

Appendix to: Tuning ProteinMPNN to reduce protein visibility via MHC Class I through direct preference optimization

HANS-CHRISTOF GASSER^{*,1}

DIEGO OYARZUN^{1,2}

JAVIER ALFARO^{*,**,1,3,4,5}

AJITHA RAJAN^{*,**,1}

¹School of Informatics, University of Edinburgh, Edinburgh, UK

²School of Biological Sciences, University of Edinburgh, Edinburgh, UK

³International Centre for Cancer Vaccine Science, University of Gdańsk, Gdańsk, Poland

⁴Department of Biochemistry and Microbiology, University of Victoria, Victoria, Canada

⁵The Canadian Association for Responsible AI in Medicine, Victoria, Canada

*Correspondence to: h.gasser@sms.ed.ac.uk, javier.alfaro@proteogenomics.ca, arajan@ed.ac.uk

**co-directed equally

March 2025

Acronyms

AA amino acid
Ab antibody
AR auto-regressive
CTL Cytotoxic T-lymphocyte
DPO Direct Preference Optimization
GAN Generative Adversarial Network
LLM large language model
LM language model
MD Molecular Dynamics
MHC-II MHC Class II
MHC-I MHC Class I
ML machine learning
MPNN message passing neural network
NLP Natural Language Processing
PDB Protein Data Bank
pI isoelectric point
PPO Proximal Policy Optimization
PWM position weight matrix
RBF radial basis function
REU Rosetta Energy Unit
RL reinforcement learning
RLHF reinforcement learning from human feedback
SOTA state of the art
TCR T-cell receptor
VAE Variational Autoencoder

A Appendix

A.1 Background on ProteinMPNN

Protein folding prediction is concerned with finding the 3D structure that an amino acid (AA) chain will fold into. After significant advances in this area - in particular by models like *AlphaFold* (Jumper et al. [2021], Abramson et al. [2024]) - the inverse problem, protein design, has also attracted renewed attention. Traditional approaches have been physics based like the *Rosetta Packer* (Leaver-Fay et al. [2011], Fleishman et al. [2011]). The current machine learning (ML) revolution is leading to a plethora of data-driven approaches using for example large language models (LLMs) (Ferruz et al. [2022], Madani et al. [2023]), Variational Autoencoders (VAEs) (Lyu et al. [2024]) and Generative Adversarial Networks (GANs) (Repecka et al. [2021]) to generate new amino-acid sequences. A particularly prominent model in this space is ProteinMPNN (Dauparas et al. [2022]). Although is a relatively small model (around 1.7m parameters) it supports protein complexes consisting of multiple chains, and parts of the produced sequence can be fixed upfront. This might be particularly interesting when the user has detailed knowledge of the desired functional mechanism and wants to ensure the necessary AAs are present.

A.1.1 ProteinMPNN data embeddings

The examples delivered by the above dataset are processed within the model to produce the following initial embedding tensors that are then fed through the encoder and decoder which are updating these.

Edge embeddings: The edge embeddings tensor will have shape $[B, S, N, D]$ (see main paper Table 1). The dimensions two and three determine which residue and neighbor the encoding in dimension four belongs to. Only during ProteinMPNN training, Normal random noise with standard variation of 0.2 Å is added to the coordinates of each residue. The nearest N neighbors (including itself) are then determined based on C_α distance between residues. Then the coordinates of the AA atoms are used to calculate inter atom distances. For each residue-neighbor pair, 25 inter atom distances are calculated. Each of these is between an atom (N, C_α , C, O and virtual C_β) of the residue and an atom of the nearest neighbor (N, C_α , C, O and virtual C_β). Distances are not encoded as a single numbers, but rather using 16 radial basis functions (RBFs) with centers spread between 2 and 22 Å (far distances - above 25 Å- will have values close to zero in all bins). So, at this stage each residue-neighbor edge will be encoded using 400 elements (25 x 16). These get concatenated with a relative positional embedding. This embedding can be one of 66 different learnt 16 element vectors. One of these vectors signifies that the neighbor is on a different AA chain, while the others signify that the neighbor is up to 32 positions before or after the residue concerned (further distances are clipped).

Overall the concatenated edge embedding tensor now has a shape of $[B, S, N, 416]$. This is reduced to $[B, S, N, D]$ by an affine transformation.

Node embeddings: These are a tensor of shape $[B, S, D]$. They start out being zero and are first updated by the encoder and afterwards by the decoder.

Token embeddings: These are only accessible by the *Sequence Decoder*. A sequence is represented as a list of AA indices (e.g. 0 is alanine, 1 is cysteine, ...). These are embedded into a D -dimensional space resulting in a token embeddings tensor of shape $[B, S, D]$.

A.1.2 ProteinMPNN information flow

Here we examine the information flow in a training forward pass. A sample step is the same for the encoder but the decoder needs to be called several times. This is a bit more complicated as it needs to be ensured that the decoder does not access information that is not yet available. For clarity, the following description leaves out normalization, dropout and activation steps.

Backbone Encoder

We now will discuss the *Backbone Encoder*. This is a message passing neural network (MPNN) consisting of a stack of three encoder layers. It receives the edge embedding, node embedding, nearest neighbor information as well as a mask that stores which residues have no known coordinates. The edge and node embeddings are updated by each layer of the stack.

Within a encoder layer, at first a message tensor is produced. This happens by combining the edge and node embeddings into a tensor of shape $[B, S, N, 3 \times D]$. For an edge between a residue specified in dimension 2 and one of its neighbors specified in dimension 3, its last dimension holds the following three embeddings (each D elements): the residue’s node embedding, the edge embedding, the neighbor’s node embedding. A sequence of three affine operations transforms this into the message tensor of shape $[B, S, N, D]$.

Elements in the message tensor that are linked to missing residues are first masked out, before summing over its third (neighbor) dimension. The result of this is then added to the node embeddings (skip connection). These updated node

embedding are then fed through a two layer fully connected network and the result added to the node embeddings again (skip connection).

To update the edge features, again a message tensor is produced as described above (using different parameters for the operations). This time, however, missing elements are not masked out. This message tensor is then added to the edge features (skip connection).

Sequence Decoder

A decoding order is sampled (missing residues are first). Based on this a backward and a forward mask are generated. An element in the backward mask is one if and only if a node’s neighbor has already been decoded. The opposite holds for the forward mask.

The decoder receives the finally updated edge and node embeddings from the encoder as well as the masked sequence embeddings. With those the tensor EXV of shape $[B, S, N, 3 \times D]$ gets produced. Its last dimension hold first the encoder produced edge embeddings, then zeros and finally the neighbor’s node embeddings (each D elements). In addition, the forward mask is applied to it (elements of neighbors that have already been decoded are set to zero).

Before the call of each of the three decoder layers, also another tensor ESV with a shape of $[B, S, N, 3 \times D]$ gets produced. Its last dimension holds initially the same elements at the beginning and end as EXV . Just the zeros in the middle are replaced by the **AA** embeddings of the neighbor. Also, the neighbor’s node embeddings at the end get updated with every decoder layer call. Finally, the backward mask is applied to ESV (elements of neighbors that have not yet been decoded are set to zero).

Each decoder layer is then called with the node embeddings (by reference), and the sum of EXV and ESV . So the following embeddings are available for each edge.

- The residue’s node embedding
- The edge embedding produced by the encoder
- The neighbor’s **AA** embedding or zero if the neighbor has not yet been decoded
- The neighbor’s node embeddings (encoder produced for not yet decoded neighbors)

The shape $[B, S, N, 4 \times D]$ tensor holding this per edge information is fed through three affine layers, resulting in a message tensor of shape $[B, S, N, D]$. The neighbor dimension of this message tensor is summed over and the result added to the node embeddings (skip connection). This updated node tensor is then fed through a two layer fully connected network and added to itself (skip connection). After this all node embeddings linked with unavailable residues are set to zero.

After the last decoder layer, there is an affine transformation that produces logits from the final node embeddings.

A.2 PWM presentation predictor example

Let us assume the peptide *SEELDKGEV* and our hypothetical patient (see main paper **Definition 1**).

Once, we need to construct the **position weight matrix (PWM)** for 9-mers corresponding to the hypothetical patient. For this we

1. Generate a million random 9-mers (5% change of each **AA** to be in each position)
2. Use *netMHCpan* (with a rank cutoff of 2%) to identify those that are predicted to be presented. We find that these are 19.6% of all random 9-mers
3. We construct the **PWM** by counting the number of **AAs** at each position within random presented 9-mers. This, divided by the number of random presented 9-mers is the probability of observing this **AAs** at that position. The log of this is recorded in the **PWM**
4. We calculate the the sum of log-probabilities for each 9-mer
5. The cut-off is then the 19.6% (see above) percentile of these log-probabilities. In our case this is -26.48

For this we first use *netMHCpan*

For the peptide *SEELDKGEV*, the probabilities for finding its **AAs** at their respective positions in presented peptides within PWM are: 6.3%, 1.8%, 3.4%, 4.2%, 4.2%, 5.0%, 2.8%, 6.3%, 10.3%. The logarithms of these are -2.76, -3.99, -3.38, -3.16, -3.17, -3.00, -3.57, -2.77 and -2.27. Their sum is -28.07 which is lower than the limit of -26.48 we calculated as cut-off (see above). We, therefore, predict that the peptide is not presented. If in contrast the second position in the peptide had been a proline, which was observed in 9.5% of all presented peptides, then the sum of the log-probabilities would have been -26.43, which would be above the cutoff point and therefore predicted to be presented.

	position								
	1	2	3	4	5	6	7	8	9
A	-2.92	-2.75	-2.89	-3.00	-3.04	-3.04	-3.09	-2.82	-3.20
C	-3.72	-4.63	-3.52	-3.72	-3.59	-3.54	-3.47	-3.32	-3.69
D	-3.48	-6.30	-2.83	-2.65	-3.17	-3.31	-3.42	-3.59	-9.55
E	-3.34	-3.99	-3.38	-2.68	-3.10	-3.20	-3.09	-2.77	-6.18
F	-2.65	-2.87	-2.76	-3.20	-3.01	-2.92	-2.95	-3.07	-1.87
G	-3.26	-3.95	-3.41	-2.85	-2.91	-2.99	-3.57	-3.30	-6.20
H	-2.75	-2.19	-2.90	-3.03	-2.77	-2.83	-2.67	-2.89	-3.89
I	-2.89	-3.46	-2.91	-3.23	-2.93	-2.81	-2.88	-3.11	-2.31
K	-2.83	-3.29	-3.05	-2.93	-2.84	-3.00	-3.18	-2.89	-4.93
L	-3.12	-2.85	-3.03	-3.16	-3.06	-2.99	-2.87	-2.96	-1.88
M	-2.86	-2.92	-2.77	-3.24	-3.19	-3.09	-2.90	-3.07	-2.06
N	-3.02	-4.36	-2.91	-2.95	-2.96	-3.03	-3.01	-3.09	-8.45
P	-4.09	-2.35	-2.91	-2.52	-2.74	-2.71	-2.88	-3.17	-4.89
Q	-3.06	-2.90	-3.04	-2.90	-3.02	-2.94	-2.85	-2.78	-5.89
R	-2.77	-2.33	-3.14	-3.12	-2.88	-2.97	-2.98	-3.02	-5.14
S	-2.76	-3.02	-2.96	-2.87	-2.99	-2.93	-3.02	-2.72	-5.16
T	-2.92	-3.16	-3.13	-3.00	-3.00	-2.85	-2.89	-2.66	-4.12
V	-2.89	-3.30	-3.13	-3.21	-2.96	-2.85	-2.86	-2.86	-2.27
W	-3.23	-2.82	-2.97	-3.11	-3.02	-3.18	-2.90	-3.53	-2.24
Y	-2.54	-2.41	-2.70	-3.19	-3.03	-3.05	-2.89	-2.88	-2.11

Table A1: PWM for presented 9-mers in our hypothetical patient: Each column corresponds to a position in the 9-mer and each row to one of the 20 standard AAs. The values are the log-probabilities of finding the corresponding AA at that position in a presented 9-mer

A.3 Protein assessment details

Figures A1, A2 show the information for the specific validation proteins corresponding to the **Figures 3, 4** in the main paper. Then follow pages with detailed information about the generated designs for specific validation, specific test and illustrative proteins (mixed and ordered alphabetically).

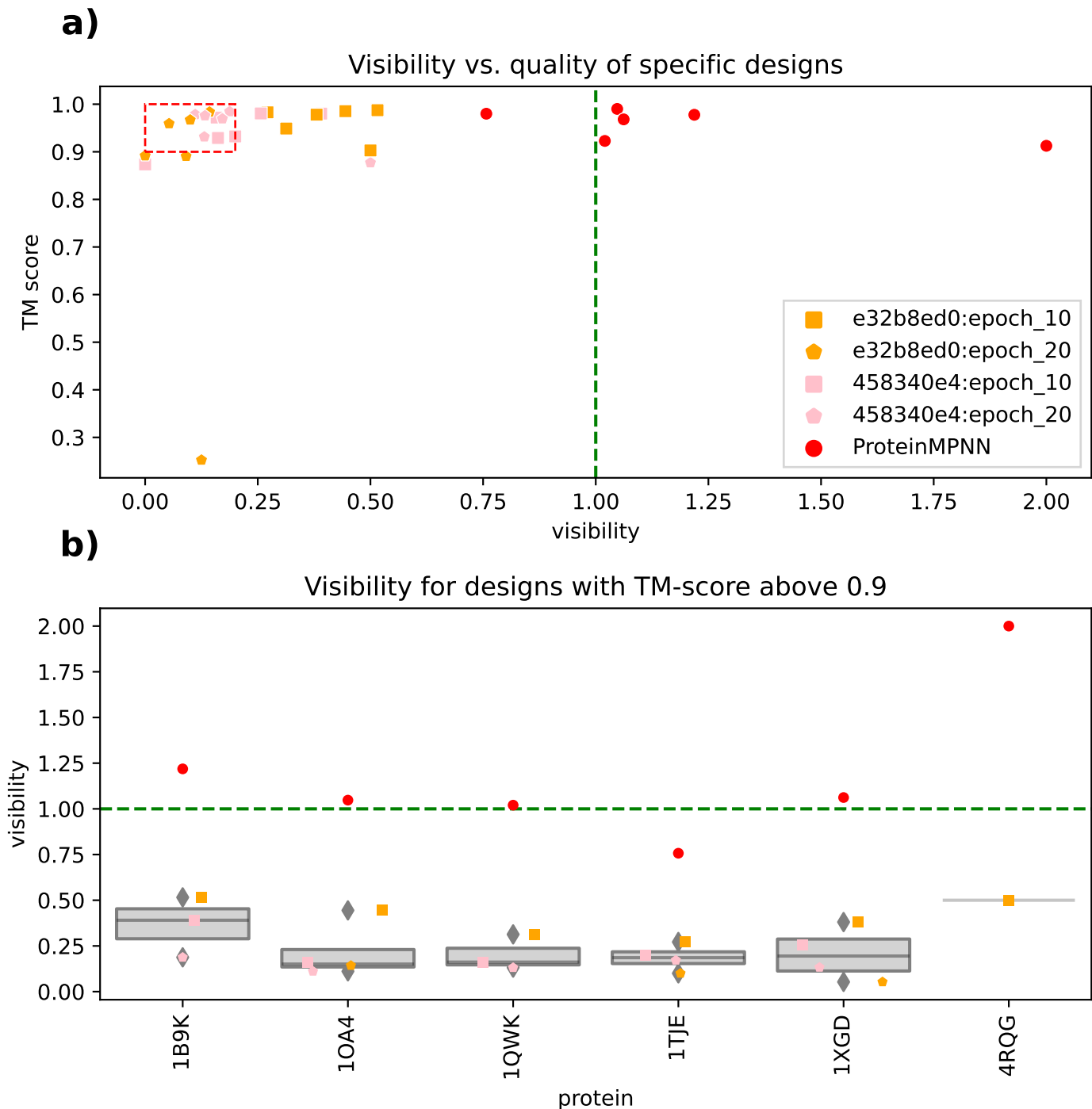


Figure A1: A broad performance range: This figure looks into specific valuation designs. In plot **a)** the quality of generated specific designs (as measured by TM-score on the y-axis) gets compared to the relative visibility (x-axis) for two [Direct Preference Optimization \(DPO\)](#) hyper-parameter selections after two epochs of tuning. Each point represents a design for a specific backbone using the checkpoint. As for the specific test designs (main paper **Figure 3**), We find that most designs have high TM-scores above 0.9. Not all of them are less visible than the original sequence (visibility < 1.0) and we also see that the **ProteinMPNN** designs in red tend to have far higher visibility. Furthermore, some designs also only show a maximum of 20% the number of visible peptides in the designs (red dashed box). In plot **b)** we then look into the distribution of visibilities of designs for specific backbones. These are subset of designs from plot **a)** that satisfy the condition that their TM-score is above 0.9. The dashed green lines represent the visibility of the PDB proteins.

