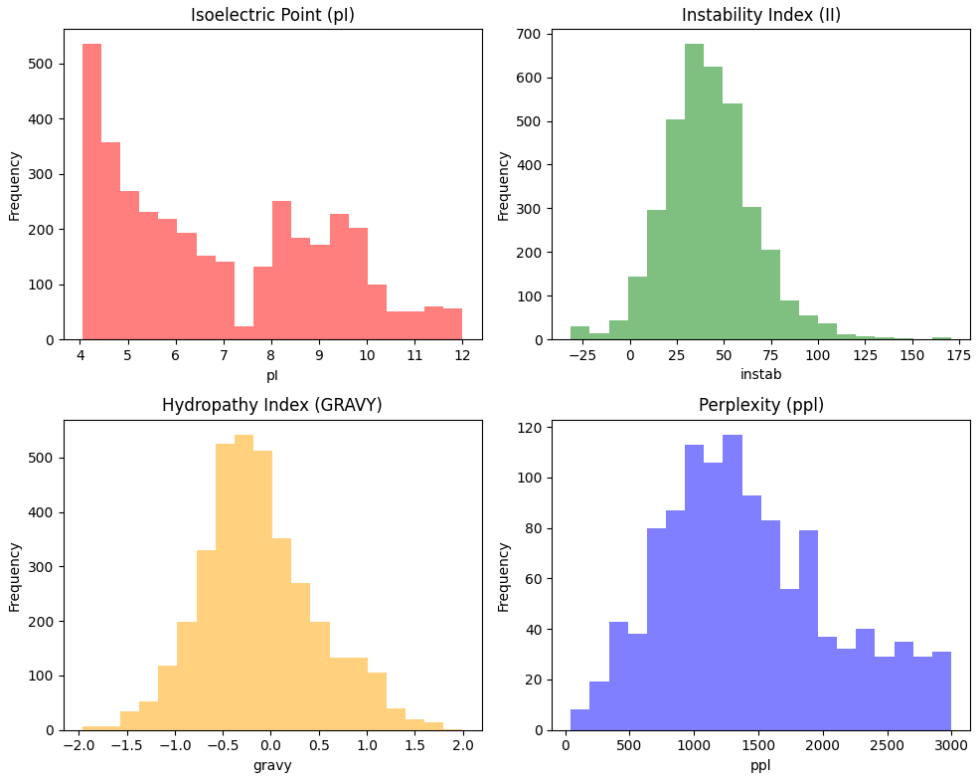

Figure 2: Histograms of preliminary metrics done on the generated sequences.

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

2.4 Docking Simulation with AlphaFold

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 Docking Methods Comparison

4 Discussion

Bibliography

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

[2] H. Pradeep, U. Najma, and H. S. Aparna, "Milk Peptides as Novel Multi-Targeted Therapeutic

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	

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

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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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.

[9] 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

[10] 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

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

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

[13] 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

[14] 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

[15] 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