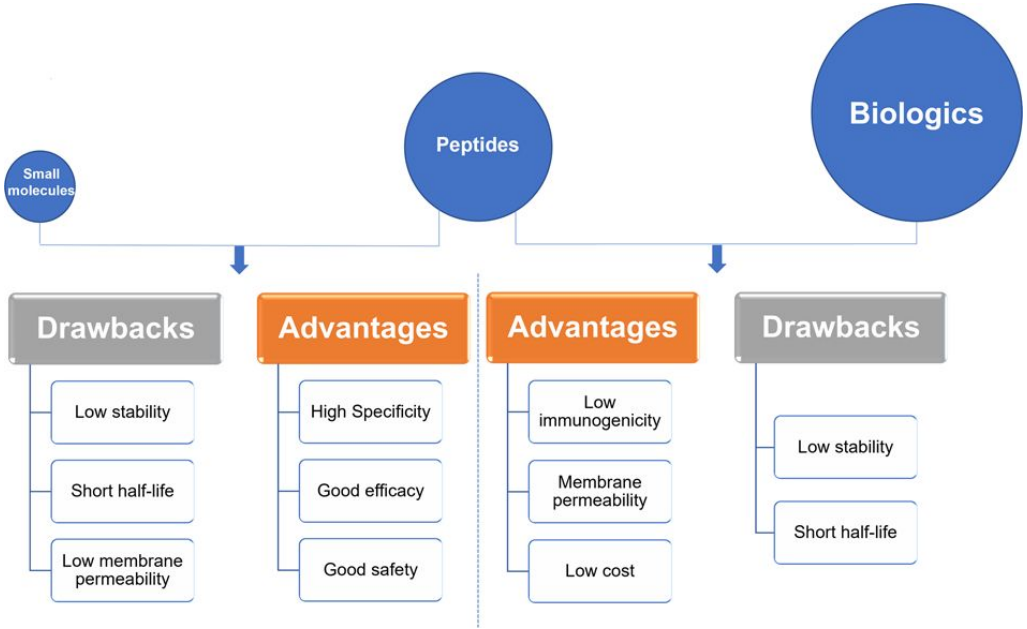
Exploration of peptide structures and generation of therapeutic peptide candidates using Protein Language Model

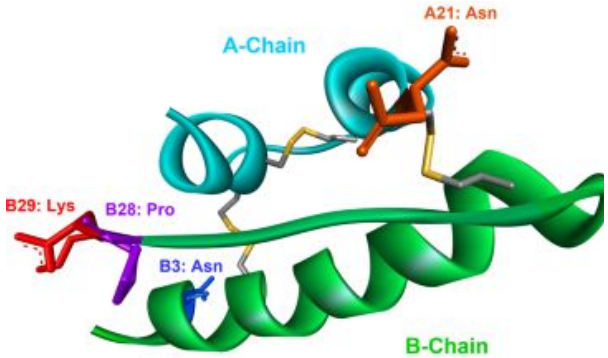
Mateusz Chojnacki, Paweł Park, Łukasz Milewski, Hubert Wąsiewicz