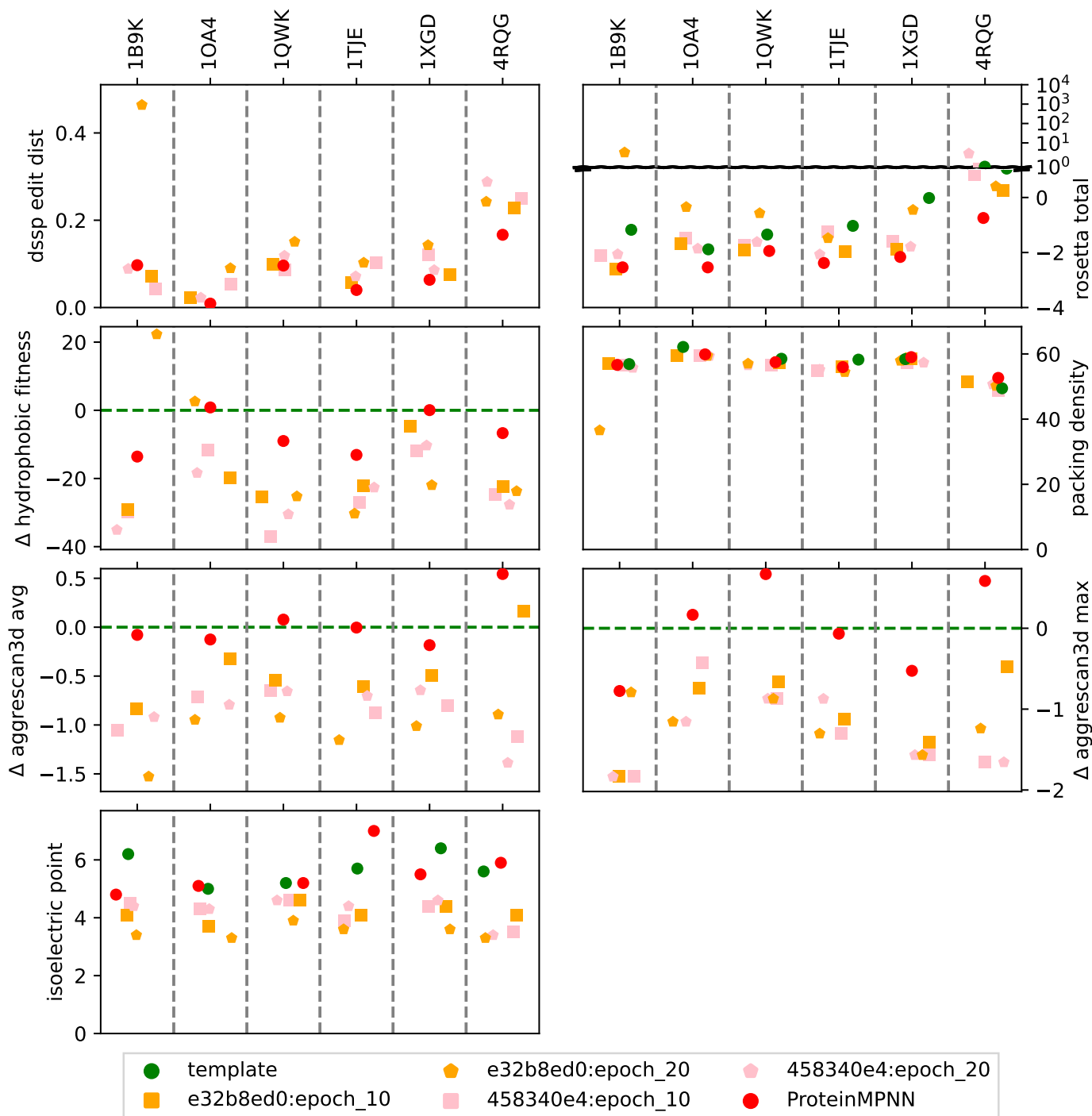


Figure A2: Quality indicators for designs: The figures above show the distribution of quality indicators for the ProteinMPNN and CAPE-MPNN designs by specific validation set protein template. Each dot signifies a design. The green bars in the background depict the range of values for identified naturally occurring peer proteins (only included if at least five were identified). For all indicators except packing density and isoelectric point (pI), lower values are considered favourable. With regards to pI, we want to be as close as possible to the template protein and with regards to packing density, higher values are considered favourable (Stam and Wood [2021]). The figure showing "rosetta total" is split in two parts. The bottom part (range -4 to 1) depicts the normal range of values. To also depict the outliers that have problematic values we also show a second log scale above.

A.4 1A3H

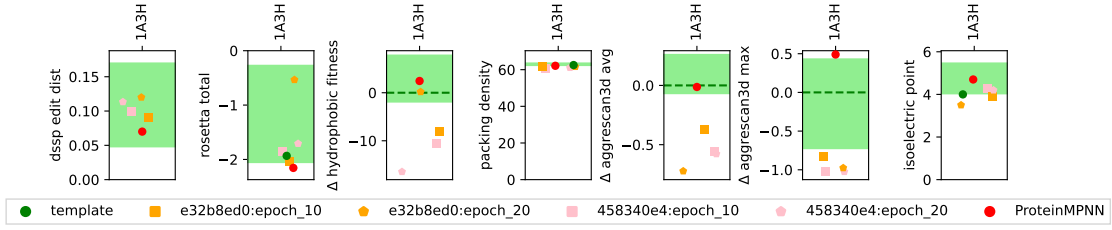


Figure A3: Focus on 1A3H

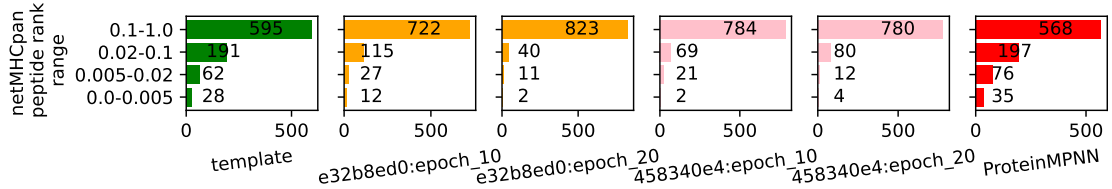
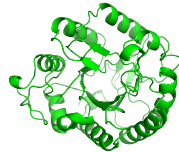
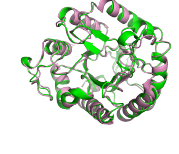


Figure A4: netMHCpan ranks for 1A3H designs

template	ISNGELVNER	GEQVQLKGMS	SHGLQWYQQF	VNYESMKWLR	DDWGINVFRA	AMYTSSGGYI	DDPSVKEKVK	
SVVEEHGQLS	ISNGELVNER	GEQVQLKGMS	SHGLQWYQQF	VNYESMKWLR	DDWGINVFRA	AMYTSSGGYI	DDPSVKEKVK	
EAVEAIDLDL	IYVIDWHIL	SDNDPNYKE	EAKDFDEMS	ELYGDYPNVI	YEIANEPNGS	DVTWGNQIKP	YAEVIPPIIR	
NNDPNIIIV	GTGTWSQDVH	HAADNQLADP	NVMYAFHFYA	GTHGQNLRDQ	VDYALDQGA	IFVSEWGTSA	ATGDGGVFLD	
EAQVVIDFMD	ERNLSWANWS	LTHKDESSAA	LMPGANPTGG	WTEAELSPSG	TFVREKIRE			
e32b8ed0:epoch_10	- TM score: 0.993, Seq Recovery: 0.51, Visibility: 0.43							
DFVTKNGKLS	IKDGKLVNEN	GEEVVLKGVS	SDNLQDNGDY	VNDSNIKDLK	DNWGINIYRA	NVDRDNGGLV	DDPSILDKAY	
DAIDACEKND	IYGVFDLHNG	NTGNPNNDID	LDKEVMKEIA	EKYGDKDNVI	YDINNPNNGA	DVTWDTDIKP	YAEVIPPIIK	
ENDKDNIIIV	GTANNSTDVV	SAADNKLDDP	NVMYAFHAAA	GEELDKKMED	VKKAQDKGAA	IFMSEWTTAN	WNGDGGVHLD	
EAQKLWDFMD	ENNISGTYWA	DSDNDRDSAL	RTSGASPDGG	WSDSQLSESG	QAVAKIRDS			
e32b8ed0:epoch_20	- TM score: 0.979, Seq Recovery: 0.38, Visibility: 0.14							
DDVDDNGKLS	VKDGKEVDEN	GNETTLKGNS	SDDNNKGDY	DNKDNIDDDK	NEDGANIWA	NNVDGGGRD	DDPSTKDKDD	
DSNDACEPDN	IYGDSNNNG	NDGNPNNDKD	KDKEDMDEKA	KKYGDKDNVI	YDIDNEPNGD	DVTWDDDIKP	NADEVIPPIK	
ENDDDNIDV	GTANNSTDVD	TAADNKLDD	NTRANNHSSA	GNDGDDDKN	VDKAQDKGEA	VNNGEWWGAN	DNGDGGVNDD	
EDDKENDYND	DNGISSNYWS	DSNDDDDASD	RTKDASPDGG	DTDDDDSENG	KKRKDDIKDA			
458340e4:epoch_10	- TM score: 0.989, Seq Recovery: 0.46, Visibility: 0.26							
DVVKNKGKLS	IKDGKIVNEK	GEEIVLKGIS	SDNLQKDGEY	VNCKNIKDKK	DNWGINIWRA	NCNIGNGGIK	DDPSILDIGN	
KGIDCCEKED	IYCVFDNNNG	TDGNCNTDID	LTKIKEINKEIA	EKYGDKDNVI	YDIDNEPNKG	DVTWKDNIKP	YAEVIPPIIK	
ENDKDNIIIV	GTANNSTDVV	EAAKNKLDDK	NVCYACHWCA	GEKGQKKKD	VKKAQDEGAC	IWCSEWSTAK	GNGDGGVNNEE	
EAEDEWDFCD	ENNISCCYCC	DSNEDKDAAL	RTKDNAPDGG	WKDEELSESG	KKVKDKIKKS			
458340e4:epoch_20	- TM score: 0.989, Seq Recovery: 0.45, Visibility: 0.18							
DFVKNKGKLS	IKDGKIVNKD	GEEVVLCGIC	SDGLQKNGEY	VNCKSIKWCK	DECGCNIWRC	NCKWGNNGWI	DDPSIKDIDC	
KGIDCCEEND	IYCVFDCNG	DIGNPNDNID	KIKECMKEIA	EKYGDKDNVI	WDICNEPNKG	DVTWEDNIKP	YAEVIPPIIK	
ENDKDNIIIV	GTGNNSTDVV	EAAKNKLDDK	NVCYACHFCA	GEKGDEKKD	VKKAQDEGAC	IFCSEWSCAK	GNGDGGCNFD	
EFDKWCDFCD	ENNISWCYCC	WSNDDKDNAC	LTKNASPDGG	WKDEELSDCG	KKVCDIIKKK			
v_48_020	- TM score: 0.993, Seq Recovery: 0.53, Visibility: 1.23							
MFVEKNGKLS	IKNGKLVNEN	GEPVVLKGS	SDALQKYGEY	LNYSMMKWLK	DEVGINIFRA	NMNMDSGGYI	EDPSIYDLVD	
KAIDAAAEELD	IYAVVDAAHN	SLYNPNPTYLD	ERIELFEKIA	KKYGDKPHVI	YEIANEPSGA	DVTWEKDIKP	YAEKVIPVIR	
KHNPDNIIIV	GTPNWSRDVV	TAANNPLAYP	NVMYAFHFAA	GEPLLEHMK	VDKALAKGAA	IYCSEFSLAK	WNGDGGVFP	
QFNKMWDYFD	EHNISYTYWS	LSTANRDAAL	LTSASPDGG	WTDSQLSAAG	QLVKAKIRAA			

Box 1: Template and designed sequences for 1A3H

A.5 1B9K

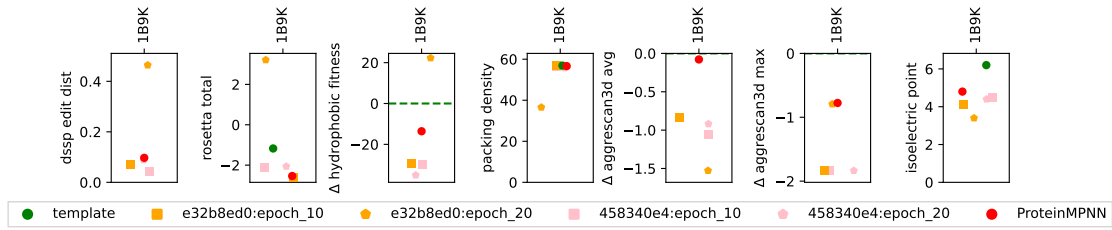


Figure A5: Focus on 1B9K

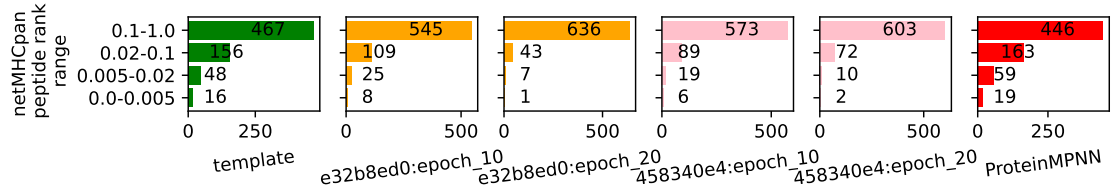


Figure A6: netMHCpan ranks for 1B9K designs

template	EDNFARFVCK	NNGVLFFENQL	LQIGLKSEFR	QNLGRMFIFY	GNKTSTQFLN	FTPTLICADD	LQTNLNLQTK	PVDPTVDGGA
	QVQQVINIEC	ISDFTTEAPVL	NIQFRYGGTF	QNVSVKLPIIT	LNKFFQPTIM	ASQDFFQRWK	QLSNPQQEVQ	NIFKAKHPMD
	TEITKAKIIG	FGSALLEVD	PNPANFVGAG	IIHTKTTQIG	CLLRLEPNLQ	AQMYRLTLRT	SKDTSVSRQLC	ELLSEQF
e32b8ed0:epoch_10 - TM score: 0.987, Seq Recovery: 0.33, Visibility: 0.52	GDNFEDFINK	DSGTLTKDDI	IEINIDKTYD	GNTGTIKLTF	KNLTDEDITD	FDPKIEQSED	LKEKLDIKYD	DGDKTIAGGE
	EKDLTIDVTA	KAPFKDKPTL	KITYTYNGEE	KEDEVKLPIV	LNDFAEPEVM	NKEEFEEKN	KVTDPDNIKT	TTKKRKNPND
	KEKIKEKIKG	FGFALLEVD	DDKNNFVGAG	TIKTEKDKIK	VLLLEKPDDE	NDSDTLTVKA	DDAEVADIVN	EELSKLF
e32b8ed0:epoch_20 - TM score: 0.252, Seq Recovery: 0.22, Visibility: 0.12	GDNTDDFIDK	DDGTLDEDD	TKVDQDGDYD	GNKGTVKLTT	TNKTDDDDADD	DDDTVQDD	LKDKLIDIED	DGDDTIDGGD
	SEDITLTDN	KONYDDPTL	DRDYDLGD	KSDSSDLPIN	NNKNKEAEE	DKDEWEKKKN	EAEDDDIEE	ETSDNKNND
	DDEIKEKIKG	FGDAVLEVD	DDNNYDGKG	TIKSEKDD	VQSKLENDEE	NDKDLTLVIA	DDAETAKTEN	EDKKKLY
458340e4:epoch_10 - TM score: 0.980, Seq Recovery: 0.33, Visibility: 0.39	GNNFENFINK	KNGTLYKDDI	IEINIDKECD	GNKGTIKIKI	KNKTDEDIKN	CDCKIECED	LKEKLDIKYD	EGDKTIKGGG
	EKDITIKVTC	KKPFKEKPIL	KIKYEYNGKE	IEKEIKLPIV	LNDFMEPKEM	NKEEWEED	KIEDEDEIKT	ITKKRKNKND
	KEKIKEKIKG	FGGALLEVD	DDKNNFVGNG	IIKCEKDKID	VKFELKPDDE	NDTDLTLIKA	DDEEVADILC	EELSKLF
458340e4:epoch_20 - TM score: 0.985, Seq Recovery: 0.30, Visibility: 0.19	GDNYEDFICK	KDGTLEFENDI	FKVDIEKTCE	GNKGTIKITF	TNKTDKDCTN	FDPKIECDDE	LKEKLDITSD	DCDKTIEGGK
	SIDKTINVTC	KDIFSGKPII	KIKCEYNGED	KEYEIKLPIN	INDFCEPKEC	DKETFWEECN	KCKDKDCICE	KTKRKNND
	KEKIKEKIKG	FGFALLEVD	DDKNNFCCCG	IIKCEKDKIK	CQCIKPDDE	NNTCTITCIC	DDSECKICC	EEISKLF
v_48_020 - TM score: 0.978, Seq Recovery: 0.39, Visibility: 1.22	GDNFEEFINK	KSGVLYENDI	IKVTISATYS	GNKGTITLTF	TNLTDVPLTN	FTPTIVLSPE	LQKLTITYA	PVDPTLAGGS
	SVTLTLNVTA	KDPFSERPIL	VITYTYKGKE	HKESITLPIV	LNDFLEPKPM	DKETFWALYN	SVTDPENVVE	ITFKAKNPFD
	RETIKAKLKG	FGFALLEVD	PDPNNFVGAG	IVKFENKDIK	VLLILKPDLE	KNTFTLTIIA	DDPEVAKILA	ELLSKLF