Under the Supervision of:
Wanda Niemyska PhD, prof. Joanna Sułkowska

Peptide therapeutics



Wang et al. 2022



Human insulin

SARS-CoV-2 nsp13 helicase

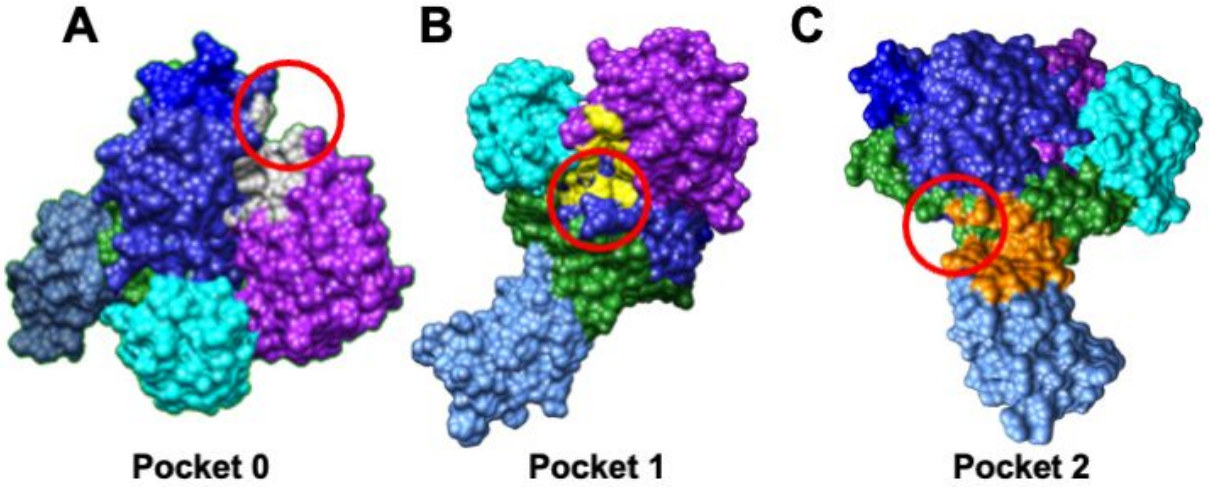
› Nat Commun. 2021 Aug 11;12(1):4848. doi: 10.1038/s41467-021-25166-6.

Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase

Joseph A Newman [1], Alice Douangamath [2], Setayesh Yadzani [3], Yuliana Yosaatmadja [4], Antony Aimon [2], José Brandão-Neto [2], Louise Dunnett [2], Tyler Gorrie-Stone [2], Rachael Skyner [2], Daren Fearon [2], Matthieu Schapira [3 5], Frank von Delft [4 2 6], Opher Gileadi [4]

Affiliations + expand

PMID: 34381037 PMCID: PMC8358061 DOI: 10.1038/s41467-021-25166-6



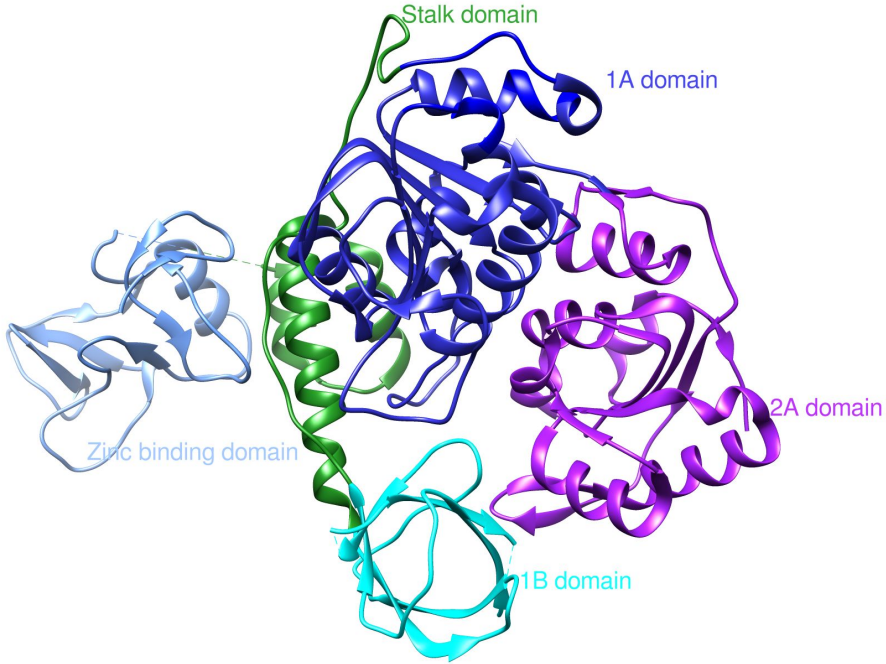
ATP binding

nt binding

Potential allosteric binding sites

Our Goal

Explore the space of possible peptide structures using a protein language model and attempt to generate plausible therapeutic peptide candidates against the nsp13 helicase