Box 2: Template and designed sequences for 1B9K

A.6 10A4

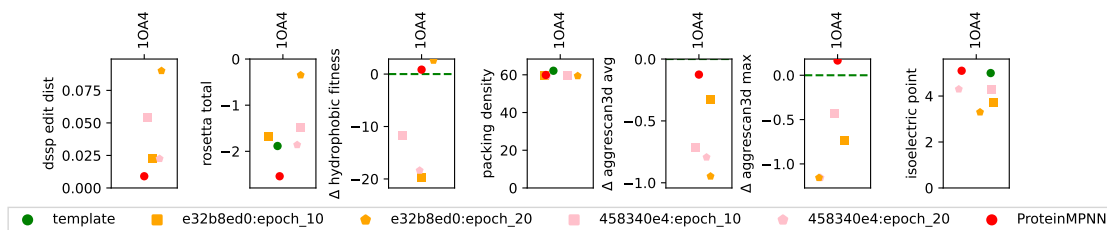


Figure A7: Focus on 10A4

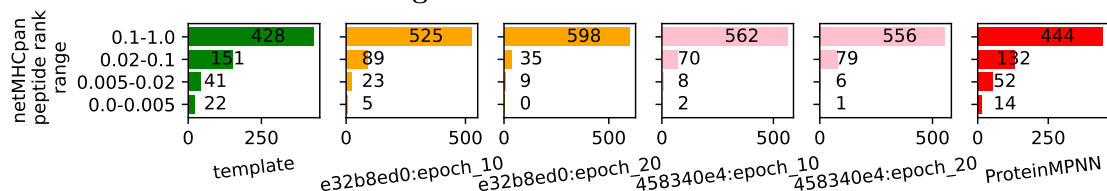


Figure A8: netMHCpan ranks for 10A4 designs

template	TTIQDRYVYVQ	NNRWGTSATQ	CINVTGNGFE	ITQADGGSVPT	NGAPKSYPSV	YDGCCHYGNCA	PRTTLPMRIS
NQQICDRYGT	TTIQDRYVYVQ	NNRWGTSATQ	CINVTGNGFE	ITQADGGSVPT	NGAPKSYPSV	YDGCCHYGNCA	PRTTLPMRIS
SIGSAPSSVS	YRYTGNGVYN	AAYDIWLDP	PRTNGVNRTE	IMIWFNRVGP	VQPIGSPVGT	AHVGGRSWEV	WTGSNGSNDV
ISFLAPSATS	SWSFDVKDFV	DQAVSHGLAT	PDWYLTSIQA	GFEPWEGGTG	LAVNSFSSAV	NA	
e32b8ed0:epoch_10 - TM score: 0.985, Seq Recovery: 0.50, Visibility: 0.44							
TTTTTEKDG	TTIGNKFVAS	NDNWGITD	AIEVGANGFT	ITKSDGSNST	DGNSKGNPHI	FDGCENGVCA	PGTTLPPKIK
DIGSAPASVS	ITTTDTGVYD	AEGDIDIDKT	AKTSGTADTE	IDIIIFDRNGD	VKPVGEKIGK	ATVNGKNWDV	YEGSDGTRNI
LSLVADSP	SESFDVKDFV	DLAIDK GKAT	NDDYLTDIGW	GFDNWDGEGG	NAVNDFSASV	NA	
e32b8ed0:epoch_20 - TM score: 0.984, Seq Recovery: 0.42, Visibility: 0.14							
TETICEKDG	TTIKDKYVAD	NDNDGSDDE	CIDVDNNGFT	ITKNDGSNDT	NGDGKGDPHI	YDGCEDGNCA	PGTNLPKIK
DIGNAKKGGK	YTKTDDGYD	AEDDIDIDAT	PKTSGTDDTE	IDNVDSNDG	VDDVGEEIGD	DEINGEDYDV	WEGDDGNKDI
LGNDADDDID	SKDYDAKDDV	DKAIDK GKAT	DDDYLLNIGS	GFBNYDGGEG	NSVDDFSADV	DD	
458340e4:epoch_10 - TM score: 0.972, Seq Recovery: 0.43, Visibility: 0.16							
KKIICEKDG	TTIKDKFVVE	NNNWGTDDDEC	CIEVGENGFK	IVKCDGSNST	DGNSKGDPHI	FDGCENCKCA	KGTTLPPKIK
DIGKCPTSVK	IKITDDGVYD	VEDDIDIDKT	AKTDGTTDDT	IDIIRDKNKD	VKPIGKKIGK	ETINGKNWDV	WEGNDGKKNI
LSLVADEKID	SESKDVKDYV	KLAIDK GKAT	EDDYLCDI	GFDCWEGGEG	REIKDFSACK	ED	
458340e4:epoch_20 - TM score: 0.978, Seq Recovery: 0.44, Visibility: 0.11							
DKTICEKDG	TTIKDKFVVS	NDNWGTDDDEC	CITVGENGFK	ITKSDGSTST	DGDCCKGDPHI	YCGCDDGICA	PGTNLPKIK
DIGKCPCSVE	FKFTDDGVYT	AKYDIDIDKE	PKTDGITDTE	IEIVFDKNKD	VKPVGEKEIGE	AEINGKKWKV	YEGSDGKRDI
LSLVAKEKIE	KDSFDIKDCI	KLCIDK GKAT	EDDYLCRIGC	GFNCNWGGEG	NEVKDFDCKC	EE	
v_48_020 - TM score: 0.990, Seq Recovery: 0.52, Visibility: 1.05							
MKVLSEKNGY	TYIKNKYVVA	NNNWGTDAVP	AIEVGENGFT	IIKSEGSNPV	TGAPKGFPHI	YYGCVGVCA	PGTTLPMKIK
DIGKAPVSAS	FTYTDTGVT	AQWDILIDKT	PKTSGTASTE	IEVIFARNGP	VKPVGKYVGE	AEVNGEKWEV	WIGSDGTRNI
LSLVAKEPIA	SKSLDAKDFL	DLAIEKGLAT	PDDYLLSISA	GFAPWEGGVG	LSVKSFSAEV	LK	

Box 3: Template and designed sequences for 10A4

A.7 1P3C

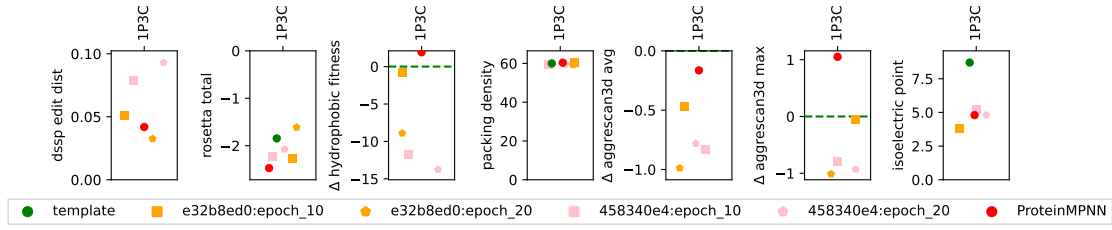


Figure A9: Focus on 1P3C

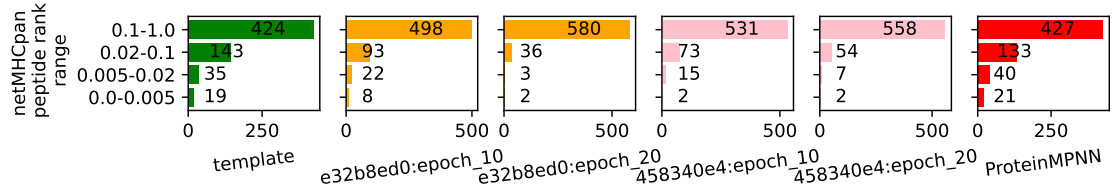


Figure A10: netMHCpan ranks for 1P3C designs

template VVIGDDGRK VANTRVAPYN SIAYITFGGS SCTGTLIAPN KILTNHCVY NTASRSYSAK GSVYPGMNDS TAVNGSANMT EFYVPSGYIN TGASQYDAV IKTDTNIGNT VGYRSIRQVT NLTGTTIKIS GYPGDKMRST GKVSQWEMSG SVTREDTNLA YTTIDTFSGN SGSAMLDQNQ QIVGVHNAGY SNGTINGPK ATAAFVEFIN YAKAQ	
e32b8ed0:epoch_10 - TM score: 0.992, Seq Recovery: 0.57, Visibility: 0.56 NIIGDDKRTE VDTTVAPYN AIAYVTFDNG YCTGTLIDNN KVLTNAHCYV NTNTNTYAGK GTVTPGQDNS TAVNGSANET DYYVPDGYIN TGDDKYDYAV IDTDTNIGDT VGKLDIKEVD TLKGENVTVA GYPGNITAT GKVSQWEESG KVTDEDEDLM YSDIDTDGGD SGAPFLDDDN KIVGIHSGGA DGGTKNYGPK ANKEVKDFID EAKKK	
e32b8ed0:epoch_20 - TM score: 0.989, Seq Recovery: 0.45, Visibility: 0.09 NIIGDDKRKE VDDTTDPNN SIADVTDGDG SGTGTLIDDN KILTNAHDDY NNNNKNADK GKSTPGEDNS EAKNGESNET DDDIPDGAKD TGDDSDDYSV RDTDKDIGK DVGKDTKEVD TLKGEDVTT GPNDEKDKT GKNSQYEESG KVTDEDDDNA YSDIDTDGGD SGADFLDKDN KIVGVNSGGA NNGTKNGGPK ANDDFKKDID DAKKK	
458340e4:epoch_10 - TM score: 0.989, Seq Recovery: 0.46, Visibility: 0.31 NIIGEDKRKE VEDTTKPNN CICYVTFDDG YCTGTLIDDN KILTNAHCYV NKETNYAGK GTATPGDNS EAKNGSSKIT DYDIPEGYKN TGDDKWDIAC IKLDKDIGK VGKLDIKEVD KLKGEKVKLC GYPGNIDKT GKVSQWCEEG EITDEDEDLM YSDIDTDGGD SGAPFIDKND KIVGIHTGGA KGGTKNYGIK ANKKVKDFID KCKKK	
458340e4:epoch_20 - TM score: 0.989, Seq Recovery: 0.45, Visibility: 0.17 NIIGEDKRKK VEDTTKPNN CICYVTFDDG NCTGTLIDDN KILTNCHCVW NCETNKWAGK GTVTPGQDNS EAKNGTTKET DYCVPEECKN TCDDKWDIAC IKTDKDIGK VGKLDIKEVD TVKGEKVKVC GYPNNIDKT GKVSQWCEEG EITDEDEDLM YTDINTDGGC SGAPFLDKDD KIVGINTGGA KNGTKNYGIK CNKKVKDFID ECKKK	
v_48_020 - TM score: 0.992, Seq Recovery: 0.52, Visibility: 1.13 AIIGEDKRTE VDTTVFPYN AIAYVTFSSG FCTGTLIADN KVLTNAHCYV NYNTKTYAGL GTVYPGMNDS TPLNGSAKMT AYYIPQEYID TGSWKYDYAV IDLTDIGKK VGYLEIKEVD TLKGETVLA GYPGNYIAAT GKVSLYEETG KIVSEDENLM YSKIDTDGGN SGAPFLDSNN KIVGIHTGSA KGGTINYGPK MTKEAKKFIE EALKK	