SARS-CoV-2 Nsp13 helicase

Project Pipeline

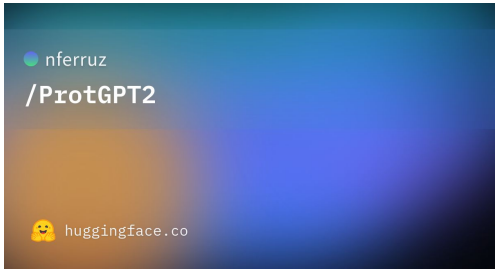
Literature Search for Known Peptide Therapeutics against the Target

Database Search for Structural Homologs to the Literature-based Peptides

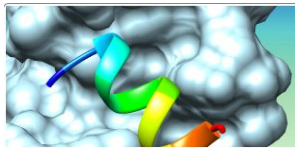
Novel Peptide Generation Using a Protein Language Model

Preliminary Selection of Peptides and Molecular Docking

Further Verification of the Best Peptides with Molecular Dynamics



ProtGPT2
Ferruz et al. 2022



HPepDock
Zhou et al. 2018

CABS-dock
Kurcinski et al. 2016

AlphaFold

AlphaFold2
Jumper et al. 2022

1) Peptides from the Literature Review

Almost no research on antiviral peptides against SARS CoV-2's nsp13

The only publication includes 45 peptides potentially binding to nsp13 helicase.

Protein J. 2021; 40(3): 310–327.
Published online 2021 Apr 11. doi: 10.1007/s10930-021-09983-8
PMCID: PMC8036162
PMID: 33840006

Milk Peptides as Novel Multi-Targeted Therapeutic Candidates for SARS-CoV2

H. Pradeep, Umme Najma, and H. S. Aparna

Sequence	Dockscore	E-model
SVFSGYRK	-12.95	-191.21
CLANGMIMY	-12.50	-172.26
RQQPGKGPRY	-12.26	-139.90
ALEATCKSL	-11.73	-153.84
LDAQSAPLRV	-11.59	-165.69
FLRQNEVL	-11.28	-156.53
IDALNENK	-11.03	-159.33
DIEQLRSQL	-10.91	-161.25
IQKVAGTW	-10.74	-124.27

2) Construction of Peptide Database

The structure homology search yielded 27 more peptides.

Docking with Hpepdock and CABS-dock showed, that they bind to at least one of described binding sites with high scores.

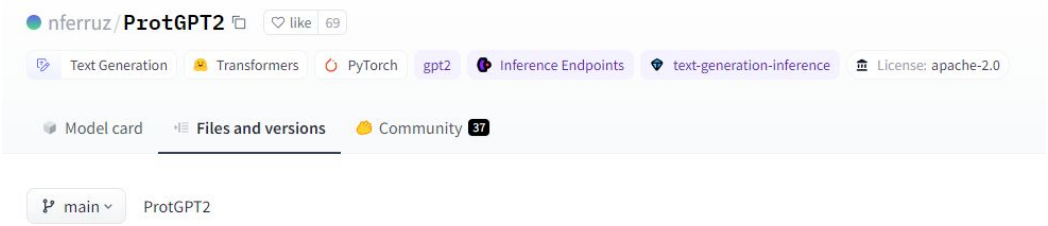
https://www.rcsb.org/

3) Generating Peptides

ProtGPT2 speaks the ‘language of the proteins’

Problems with fine-tuning - not enough data to generate completely novel peptides without overfitting