Box 4: Template and designed sequences for 1P3C

A.8 1PGS

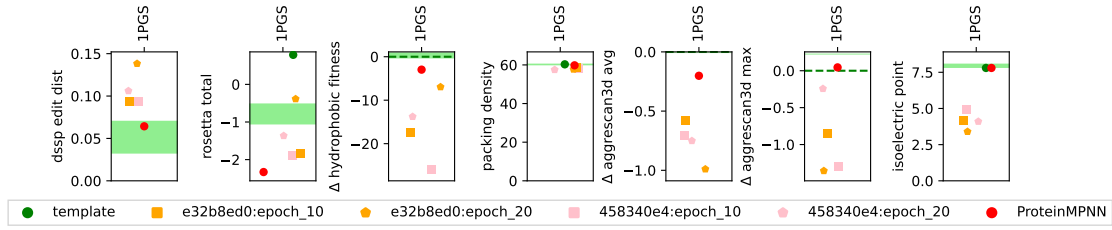


Figure A11: Focus on 1PGS

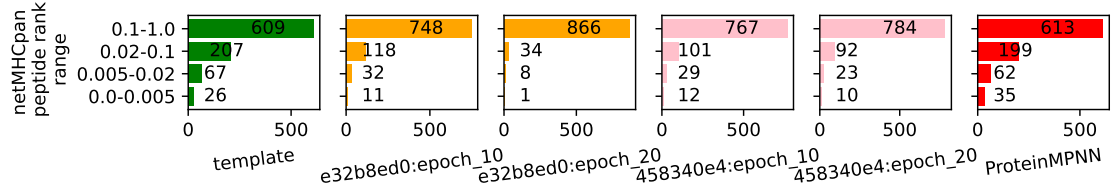


Figure A12: netMHCpan ranks for 1PGS designs

template DNTVNIKTFD KVKNAFGDGL SQSAEGTFTF PADVTTVKTI KMFIKNECPN KTCDEWDRYA NVYVKNKTTG EWYEIGRFIT PYWVGTEKLP RGLKIDVTDF KSLLSGNTL KIYTETWLAK GREYSVDFDI VYGTDPYKYS AVVPVIQYNK SSIDGVPYK AHTLGLKKNI QLPTNTEKAY LRTTISGWGH AKPYDAGSRG CAEWCFTHT IAINNANTFQ HQLGALGCSA NPINNQSPGN WTPDRAGWCP GMAVPTRIDV LNNSLTGSTF SYEYKFQSWT NNGTNGDAFY ATSSFVIAKS NTPISAPVVT N									
e32b8ed0:epoch_10 - TM score: 0.985, Seq Recovery: 0.49, Visibility: 0.46 EETETINAFE NAKTSSGNGG SKSAEGTFKF PEDTTKVNI DMHIKEDCPN NSCDPNCNYC KVEVKDKETG KWLTIKWIT PNNVGTEELK DGLKSDVTDA KDLIKGDTL KISCVDNSK GRFTSTFSDI TYGDPEYKYS EIEEVIDKDE SENDGIPYGT DNDIKTEKKV TLPANTEKAY LKVQTYGSGD ATPTAAGNKG AAEDRETDT IEVGDKPTFS INQGLGCAN NPINNQSPGN YTDDHAGWCC GQVVPITTFD LDDSLKGKTF TIKYKQEPYT NDGTNGDAYK SVTIDVVAES NTKIDKPKVT D									
e32b8ed0:epoch_20 - TM score: 0.969, Seq Recovery: 0.38, Visibility: 0.10 MTTNTNNGFD NAKCSDGNGG SKSDENTNDY PDDTTDDDI DQNTKENCND NNCDGNCNDK KKEVKDDDDG DWKTISEWRS NNNGTENDE NGLKDNVTDD KDLIKGDVDE KISCDDNSN GRYTSTSDI TEGDPPNKYS ETEDIIDND GDDGDIDVGD DDDDDTKKNV TTDDNTDDAD LNNKSGSGD ATPNDAGDK AAEDKDDTDT IKVDGEDTFD DDDGGIGCAN NKNNNGSPGN DNDDHAGWCC GETVPKTTNK LDDDKKGKTF TVESEQEDYT NDGTNGDADD DESTDIVDKS NTEIDKPEVT D									
458340e4:epoch_10 - TM score: 0.990, Seq Recovery: 0.49, Visibility: 0.44 EETKKIDCFK DVKCSGNGG SKSAEGTFKF PKDTTKVNI DMNIKEDCPN NSCDPNCNYC KVEVKDKETG KWLTIKWIT PNNVGTEELK DGLKDVDTDF KDLIKGDVKL KITCDVDNEK GRFTTISFDI TYGDPEYKYS EIEEVIDKDE GDNDGIPYK DNDIKTKKKV TLPKNTKAY LKIIKIGSGD ANPKDAGNKG AAEDRKTDT IKIDGKDTFK IDKGGLGCDK NKINKQDPGD WKDKAGWCC GEVVPITTCG LDKELIGKTF TIEYKFEDYT NDGTNGDAYK SITINVVCES NVKIEKPKVT D									
458340e4:epoch_20 - TM score: 0.980, Seq Recovery: 0.48, Visibility: 0.35 DEVKKIDCFK DVKCSGNGG SKCAEGTFKF PEDTTNVVKI DCYIKEDCPN NTCDPNCNYC KVEVKDKETG KWLTIKWIT PNNVGTEELK DGLKDVTCCK KDLICGETKL KICCDVDNEK GRFTSISFDI TYGDPEYKYS RIVEVIDYDE DDDGDIDYGV DDDKLKKKV TLPKNTKAY LKIIKIGSGD ACPKDEGNKG CCEDCERTDK IKINGKETFD IDMGGLGCEN NPINNCEPGD CKDDKAGWCC GQVPEKICE LDKELIGKTF TIEYEFEDRE NDGTGDCDV CITIDVVCES NVEIEKPKVT D									
v_48_020 - TM score: 0.994, Seq Recovery: 0.56, Visibility: 1.04 MKTITINAFQ NVKTASGNGG SKSAEGTFNF PEDTTNVVKI DMVYKESCPN NSCDPNSRYC YIQVKDKETG EWLLIAAWIS PYNVGTEELP NGLKTDVTDF KELLKGETEL KIYCEVTNSK GRITFSVSDI TYGTPEYKYS RIVPVIQYDE SSTKGIPYGV PHNILLKRKV TLPKNTKAY LIVRAYGYGN ASPYAPGNRG AAERWRPTDT ILINGKPTFS LPHGALGCAK NPINKQAPGN WKPDRAWCP GEPVPTHRFE LPSSLFGKTF TIEFKLEPYT NDGSNGTAYI ALSIYVVVES NVPIEKPVVT D									

Box 5: Template and designed sequences for 1PGS

A.9 1QKD

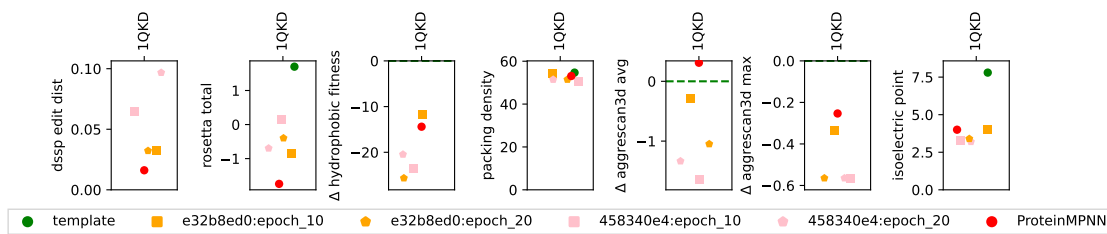


Figure A13: Focus on 1QKD

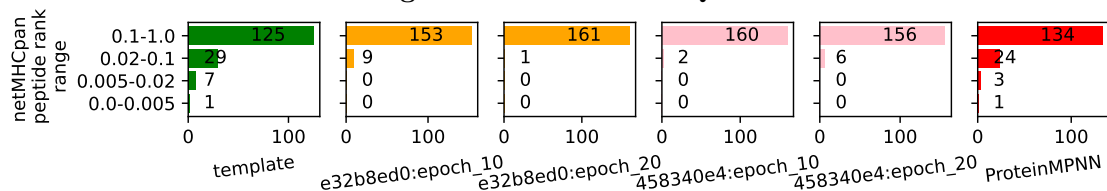
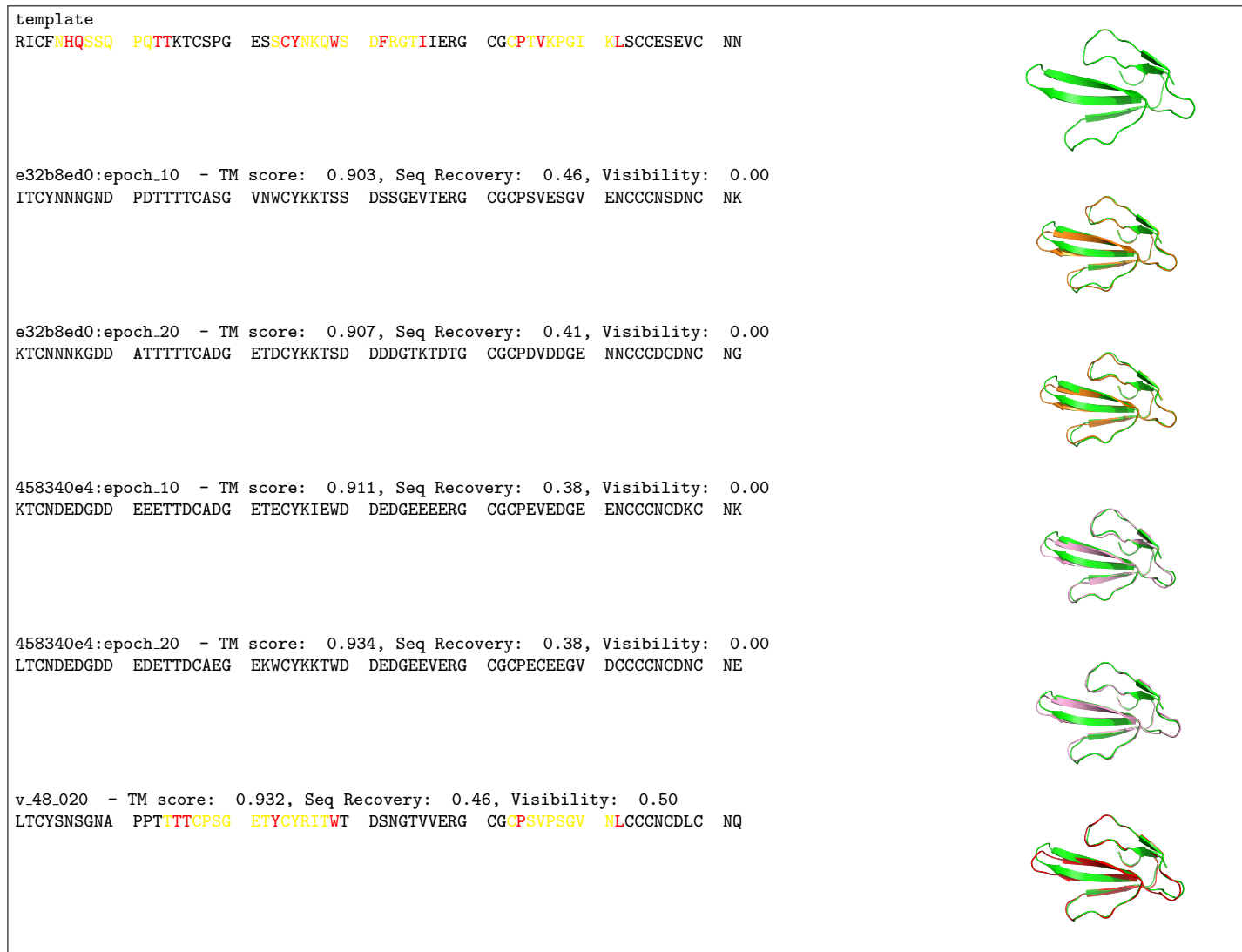


Figure A14: netMHCpan ranks for 1QKD designs



Box 6: Template and designed sequences for 1QKD

A.10 1QWK

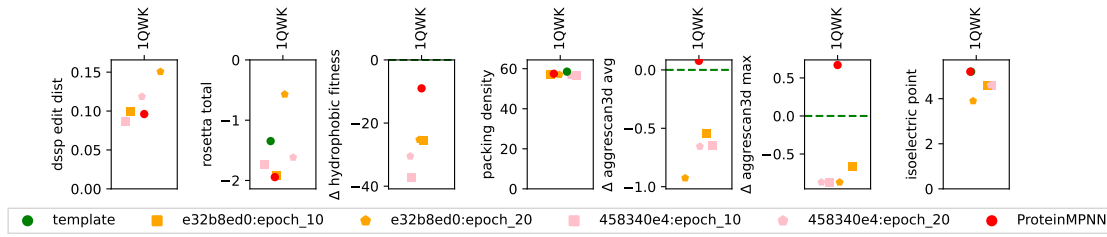


Figure A15: Focus on 1QWK

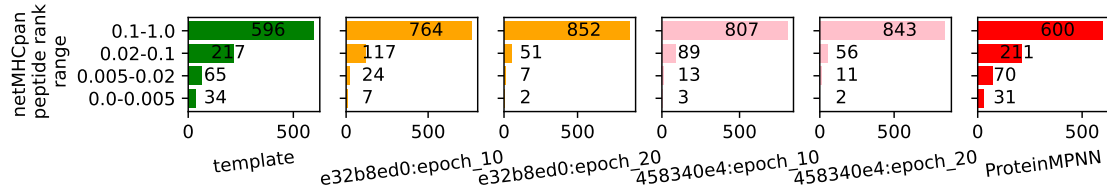
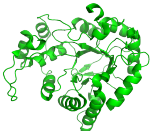
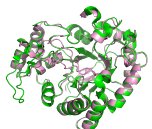


Figure A16: netMHCpan ranks for 1QWK designs

template	TASIKLSNGV	EMPVIGLGTW	QSSPAEVITA	VKTAVKAGYR	LIDTASVYQN	EAAITGAIKE	LLEEGVVKRE	ELFITTKAWT	
HELAPGKLEG	GLRESLKKLQ	LEYVDLYLAH	MPAAFNDMDMS	EHIASPVEDV	WRQFDAVYKA	GLAKAVGVSN	WNNDQISRAL		
ALGLTPVHNS	QVELHLYFPQ	HDHVDFCKKH	NISVTSYATL	GSPGRVNFTL	PTGQKLDPWP	APSDLQDQNV	LALAEKTHKT		
PAQVLLRYAL	DRGCAILPKS	IQENRIKENF	EVDFDSLTEE	DIAKLEESKN	SQRFLQLQDFM	TGHPEDAFAA	ER		
e32b8ed0:epoch_10 - TM score:	0.949,	Seq Recovery:	0.42,	Visibility:	0.31				
MKTITLSNGK	KLPILMLGTG	GATPEEAKEA	VKTAIKNGIK	GIDTAEGNGN	EKAVGEAIKE	AEEEGIVKRE	DLNIITKIGP		
AGLRPGKARE	AVEKSLEELG	RDKADVVLAGE	NTCATNDDGT	EFVDAVPTEI	HKELDEIKKE	GLADAIGVSN	ANNDQIDEAE		
ATGGTPIDVW	KDEINIDYPN	KDLIAHCKER	NIGVIADKPL	GKKGKEENV	EDGTKGESTD	KESQQTDENV	KKLAKEKGKT		
PEQICIKYCI	DKGCAVTPGA	KDKKDIEENG	KVFDFSLEE	EIKELEKSGK	NERQDKREET	KGHPNDPRKE	ER		
e32b8ed0:epoch_20 - TM score:	0.891,	Seq Recovery:	0.33,	Visibility:	0.09				
DDSIELSNGK	KKPKTKQGTT	GADEEEAKEA	VKDAIKNGIK	GIDTGEGNGN	NKAVGEAIDE	AEKNGDVNED	DLIEDKIGG		
DGNRRGKGED	AVNKSLEDLN	KDKVDVNKNG	NSDATNDDGT	EDVDEDVTBQ	HKEQDKIKKD	GKSDSIGTSN	ANNDQIDEAE		
ATGGDKVDDH	DDEINNGNKN	DDLIDHCKDK	DINVTAOKPN	GKDDTEDNTD	DDGTGDDTD	KESDGNDNV	KKKADEKGKT		
ADQINNQC	DKGCSVEPGE	KDKDDIKDN	DDNDSLTDD	EIDDLKSGK	NEKEDKRDN	KGHKNDANKD	ED		
458340e4:epoch_10 - TM score:	0.929,	Seq Recovery:	0.39,	Visibility:	0.16				
TKTIELSNGK	KLPILKLGTG	GCTEEECKEA	VKTAIKNGIK	GIDTAKDNNGN	EKAVGEAIKE	CEEEGICKRE	DLCIETKIGA		
DGLNPQKGVK	EVEEMLKELG	VDKVDICLAG	NCCCKNKDGT	EDIDVPVTEI	WKELDEIKKE	GLCDSIGVSN	CNIDQIDECE		
KTGGTKIDVV	KTEINIYNPQ	KELIEHCCEK	NINVIADKPL	GKKGKKENVK	EDGTKEEED	LESKTDENV	KELAKEKGKT		
EEQICIKWCI	DKGCSVTPGE	KKEEDIKEVG	KIDDFKLTTEE	DIKKLEESGK	NEIQDKREEN	KGHKNDPRKE	ER		
458340e4:epoch_20 - TM score:	0.931,	Seq Recovery:	0.38,	Visibility:	0.13				
TKCIELSNGK	KLPILIGLGL	DCTKEEAKEA	VKEAIAKAGCK	CICTASENGN	EAAVGEAIKE	CIEEGICKRE	DLCIITKIGS		
ECCRKGKAKE	CVDKDLEKLG	VDKVDVCLCG	NCCCFNEDET	ECIDEPITDI	WKQMCDIKKE	GKADAIGCCN	CNNKQIDECE		
ATGGEKCDIV	KVECNIDNPQ	KDLIDYCKKK	NIGVICDKPL	GREGKECNTK	EDGTGKCKD	KESEKTNEEV	KKLAKEKGKT		
PEQICIKWCI	DKGCCVEPEE	KDKKIKDKCL	KVDDFSLTDE	EIEKLEEDGK	NEKKDKRDWC	KGCKNDPKDK	EW		
v_48_020 - TM score:	0.923,	Seq Recovery:	0.49,	Visibility:	1.02				
MKTITLSNGL	KMPLLMLGLG	GATPEEAVEA	VKTAVKAGMR	GISTAASYGN	EAAVGEAIKE	LLEEGVVKRE	ELFVETKIPP		
HGLRPGRARE	TALAMLEELG	LDRVOLLLAG	NPCAFNEDLT	EFVPPVPVTEI	HRQLDEIKAE	GLADAIGVSN	ANIEQIEAAL		
ALGLTPIDVL	RTEINIEYPQ	TELIAFCCKH	NILVIADRPL	GRGRRRFVL	PSGVRLASVP	LPSPLTLPEV	LALAKELGKT		
PEQILIRWVL	ERGCAVTPGA	REPEAIKENL	KVFDFSLEA	DLARLLASGR	NELLDKREFT	KGHPNDPYKE	LR		

Box 7: Template and designed sequences for 1QWK

A.11 1UBQ

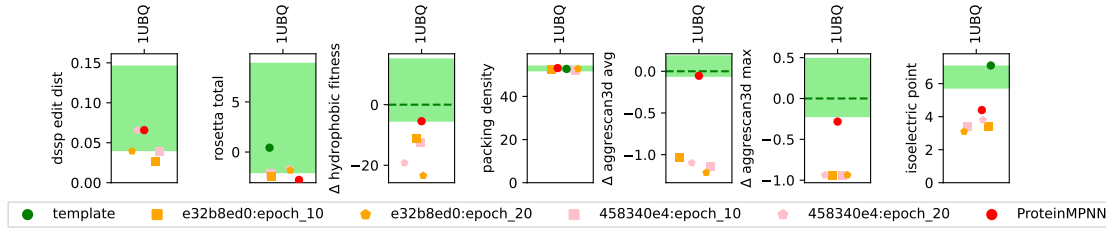


Figure A17: Focus on 1UBQ

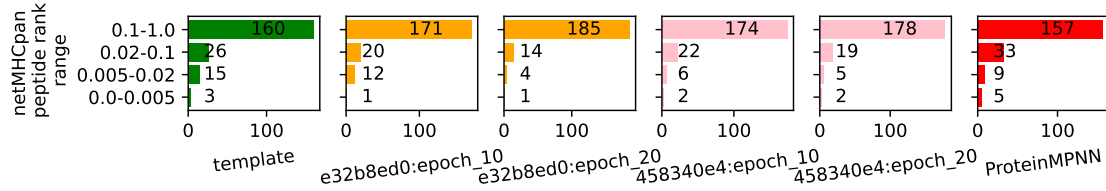


Figure A18: netMHCpan ranks for 1UBQ designs

template MQIFVKTLTG KTITLEVEPS DTIENVKAKI QDKEGIPPDQ QRLIFAGKQL EDGRTLSDYN IQKESTLHLV LRLRGG	
e32b8ed0:epoch_10 - TM score: 0.946, Seq Recovery: 0.49, Visibility: 0.56 MKIKIKDEDG NTIELEVDED DTIAEVKKKI EEKTGIDPEE QKLIYNGEEL EDDKTLKDYN IKEGDTIELE LEDDGE	
e32b8ed0:epoch_20 - TM score: 0.909, Seq Recovery: 0.42, Visibility: 0.17 MNINVEDDGG NTIELEVDDD DTVEDVKKKI KEKDGTEDEE DKLIVDDEEL EDDDTLDDYD IEEDDTLERE KEKDGG	
458340e4:epoch_10 - TM score: 0.956, Seq Recovery: 0.45, Visibility: 0.44 MKIKIKDEDG NTIELEVDED DTIEEVKEKI EEETGIDEEE QKLIYKGEEL EDDKTLKDYN IKEGDTIDLE IEDDGE	
458340e4:epoch_20 - TM score: 0.960, Seq Recovery: 0.47, Visibility: 0.39 MKIKIKGEDG KTITLEVDED DTIEEVKEKI EEKTGIDEEE QKLIYKGEEL EDDKTLKDYN IKEGDTIDLE IEDDGE	
v_48_020 - TM score: 0.960, Seq Recovery: 0.58, Visibility: 0.78 MTIKVKFTDG TTILEVSPS DTIAELKKKI QEKTGIPPEE QRLIYKGVVL EDDKTLADYN IQEGDTIELE LVPKGG	

Box 8: Template and designed sequences for 1UBQ

A.12 1SQ3

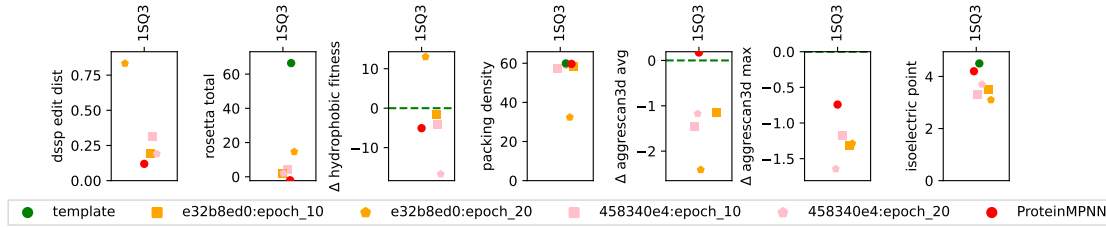


Figure A19: Focus on 1SQ3

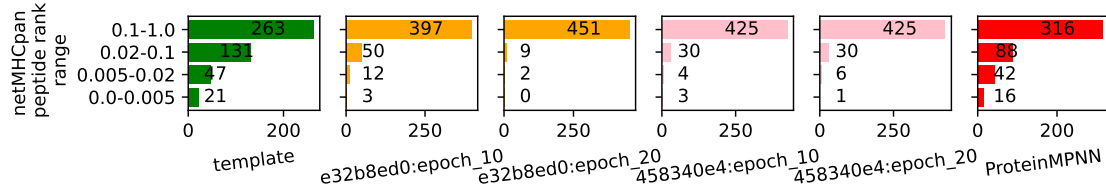


Figure A20: netMHCpan ranks for 1SQ3 designs

template SISEKMVEAL NRQINAEIYS AYLYLSMAS Y FDSIGLKGFS NWMRVQWQEE LMHAMKMFD V SERGGGRVKL YAVEEPPSEW DSPLAAFEHV YEHEVNVTKR IHELVEMAMQ EKDFATYNFL QWYVAEQVEE EASALDIVEK LRLIGEDKRA LLFLDKELSL RQ	
e32b8ed0:epoch_10 - TM score: 0.966, Seq Recovery: 0.43, Visibility: 0.22 MMSEETEEKL NEQLNNERYA GNTEEADASW YRSKGREGFA KEEVEEAEE LEHADDIADF IEERGGEVKL EQIEQPKQEW DSPLAAEEDH LEQEETTEE EEDIVEEAEE EGDETEEFM QEYVEQVEE EDEEEDVDK LKEIGDDKEA EKKLDEEEGE EE	
e32b8ed0:epoch_20 - TM score: 0.202, Seq Recovery: 0.28, Visibility: 0.03 DMTDDNDAE NDQVNNEEQA GKDEDADADW ENDNGNDGEA KEEVEEAEE QEHADDKEDD IEDRGGDVDE DQIDQKQDD DSLDAAEEDD NNDEKDTTNN EDDNVDKAKD DGDDGTADNE QEDTDDQDEE EDDEDDVDK LKNGDDKEA KDDNDKEEGK EE	
458340e4:epoch_10 - TM score: 0.971, Seq Recovery: 0.36, Visibility: 0.10 MMSEETEEKL NEQVNDERWT GNEEEADASW CKDKGKEGWA KEEVEEAEE NEHAEDWEDF IEERGKVKL EQIEQPKQEW DSELALAEEDD LEDEEETTEE EEDIVKEAEE EGDETEEFN QEDIEDQEE EDDEEDVDE LEECGDDKEC QDDEDKEEGE EE	
458340e4:epoch_20 - TM score: 0.963, Seq Recovery: 0.37, Visibility: 0.10 KMSEETCKKL NEQINDERWT GNEEEADASW CKSKGKDGWA KEEVEATEE NGHADDWKDF IEERGKVEL DKIDKPKQEW DSELAMEEDA EEEEEETTEE EEDIVKEAEE EGDETEFEFN QEDIEDQEE EDDEEDVDK LKECGDDKEC QEELDKEEGK EK	
v_48.020 - TM score: 0.988, Seq Recovery: 0.43, Visibility: 0.85 AMSAALVAAL NAQVNAELTA SYTELAASW YRSHGYDGF QWEKVAAEE IAHALALADY IHDRGGTLVL QQIPQPKQW ASPLAAEADA LAEEQATTAE VQALVAQAKA AGDTATETFL QAYIAAQVEE VAEQAIVDK LKAIGDDKAA LLALDKELGK VK	

Box 9: Template and designed sequences for 1SQ3

A.13 1S5T

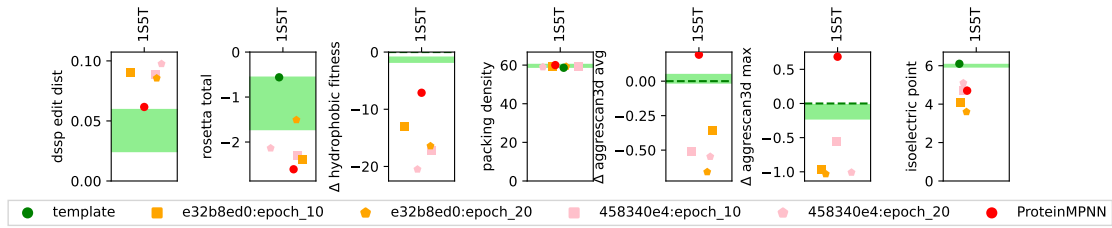


Figure A21: Focus on 1S5T

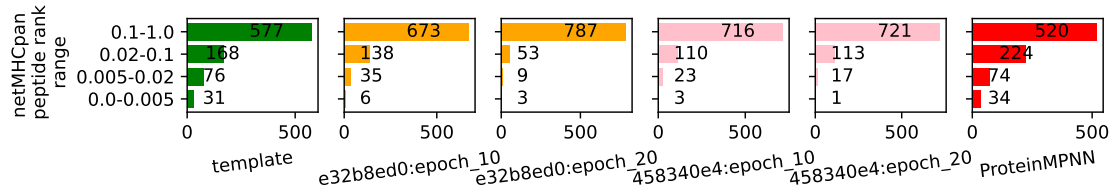


Figure A22: netMHCpan ranks for 1S5T designs

template								
MFTGSIVAIV	TPMDEKGNVC	RASLKKLIDY	HVASGTSIAV	SVGVTGESAT	LNHDEHADVV	MMTLDLADGR	IPVIAGTGAN	
ATAEAISLTQ	RFNDSGIVGC	LTVTPYYNRP	SQEGLYQHFK	AIAEHTDLPQ	ILYNVPSRTG	CDLLPETVGR	LAKVKNIIGI	
KEATGNLTRV	NQIKELVSDD	FVLLSGDDAS	ALDFMQLGGH	GVISVTANVA	ARDMAQMCKL	AAEGHFAEAR	VINQRLMPLH	
NKLFVEPNPI	PVKWACKELG	LVATDTRLRP	MTPITDSGRE	TVRAALKHAG	LL			
e32b8ed0:epoch_10 - TM score: 0.993, Seq Recovery: 0.52, Visibility: 0.38								
MFEGAIAAM	TPRDADGKMD	KESLKKMVD	MIEDGIKAIV	AVGTTGEADV	LDDDERIEVI	NTTIELADGK	IPVIAGTSAS	
DTDEAIALTK	KVEGSGVEGI	LSGTPGGVKP	TQEGLKDHYT	AVADNTDLPI	ILDNNPARTG	TKITPETVGE	LAKNENIIGI	
KNGSGDLSEV	DAIKELVDD	FKILSGTDDT	ALDEMKAGGD	GSISGTANVA	GKDCAEMADL	AAKGEYDEAQ	KIDDKLKDNL	
INKNAEPNPV	CTKWAAKELG	LIDNDTVRAP	YTPATEEAKE	KIRAALKKAG	KL			
e32b8ed0:epoch_20 - TM score: 0.987, Seq Recovery: 0.43, Visibility: 0.11								
MNDGAIAAAT	NPKDDDGND	DDSRKKQVDD	DIDNGVDGIV	ADGTTGEADT	NDDDTAKDNV	DKTIDDADGK	IPVIAAGTDN	
DTDNAIKKAK	TVEDSGVDGN	DTGNPGGNKD	TQDGKIDHYT	DVADNTDLPI	ITDNKPDRGT	TNMEPDVGE	LAKNDNIVGT	
KDETGDTDNV	DDIKDKVDD	YKILSGNDDT	ALDDQAKGGD	GNVSGTANVA	GKDEAEQADK	AADGDNDAK	EIDDKNAKLN	
KDKGSEKNPV	NTKEGEKTLG	KTDSDTVNEP	DTPATDEBKE	KIDDAALKDAG	KL			
458340e4:epoch_10 - TM score: 0.986, Seq Recovery: 0.46, Visibility: 0.24								
MFEGCICAI	TPKDKNNGYD	EKSMKMKIDK	CIEDGIDAIV	ACGGTGEKDV	LNDEEKKV	DTTIKEADGK	IPVIAGVSDN	
DTDKAIEKAK	KVEGKGVGL	LCGAPKGKIP	TQKGIDHYT	KVADNTDLPI	ILDNDPKVNG	TDIEPDVVG	LAKNENIIGI	
KDSTGKLDNV	KAIKEKVGDD	FKILSGKDEV	ALDAIKEGAD	GCISTVCNVA	GKDCSEMCKL	CQEGKFDEAK	KIDDKLKDLK	
INFGSQPNPV	CTKWACKKLG	LIDNDTVNEP	YKPATEEDKK	KIEEALKKAG	KL			
458340e4:epoch_20 - TM score: 0.990, Seq Recovery: 0.47, Visibility: 0.17								
MFEGCICACI	CPKDEKGYD	EKSCCKIIDK	MIEDGIDAIV	CCGTTGEADV	LNDDERIEVI	KTCKIEANGK	IPVIAGVGSS	
DTEEAIKLTK	KVEGTGVEGC	LCGVPKGKIP	TQEEIIDHYT	KVADNTDLPI	ILDNDPKGTG	TDIKPETIGK	LSKNKNIIGI	
KNGTGNLDEV	EKIKEKVDKD	FKILSGKDEV	ALDCMKKGGD	GCISGVCNVC	GKDCAEMCKL	CQEGKYDEAK	KIDDKLKDDC	
KKNKCEKNPV	CIKWACKELG	LIDNDTVNEP	YKPISEAAK	KIKEALKKAG	KL			
v_48.020 - TM score: 0.993, Seq Recovery: 0.56, Visibility: 1.01								
MFTGAIAAIV	TPRDAAGNFD	RASLARLIDR	YIAEGIAALV	AVGTTGEAAV	LSDEERTAVV	LATLELAAGR	IPVIAGVGAS	
DTEKAIALAK	RFEGSGVVGL	LLAVPAGVKP	SQAGLIAFYT	AVADATTLPI	ILYNYPALTG	TRLEPAAVGE	LAKHPNIVGL	
KDATGDLANV	PAIRALVPAD	FRLLSGVDAV	ALEYMELGGD	GVISAVANVA	PAQMAEMAKL	AAEGKYEEAK	KIADKLAPLA	
VALFSQPNPV	PIKWALKTLG	LIDSDTLRAP	YTPATAEDKA	TIEAALKAAK	LL			