The literature peptides are fed into the pre-trained model, treated as ‘context’ and new amino acids are appended to the known peptide



https://huggingface.co/nferruz/ProtGPT2

Ferruz et al. 2022

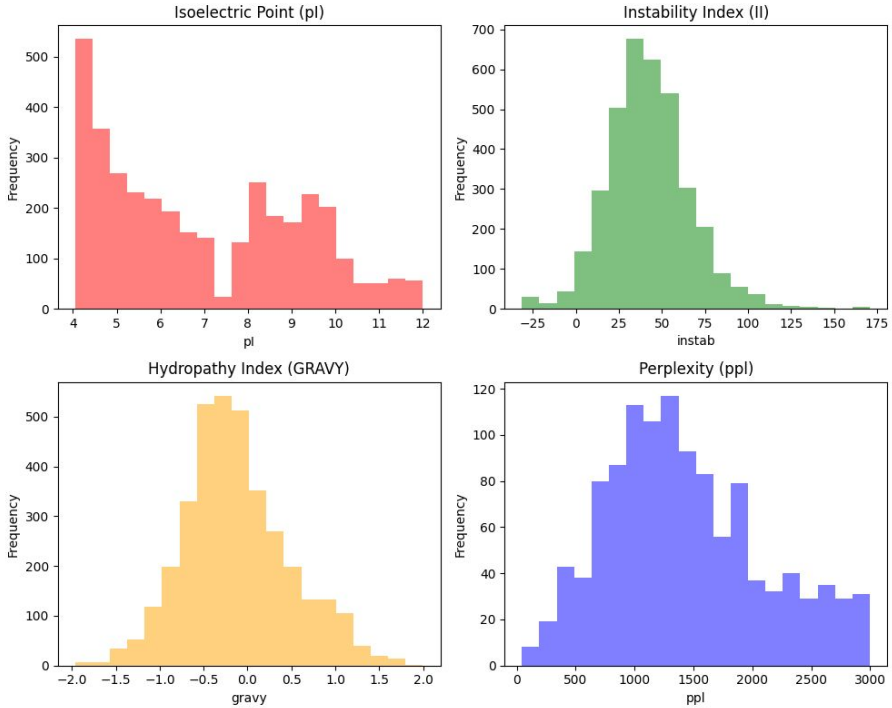
4) Preliminary Filtering

Isoelectric point - 6-8 to mimic physiological conditions

Instability index - anything below 40 can be considered stable (Guruprasad et al 1990)

GRAVY - balance between hydrophilic and lipophilic (-1, 1)

Perplexity - predictive performance of a language model (the lower the better)

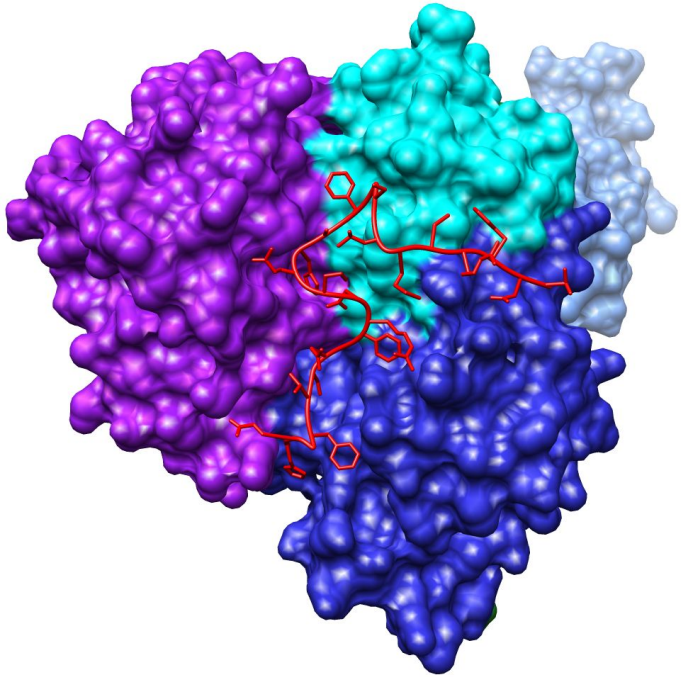


Histogram of the preliminary metrics used to filter out undesirable peptides

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

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

5) Docking Results



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	

Number of G peptide models that docked to specific binding pockets (0 - ATP binding pocket, 1 - RNA binding pocket, 2 - unnamed potential pocket between Zinc/stalk domains) from docking runs performed using HPepDock and CABS-dock to helicase 6zsl with 10 results with the best binding score.