Box 10: Template and designed sequences for 1S5T

A.14 1TJE

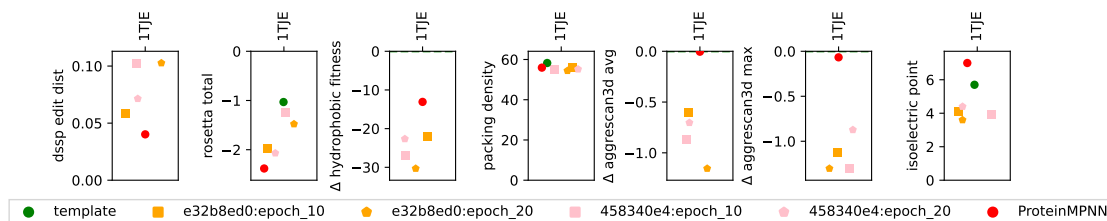


Figure A23: Focus on 1TJE

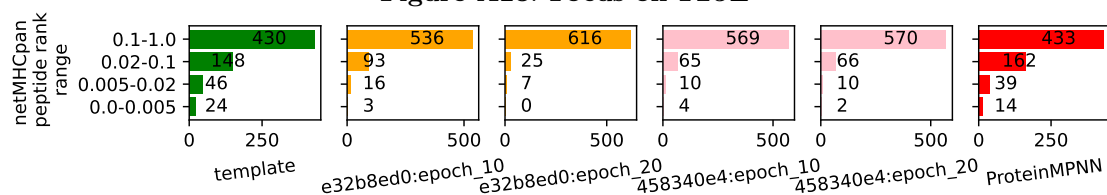


Figure A24: netMHCpan ranks for 1TJE designs

template											
MPVITLPDGS	QRHYDHAVSP	MDVALDIGPG	LAKACIAGRV	NGELVDACDL	IENDAQLSII	TAKDEEGLEI	IRHSCAHLIG				
HAIKQLWPHT	KMAIGPVIDN	GFYVDVLDLR	TLTQEDVEAL	EKRMHLEAEK	NYDVIKKKVS	WHEARETFAN	RGESYKVSIL				
DENIAHDDKP	GLYFHEEYVD	MCRGPHVPM	RFCFHFKLMK	TAGAYWRGDS	NNKMLQRIYG	TAWA					
e32b8ed0:epoch_10 - TM score: 0.983, Seq Recovery: 0.48, Visibility: 0.27											
MPTITLPDGS	TIDKDEPVS	KDIAADIGKD	EADKCIAGKA	NGELVDADDD	IEDDADVEIV	TKDDEEGQKI	IQHTCGHLVE				
AAIKKLYPDT	KIGDSPVIDG	GFYVDIDTDK	DLTDEDIKEI	EDECKKIANK	DEDIKTETVT	KEEAKKIFE	QGDELKVDEL				
DNEVDDDEKP	TLYKIGDTT	MSDGPVCPNT	KFCNFKLTD	KDDRYKNDK	NNKKLQRIKG	SAWA					
e32b8ed0:epoch_20 - TM score: 0.967, Seq Recovery: 0.42, Visibility: 0.10											
APTITLDDGS	EIDKDEPVS	DDVAKDIGDD	EADKCIAGKA	NGENVNADDD	IDDDADVEII	TDDDEEGKEI	IEDTCGDIDN				
AAIKEEDPDT	KIGDSDDIDG	GYETDIDTDK	DLTDEDIKKI	DDKRKEIADK	DEDIENEEVT	KDEAKKTFED	KGDDDGVDDM				
DNNVDDDDKP	TENKIGDDTT	KDGDPA	KNCKNFKLTD	KDDEDKDGDS	NNKTQRIKG	SADA					
458340e4:epoch_10 - TM score: 0.932, Seq Recovery: 0.46, Visibility: 0.20											
MPTITLEDGS	KIEKD	KPVSV	KDIAKDIGKD	EADKCIAGKV	NGEIVDADEK	IEEDADVEII	TKDDEEGKKI	IQDTCGHLCK			
AAIKKIYPDT	DIDDCDVIDG	GFYCDIDTDK	ELTKEDIEKI	EKECKEIVNK	DEDIKVETVT	WEEAKKIFEE	KGDEDEVDDL				
DDDVDKDEKP	TLYKIGDEV	T	KDGPVCPNT	KFCNFKLTD	TDDEYDDGDK	DNDKEIRIKG	KAWA				
458340e4:epoch_20 - TM score: 0.969, Seq Recovery: 0.43, Visibility: 0.17											
GPTIKLEDGS	EIECDEPVTI	KDIAKSIGE	EEEDCIAGIV	NGKKVDADDK	IEEDAEVKII	TKDKDGED	IRDTCADICC				
CAIKKIDPDT	KIGDCDVIEG	GFYCDIKTKK	KLTKEDIDKL	EKKCKEIANE	CKDIKVEKVT	WEEAKKIFEE	KGNDLKCDKI				
DNDIDKDDKP	NIYKIGDCIT	MGDGPVVPNT	CFCNFKLTD	ISDKYKNDK	NNKQLQRIKG	KAWC					
v_48.020 - TM score: 0.980, Seq Recovery: 0.45, Visibility: 0.76											
APTITLPDGS	TVKADRPVTP	AELAASIGAE	TAAKLIAGRV	NGKLVDASTP	IKEDAKVEFI	TKDDPEGLKI	IQHTCAHLVK				
AALKKLYPET	IFGEAPVIEG	GFYVDVLT	ELTPEDLKAL	EAEIKKIVAK	NLPIKVETVT	WEEALEIFKA	RGDFLQVDRF				
LNEVPKDEKP	TLYRIGD	T	VSKGPVCPNT	KFCQNFKLTD	ISYRYKNDP	NNPKLIRIYG	SAWA				

Box 11: Template and designed sequences for 1TJE

A.15 1XGD

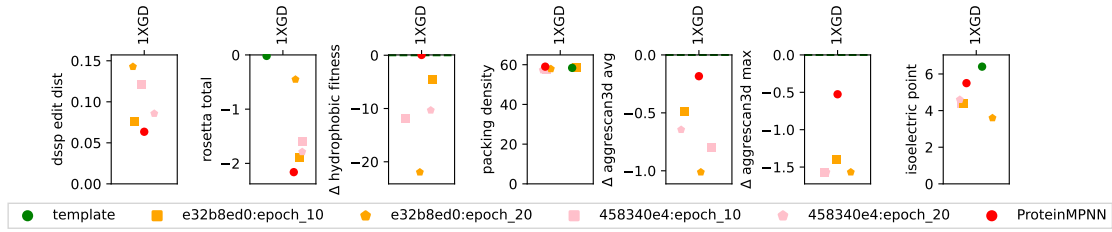


Figure A25: Focus on 1XGD

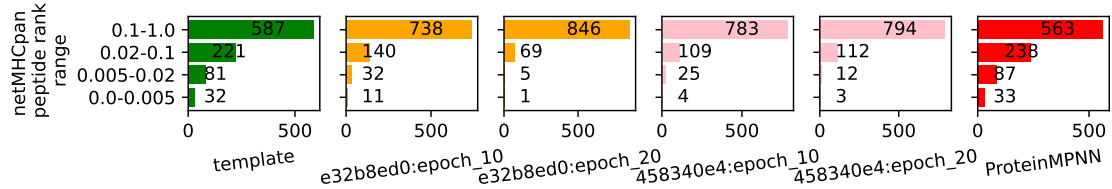


Figure A26: netMHCpan ranks for 1XGD designs

template ASRLLNNGA KMPILGLGTW KSPPGQVTEA VKVAIDVGYR HIDCAHVVQN ENEVGVAIQE KLREQVVKRE ELFIIVSKLWC TYHEKGLVKG ACQKILSDLK LDYLDLYLIH WPTGFKPGKE FFPLDESGNV VPSDTNILDY WAAMEELVDE GLVKAIGISN FNHLQVEMIL NKPGLKYKPA VNQIECHPYL TQEKLIQYCY SKGIVVTAYS PLGSPDRPWA KPDPSPLEED PRIKAIAAKH NKTTAQVLIR FPMQRNLVVI PKSVTPEAIA ENFKVDFDEL SSQDMTLLS YNRNWRVCAL LSCTSHKDYP FHEEF										
e32b8ed0:epoch_10 - TM score: 0.978, Seq Recovery: 0.44, Visibility: 0.38 MKTIKLNNNG EAPVLQFSTE GVKPEEAEEL VKEAIKEGYK GIKSAKDNGN EAEVGKAIINE AIKEGIVKVD ELFIETKIPP NKHEKDEVEA AVKESLKDGL IDYADLVLRD TPVGLKSGDD LEPKDDNGKP IFSDTTIEET WDAMEELVDK GLAKAIGVSN FSKEQIDTIL NKEGLKHKPA VNEVERNVEN IQKELTDYCK SKGIVVQSTK PEGDKNDPDK DDDDPGILDD PEIKKIADKH GKTSSQVIVK YPIEEDEIVA PENTTKEVKV ENKKVDFDFKL SDEIDIELNS YDKNRIYDD SEYKDHPNYP YKTKD										
e32b8ed0:epoch_20 - TM score: 0.959, Seq Recovery: 0.38, Visibility: 0.05 NDTLKDNNGN DTPNLEMSTD DVSADEAKDK VKKGIDEGYK GIDSAGDNDN NKEVGDGINE AIDEGKVNEE DLNISSEVDN NDHDKDDVDK AVDDDLKDLG KDKLDNVKDD DPVGKESGDD DDPKDDNGNN KESDTTNNDD WDAMEDTVDD GKANSIGVSN YNNDQIDIL NKDGKKNDPA ENEIENNVRN TEKEKNDDYK SKGINVKNDK PGDDDDDDDD DDDDDNIEDD EDIKDIADKK NKTSKQVLNK YGQEENNKVD PDNTTEEDIE ENNKCNDFEL DDEKDELD S YDKNERDDDD DDNDKHPDYP NDKDD										
458340e4:epoch_10 - TM score: 0.981, Seq Recovery: 0.46, Visibility: 0.26 KKEIELNNGT KMPTLQFSTE GVEPEKCEEL VKKAIKIGYK GIDTAKDNKN EKEVGKAIINE ALKEGIVKED ELCITTKIDN NDCEKDKVKE AVDKSLKDLG VDKLDLVLRD NPVCLKSGED DEPKDDNGNV IEGDETIEFE WEAMEELVDE GKAKAIGVSN FNKEQIDKIL NKEGLKHKPA VNEIECNVKN KQKELKKYCK EKGIVCICDK PEGCEDDEDK DEDDPGILDD PEIKKIADKK GKTRQVCIK YPIEEGCIIVK PECKTEEKIE ENYKVDFFEL TEEDKEILDS YDKNKICDKD EYKDHPNYP YKEKD										
458340e4:epoch_20 - TM score: 0.976, Seq Recovery: 0.42, Visibility: 0.13 GKEIELCNGG KMPVLQFSTE GCPAEKICEL VKKAIKIGYK GIKCAKDNKN EKEVGKAIINE CIKEGIVKED ELCITTKIPP NCCEKDDVEK CVKDSLKDLG VDDLDLVLRD TPVCLKSGED DEPKDDNGKP ICGDCKIEET WEAMEKVDN GDCKAIGVCN FCKEQIDKIL NKEGLKHKPC VNEIECNVKN CQKELIKYCK EKGIVCCCKE PEGCEDDEDK DEDDPGILDD PEIKKIADKK GKTRQVCIK WPMEEGCIIVC PECKTEEKVE ENNKCFDFEL TEEDKEILDS YDKNKICDKD EYKDHPDYP YKKKC										
v_48_020 - TM score: 0.968, Seq Recovery: 0.53, Visibility: 1.06 MKEIELNNGL KMPRLLFSTY GVPPEKITEL VKLAIKIGYK GIDSALSYGN EAEVGKAIINE ALKEGIVKRE ELFIITKIPP TRHEPGLVRE AVLEALKELQ VDYLDLVLRH LPVGLKPGEA LFPLDENGKP LPSDVILLDT WKAMQELVDE GLAKAIGVSR FSKEQIDTIL NMPGLKYKPA VNRIECNPYN IQKELIEYLY SKGIVVIAYK PLGSPNDPDK SPDDPSLLEH PEILEIAKRH GKTPRQVLK YPMERDMIVS PEMTTEEKVR ESYKVDFDFKL SEEDLKILDS FDKNRLYDTF PELKDHPLYP YKTKY										

Box 12: Template and designed sequences for 1XGD

A.16 1X0M

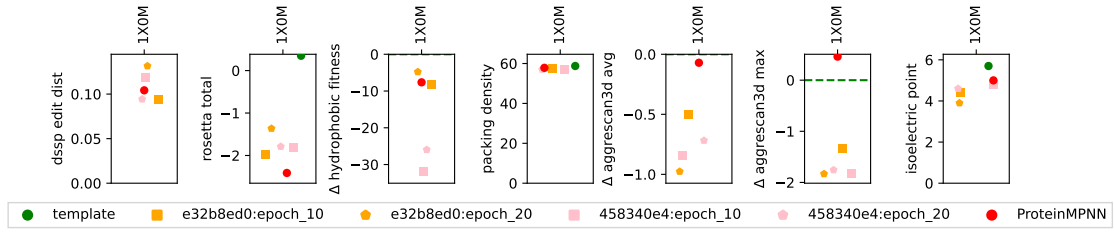


Figure A27: Focus on 1X0M

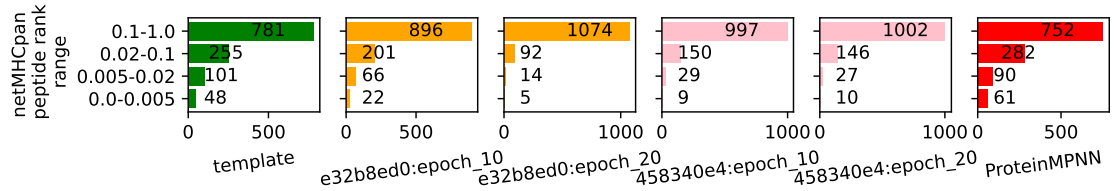


Figure A28: netMHCpan ranks for 1X0M designs

[illegible]

Box 13: Template and designed sequences for 1X0M

A.17 2BK8

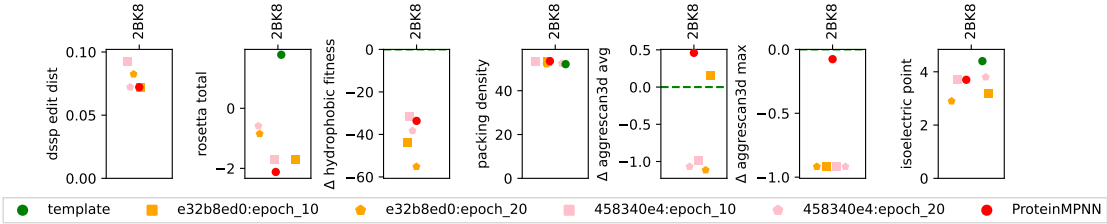


Figure A29: Focus on 2BK8

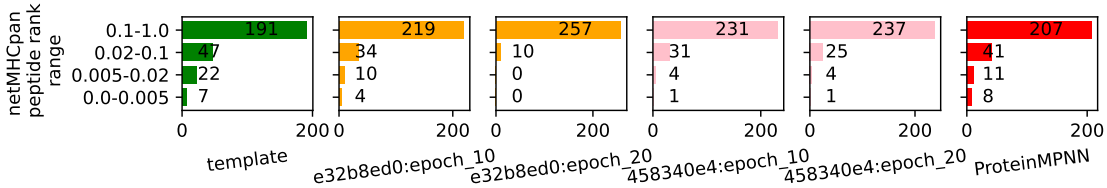


Figure A30: netMHCpan ranks for 2BK8 designs

template	
GAMVSGQIMH AVGEEGGHVK YVCKIENYDQ STQVTWYFGV RQLENSEKYE ITYEDGVAIL YVKDITKLDD GTYRCKVVND	
YGEDSSYAEL FVKGVRE	
e32b8ed0:epoch_10 - TM score: 0.931, Seq Recovery: 0.43, Visibility: 0.48	
SGPTAGSISD QTGTNGGSVT FTCNISGDDD STTWTWSKGS TTLSNSSKYS ISSSSGVSTL TVNNITAADN GTYTCNVTTD	
SGSDSASANL TVDGVNA	
e32b8ed0:epoch_20 - TM score: 0.923, Seq Recovery: 0.39, Visibility: 0.00	
GGPTAGDIDD EDCDDGGSVE FTCDIDGDDD TTDVDWKKGD EDLSNDDKND ISDSDGTSKL TVNDCTDDDD GTYTCNVTTD	
TGSDSSSGNV DVDGVDD	
458340e4:epoch_10 - TM score: 0.924, Seq Recovery: 0.54, Visibility: 0.17	
GGPKAGKIED KTCEEKGSVT FTCKIEGDDD STKVTWYKGD TELENNDKYD ITYEDGVATL TIKNCTKDDD GTYTCKVEDD	
EGSDEAKANI DVEGVDD	
458340e4:epoch_20 - TM score: 0.925, Seq Recovery: 0.55, Visibility: 0.17	
GGPKCGKIED KTCEEKGSVT FTCKIEGDDD STKVTWYKGD KELENSDKYD ITYEDGVATL TIKNCTKEDD GTYTCKVEDD	
EGSCEAKANI DVEGVDE	
v_48_020 - TM score: 0.949, Seq Recovery: 0.44, Visibility: 0.66	
SGPVAGSISD QTGTLGGSVT FTCNISGDNE STTWTWSKGS TTLSNSSKYS ITYSSGVATL TINNITAADN GTYTCNVTTD	
SGSDSASAKL TVPGVDP	

Box 14: Template and designed sequences for 2BK8

A.18 2QMT

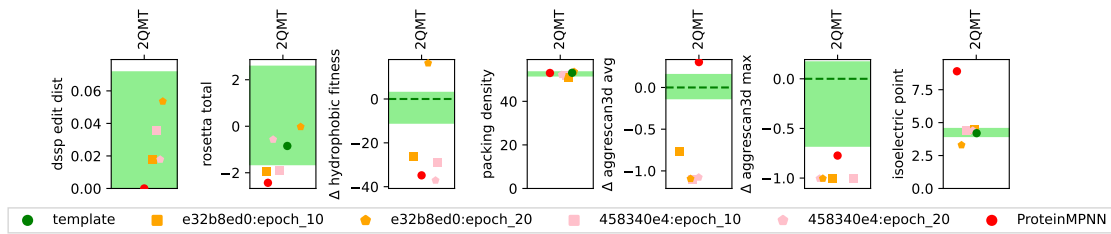


Figure A31: Focus on 2QMT

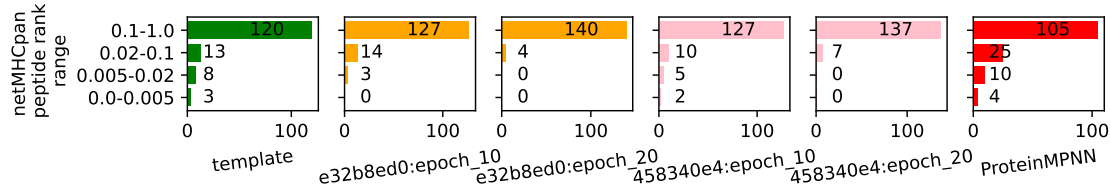


Figure A32: netMHCpan ranks for 2QMT designs

template

MQYKLILNGK TLKGETTTEA VDAATAEKVVF KQYANDNGVD GEWTYDDATK TFTVTTE

e32b8ed0:epoch_10 - TM score: 0.971, Seq Recovery: 0.52, Visibility: 0.27

ATYTLKINGK NVKGEKSVEA DNAEEAKEKF EKYCKDNGIN GEWTFDEETK TATVTE

e32b8ed0:epoch_20 - TM score: 0.924, Seq Recovery: 0.45, Visibility: 0.00

ADYTNTNGK DDKGTTSTDD DNDDDAKDDF DKNNKDNGVD GDETWDDDTK TATSTE

458340e4:epoch_10 - TM score: 0.972, Seq Recovery: 0.50, Visibility: 0.64

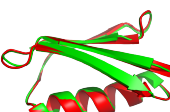
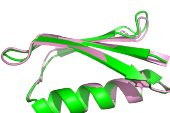
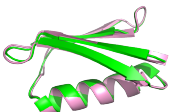
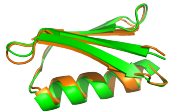
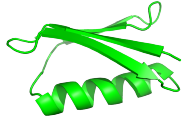
AKYKLKIEGK DIKGEKEVEA ENEEAKEIF EKWKENGID GEWTYDEETK TFTVKE

458340e4:epoch_20 - TM score: 0.944, Seq Recovery: 0.52, Visibility: 0.00

GKYKLNINGK EIKGEKECEC ENIEEKEKVF KEWKENGID GEWTCDEETK TFTCTE

v_48_020 - TM score: 0.974, Seq Recovery: 0.57, Visibility: 1.27

ATYTLNLNGK TVRGTLTVEA ANAAEAQKQF KAYCKSHGIN GTWTFDPSTK TFTVTTE



Box 15: Template and designed sequences for 2QMT

A.19 3O6A

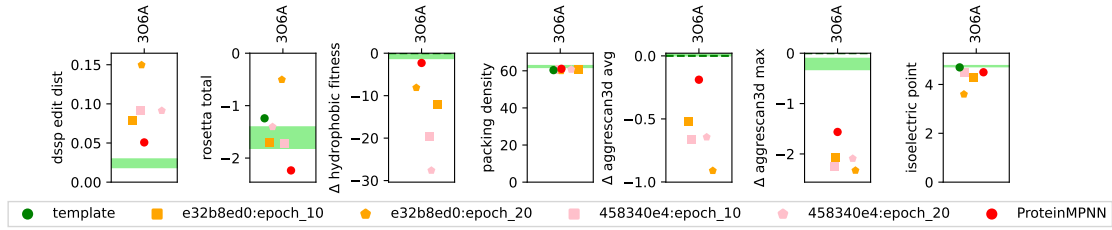


Figure A33: Focus on 3O6A

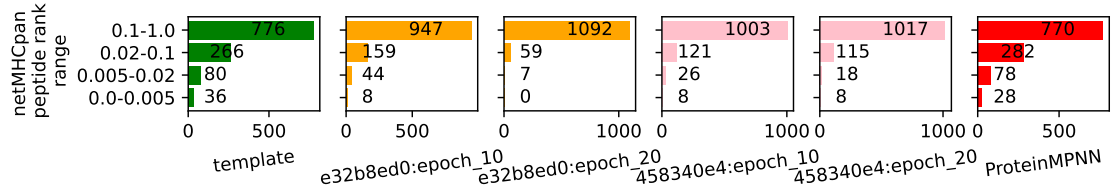


Figure A34: netMHCpan ranks for 3O6A designs

template	AWDYDNNVIR	GVNLGGWFVL	EPYMTPSLFE	PFQNGNDQSG	VPVDEYHWTQ	TLGKEAALRI	LQKHWSTWIT	EQDFKQISNL	
	GLNFVRIPIG	YWAFQLLDND	PYVQGQYQL	EKALGWARKN	NIRVWIDLHG	APGSQNGYDN	SGLRDSYNFQ	NGDNTQVTLN	
	VLNTIFKKYG	GNEYSDVVIG	IELLNEPLGP	VLNMDKLKQF	FLDGYNLSRQ	TGSVTPVIIH	DAFQVFGYWN	NFLTVAEGQW	
	NVVVDHHDYQ	VYSGGELSRL	INDHISVACN	WGWDAKKESH	WNVAGEWSAA	LTDCAKWLNQ	VNRGARYEGA	YDNAPYIGSC	
	QPLLDISQWS	DEHKTDTRRY	IEAQLDAFEY	TGGWVFWSWK	TENAPEWSFQ	TLTYNGLFPQ	PVTDQRFPNQ	CGFH	
e32b8ed0:epoch_10 - TM	score: 0.988, Seq Recovery: 0.48, Visibility: 0.45								
	NWDWDNDKIK	GVNLGGWFVD	VPDLRPSRWD	KFKDGGDTSN	IPKDQEEWDK	KKGEEEAKE	RDEWLSTWIT	EEDFKEIKDD	
	GLNFVRIPIG	EWEIEKEEND	SRVTGGEKIY	DKMLEWCKKN	NIKANLDLHG	VPGSQNGWYD	SGESGKLNWQ	EGDNRERSLE	
	VLDKVAKKYC	GDKYKDVVIG	LETANEPDGS	KLDLDKLKEF	EKECYDRLKA	TGSDTIVVID	DAGQETGFW	DFLTEENGET	
	NVVLRRHYQ	IDSCEKLAKN	TDEKIKDICD	EGKKITKEKH	KSVVGETSAA	TDDCCDDYNG	VGKGSRTGK	NDGCPYIDDC	
	EPRKDISNWT	DEKDDIKKY	IEAQIEAANA	SGGYSFWNWK	SENCWENSYK	TLKDNGLFPN	PIDKKDYPNL	CNFE	
e32b8ed0:epoch_20 - TM	score: 0.977, Seq Recovery: 0.34, Visibility: 0.06								
	NWDWDNDKAN	GVNNGGWNK	EADKDKSKYD	GEKDGDDDSG	ISKDQEEYDK	KKGDDKAKEE	EDKWRDTKVT	KDDYDDIKDD	
	GLNVVRIDFG	DWEIEKDDDD	NNVTGGQEKI	DEDMENCKDN	DIKATLDNHG	SGSGNNGGED	DGKSGTNGWE	DGDNADKDND	
	VNDKINKEYG	GDKNKDVVDG	IETENEPDGD	KLDQDKVDDY	DDKCNKNKKD	TGSDTDTED	DNGNDYGADD	DKLTEDDGGD	
	NVVIGRHWGD	DNNGDDIAKN	DDEKIDDDCK	EGEDAKKEKH	KNVVTETSSA	DDCCDDENG	VGKGCADGT	NDGCDKIDDC	
	DDKKDEKDDT	DDDKENRKKK	IEAQIDAANE	SGGYDFWDW	SEDCWDNSYK	DQKNDGDPD	PIDDKENPDQ	CGYD	
458340e4:epoch_10 - TM	score: 0.989, Seq Recovery: 0.46, Visibility: 0.29								
	NWDWDNDKIK	GVNLGGWFVD	EPWIRPSRYK	KFKKGDDSN	IPKDQEEYCK	KLKKECKCEE	RDKWLSTWIT	EEDFKEIKDD	
	GLNVVRIPIG	EWIEIEKEEND	SNVTGCEEKL	DKCLDWCKKN	NIKVNIDLHG	CPGSQNGDTD	SGKSGVLNWK	EGDNKEKTKE	
	VLKKLCEKYC	GDKYKDVVIG	LEVCEPNEDK	KLDKDKVKEF	EKECYEDLKA	TGSDCNVIID	DGGEVGEWD	DFLTEEDGCN	
	NVIINRDYYQ	IDSCEKLKKN	TDEKIKDICD	EGKKRKEKH	KVICGEWSAA	TDDCCDWYNG	VGKGSCTEGE	NNGCEKIDDC	
	EPKKDIKNWT	DEKDDIKKY	IEAQIEACNS	FGGYCFWNWK	TENCWKNSYK	ELKDNGLIPN	PIDKKDYPGL	CGFE	
458340e4:epoch_20 - TM	score: 0.985, Seq Recovery: 0.44, Visibility: 0.22								
	NWNWDTDKIK	GVNCGGWFL	EPDKRPSLFK	PFKKGDDESG	IPKDQNEWCK	KKGEECKKE	WDKWLSTWIT	EEDFKEIKDD	
	GLNVVRIPIG	WCCIEKEEND	PWVTGCEEKL	DKCLDWCKKN	NIKNIDLHG	CPGSQNGDYD	SGDSGILNWK	EGDNKECKKE	
	CFKKLCEKYC	GDKWDDVVIG	LECCNEPDGK	KLDKDKWKEW	CKECYEDLKA	TGCDCTCIH	DGGERGAWD	DFLTEEDGCT	
	NVIIDRDWED	IDSCEKIKKN	TDEKIKDICD	EGKETCKEKH	KCICGEWSAA	TDDCCDWYNG	VGKGSCTEGK	NNGCEKIDDC	
	EPKKDIKNWS	DEKDDIKKY	IEAQIEACNK	FGGYCFWDWK	TENCWKNSYK	TLKDNGLIPN	PIDKKDYPNL	CNFE	
v_48_020 - TM	score: 0.990, Seq Recovery: 0.52, Visibility: 0.91								
	NWDWETDTIR	GVNLGGWFVL	KPYLRPSLFE	PFKVGDDTST	IPRDQYEECK	KLKKEKAKEI	LDKHWSTWIT	AEDFKEIAAD	
	GLNFVRIPIG	YWAIKLLPDD	PFVTGEMEYI	DKALEWCKEN	NIKAILDLHG	VPGSQNGWYD	SGLSGVLNFK	EGDNLDFALE	
	VFDIVAKKYC	GDKYKDVVIG	LELINEPYGA	KLDRDKLLAF	YERGYNMLKA	TGSKVTFIVD	DAWQPFQYWD	DFLTPADGKE	
	NVVLSSLHLYL	IDSNEKLAMT	MEEKIAAACA	FGREARKLKH	KVIFGEFSAA	LTDCCPWYFG	VGKGSIEGK	YDGGKFIGEC	
	EPLLDIANWS	EQQLIEDIRRF	IEAQLEAAEA	VGGYIFWAWD	SENCPLLSYK	TLRAAGIYPD	PLTARRYPGL	CGFP	

Box 16: Template and designed sequences for 3O6A

A.20 3TIP

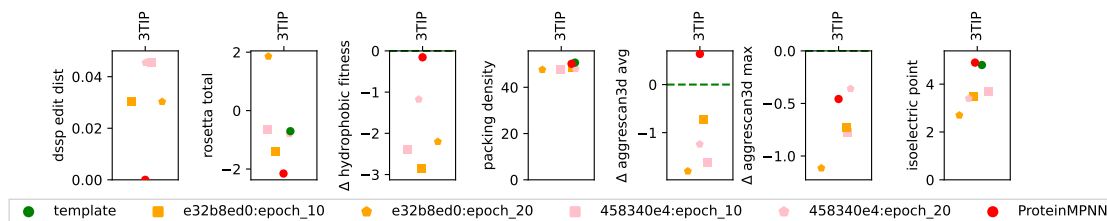


Figure A35: Focus on 3TIP

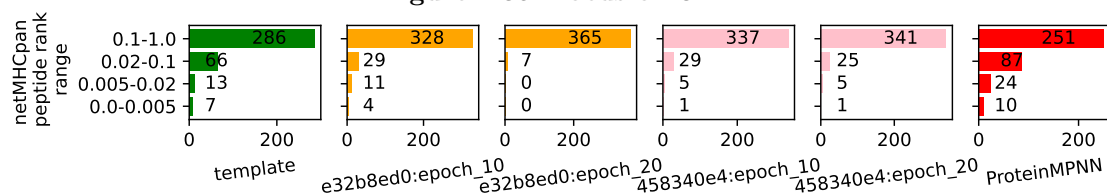
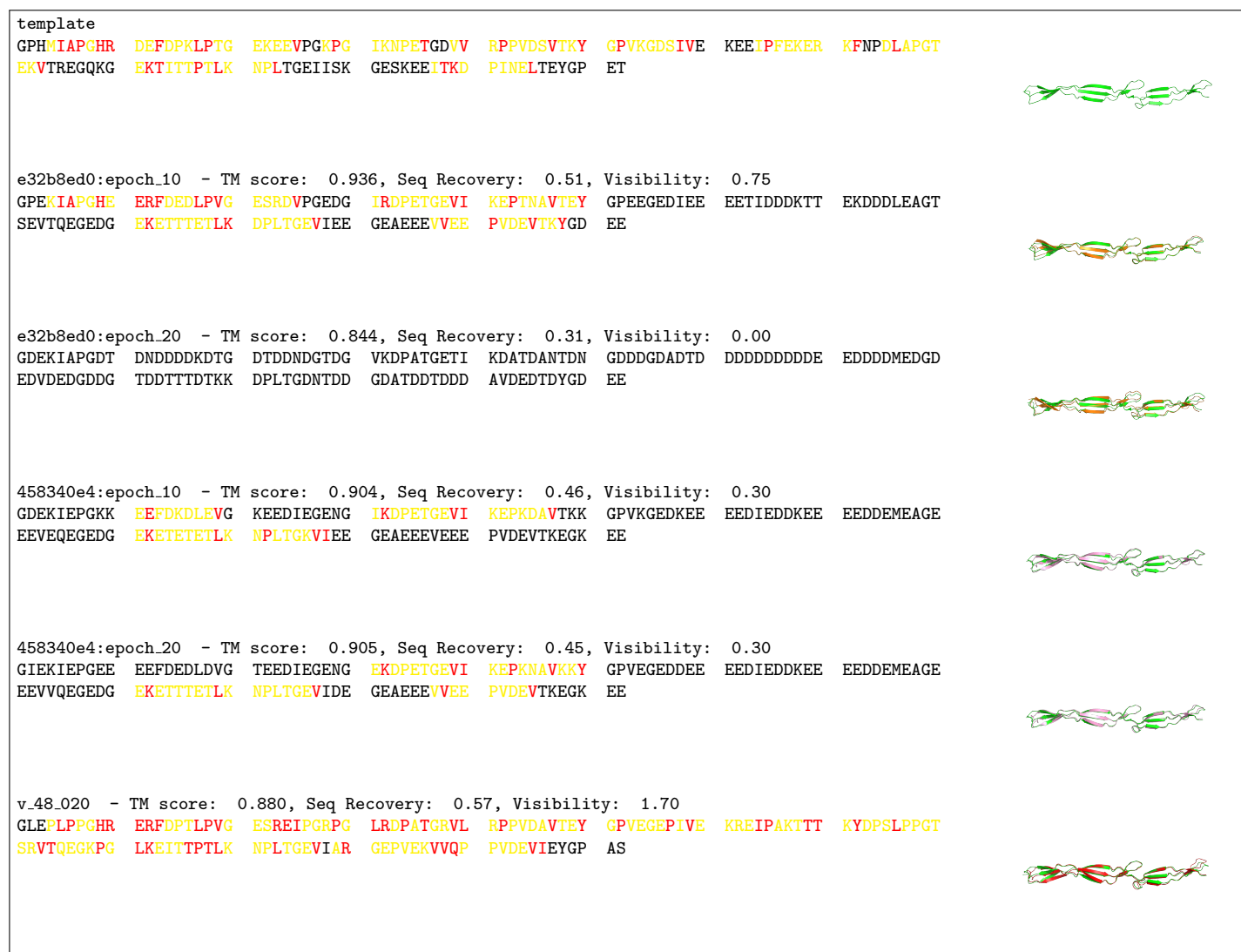


Figure A36: netMHCpan ranks for 3TIP designs



Box 17: Template and designed sequences for 3TIP

A.21 3WOY

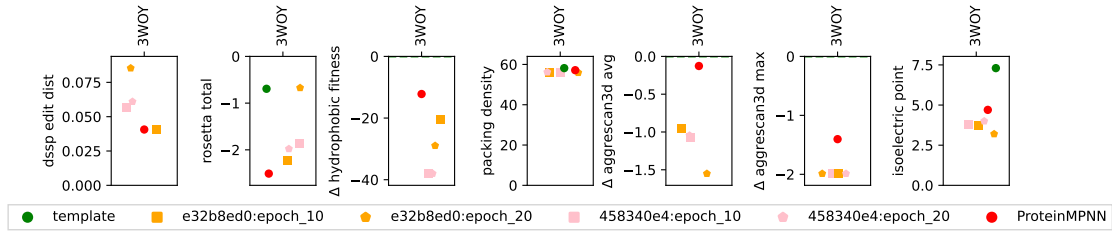


Figure A37: Focus on 3WOY

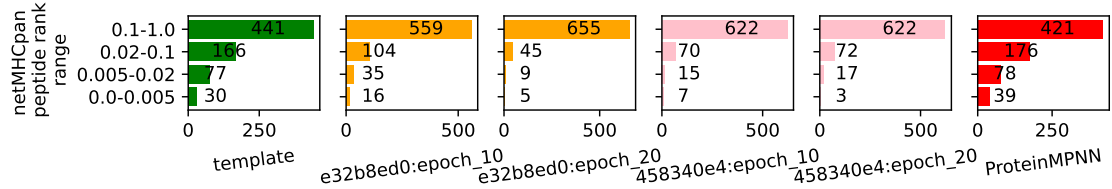


Figure A38: netMHCpan ranks for 3WOY designs

template GAVDEDDFIK AFTDVPSIQI YSSRELEETL NKIREILSDD KHDWDQRANA LKKIRSLVA GAAQYDCFFQ HLRLLDGALK LSAKDLRSQV VREACITVAH LSTVLGNKFD HGAEAIPTL FNLVPNSAKV MATSGCAAIR FIIRHTHVPR LIPLITSNCT SKSPVRRRS FEFLDLLLQE WQTHSLERHA AVLVETIKKG IHDADAEARV EARKTYMGLR NHFPGAEATL YNSLEPSYQK SLQTYL	
e32b8ed0:epoch_10 - TM score: 0.950, Seq Recovery: 0.36, Visibility: 0.48 EEEEEEEMDE ECDDVEETEI ESEEDLEEL DEIGEILSDP NTPKEERIEA LREIRSLNA GAHEYDEFYE ELEELNDELN NALNDEDPEV VKEAAVTVC IAKKLGDKAK DTAEAVAPTL LKNVGDDEE VAEAGITALK EIVKNVKNPN LVPLITDACS DEDTKVRTEA YRLLAEALDT WDKEDLKDHE DDIADTIKAG VTDEDAAKE EAKKAFDGLK KHFPEKAKEM RDGLDEDTQK ELAPAS	
e32b8ed0:epoch_20 - TM score: 0.931, Seq Recovery: 0.26, Visibility: 0.13 DDDDDEEQDK DYDDVDDTEI DSDDNDNDN DDINDTLDD NNDKDEQKDA LDDINSNINA GAGDNDSDDD NNDKNNDGLN KALNDDSDV VKSGCTTVGA IAKEEGDNAD KTADDVNPTL QDNVGDDEE DAEAGETAIE EINKNSKNKN NIDDITDNTK DDDEDVRTEA NDDLADALDT WDKEDLEDK DKINEAVDEG LEDDEDKAK EANEAFDGVK KHFPEADET EAKKDEDTKE DTKDEE	
458340e4:epoch_10 - TM score: 0.955, Seq Recovery: 0.30, Visibility: 0.21 EEEEEEEMDE ECDDVEEVEI ESEEDLEEL DEIKEICSDK NNDEERIEA LKKICSLINA GAHEYDEFYE ELEELKECLC NCCDEDPEV VKECCVCCGC ICKKLKDKAK ELADAIFFVC LKNGDDEE KSESGICCK EIKNCKNPN LVDKICDADC DEDTKVRIEA WDLLEALDT WEKEDLEGK EKICDCVKKG CEDEDEECK KAKEAFKGVK KHFPEEAEM KKGLDEETQE DLEDEE	
458340e4:epoch_20 - TM score: 0.976, Seq Recovery: 0.31, Visibility: 0.19 KEEKEECEK EYEDVEVEI ESEEDLKEEL DEIGDILDDP ETDKEERIEA CKDICSLINA GAGEYDVWKE CLDKLCDCLC NCCDEDPEC VKACCVCGC ICKECGDDFK NCAEKCIPVL IKNVGDENEE KSEAGDVAIE EIKNCKNPN LIPKICEACE DENEKVRTKA WDWLATALDT WKKEDLEGK DKICECIKG LKDENEECRE EAIKAFKGVK KHFPEEADEM KEELEEEIQE ECKEEE	
v_48_020 - TM score: 0.974, Seq Recovery: 0.38, Visibility: 1.09 REEVLKEMDK AYTDVPKVTI NSEEDLVKEL DEIGEILADP STPPEERTAA LNRIRSLNA GAEKYPAFYE RLEKLIPQLQ LALSPPDEL RKAATITVAA IAKTLGERAK KLALAVFPTL LKGLPDPDPT VAEAAALAE EIITHVHVE LIPLILDACQ SPDTRVRIEA YRLLALALET WPKELLLAHE TAIVETIKAG LQDPDPAARE EAIKAFKGLK KHFPEAAAAM LASLPPEIQA ALAPPS	

Box 18: Template and designed sequences for 3WOY

A.22 4RQG

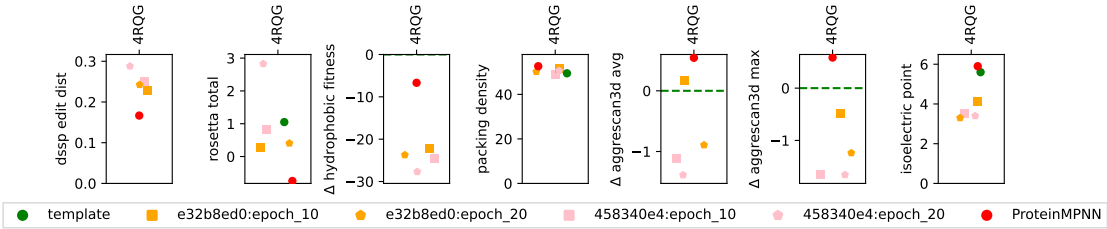


Figure A39: Focus on 4RQG

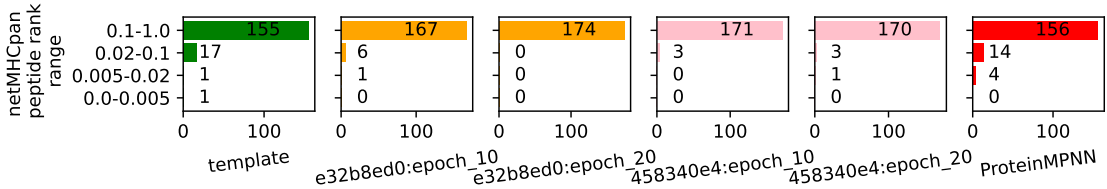
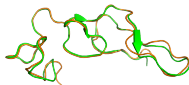
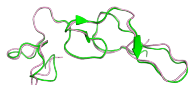
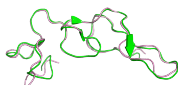


Figure A40: netMHCpan ranks for 4RQG designs

template	ECDCSSPENP	CCDAATCKLR	PGAQCGEGLC	CEQCKFSRAG	KICRIPRGDM	PDDRCTGQSA	DCPRYH	
e32b8ed0:epoch_10	- TM score: 0.903, Seq Recovery: 0.49, Visibility: 0.50							
ERDCSDSSNP	CCDPETGQLK	SGCDCCEGPC	CSNCKYCAAG	TVCRTPTGSA	PNDTCTGTCC	CCPKVS		
e32b8ed0:epoch_20	- TM score: 0.892, Seq Recovery: 0.46, Visibility: 0.00							
DKDCSDDNP	CCDKDTGQLK	DGNDCCCEGDC	CENCQNCDAG	TVCDTPTGND	KDDTCTGDCC	CCPKEN		
458340e4:epoch_10	- TM score: 0.873, Seq Recovery: 0.55, Visibility: 0.00							
ECDCSDDNP	CCDKETCKLK	EGCDCCEGEC	CENCKYCEKG	TVCDEPEGDE	PDDTCTGECC	CCPKES		
458340e4:epoch_20	- TM score: 0.877, Seq Recovery: 0.50, Visibility: 0.50							
ECDCEDDNE	CCDKETCKLK	EGCECCEGDC	CEDCKYKEEG	TICEEPEGDE	EEDKCTGDCC	CCPEEE		
v_48_020	- TM score: 0.913, Seq Recovery: 0.58, Visibility: 2.00							
SCDCPDPSNP	CCDPATCKLK	PGYQCCSGPC	CENCKYKPAG	TLCRTPTGNE	PPDYCTGTCC	CCPRVY		

Box 19: Template and designed sequences for 4RQG

A.23 4KW4

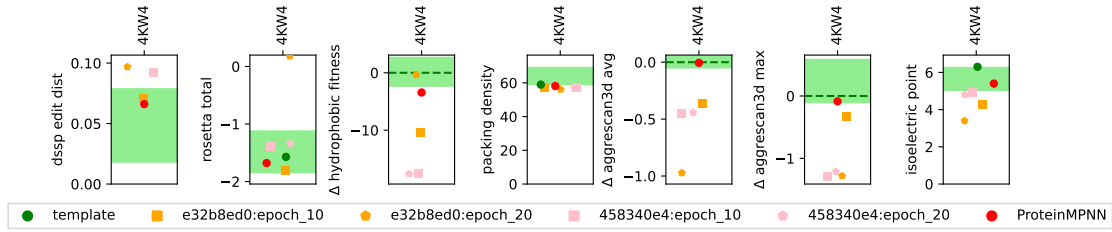


Figure A41: Focus on 4KW4

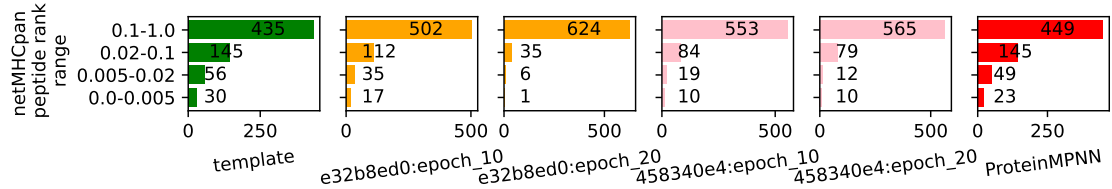


Figure A42: netMHCpan ranks for 4KW4 designs

template SKGEELFTGV VPILVELDGD VNGHKFSVSG EGEGDATYKG LTLKFICTTG KLPVPWPTLV TTLXXXVQCF ARYPDHMKQH DFFKSAMPEG YVQERTIFFK DDGNYKTRA E VKFEGDTLVN RIELKGIDFK EDGNILGHKL EYNNHHKVVY ITADKQKNGI KVNFKTRHNI EDGSVQLADH YQNTPIGDG PVLLPDNHYL HTHSKLSKDP NEKRDMVLL EFVTAAGITL	
e32b8ed0:epoch_10 - TM score: 0.992, Seq Recovery: 0.68, Visibility: 0.72 SEGAELEFGV VPVVVELDGD VNGKKFSVKG EGEGDATEGR IDLKFVCTTG ELPVPWPTLV TTLXXXIDCF ARYPEHMKDH DFFKSAMPEG YVRERTNDWE DDGKDDSRGE VKMDGDTLVN RIELKGTDYK EDGNILGHKL KYSAGSSKKN VTADEANNGI KADFTEELEV EDGSTQKVKN DEKNTPIGDG PVLLPKDHYH ETSAELSKDP NEDRDHMLR ETIEAGGIEE	
e32b8ed0:epoch_20 - TM score: 0.978, Seq Recovery: 0.50, Visibility: 0.34 SDGDDLFDGK VIDNDSGDG VNGEKYSNDG EGEGDATNGK TDEDNDNTTG ENPVDWETED DNNXXXIDCN SKYPEHMKDH DFFKSAMPEG NKDEKDNDD DDGNDDSRGE VKMDGDTLVN ENEKDGDDYD DDGNILGHKL KDSCGSSKKN IDADEKNNGI KEDGTDKNDV EDGSTQSDN DGEKTPIGDG DVKLPDHYD DTSADNSDDD DEDKDNKTT ETEEAGGIDE	
458340e4:epoch_10 - TM score: 0.992, Seq Recovery: 0.64, Visibility: 0.53 SEGEKLFEGK VPIVVEIDGD VNGKKFKVKG EGECDATGKG IDLKCVCCTG KLPVPWPTLV TTAXXXIDCC TKYPEHMKDH DFFKSCMPEG YVDERTNDWE GDGDKTRGE VKMDGDTLVN RIELKGTDYK EDGNILGHKL KDSCGSSKKN IKADEKNNGI KGDFTEELEV EDGSTQKCKN KEKNTPIGDG PVKLPKDHYD ETDCKLSKDK NEKRDHCVIN ETEVAGGIEE	
458340e4:epoch_20 - TM score: 0.992, Seq Recovery: 0.61, Visibility: 0.47 CEGKLFEGK VPIVVECDGD VNGKKFKVKG EGECDATGKG IDLKCVCCTG ECPVPWPTLV TTLXXXIDCC TKYPEHMKDH DFFKSCMPEG YIDERTNDWE GDGDKTRCE VKMEGDTLCN RCELKGTDYK EDGNILGHKL KDSCGSSKKN IKADEKNNGI KKDFTEELEV EDGGVQKCKN KGKNTPIGDG PVKLPKDHYD ETKCKLSKDE NEKRDHCCIE ETEVAGGIEE	
v_48_020 - TM score: 0.986, Seq Recovery: 0.73, Visibility: 0.88 SAGAEALFTGV VPVVVELDGD VNGHKFKVVG EGEGDATTGR LVLKFVCTTG KLPVPWPTLV TTLXXXIQCF ARYPHMAQH DFFKSAMPEG YVRERTLDVE DDGFTTRA E VKMEGDTLVN RIELKGTDFK PDGNILGHKL KYDYGSSRVH ITADPARNGI LAEFTEELPV EDGSTQKVKH KERNTPIGDG PVLLPEPHYH ETAAKLSKDP NEKRDMVLE ETIVAAGIPE	

Box 20: Template and designed sequences for 4KW4

References

- Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Židek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, pages 1–3, May 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL <https://www.nature.com/articles/s41586-024-07487-w>.
- J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, October 2022. doi: 10.1126/science.add2187. URL <https://www.science.org/doi/10.1126/science.add2187>.
- Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348, July 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-32007-7. URL <https://www.nature.com/articles/s41467-022-32007-7>.
- Sarel J. Fleishman, Andrew Leaver-Fay, Jacob E. Corn, Eva-Maria Strauch, Sagar D. Khare, Nobuyasu Koga, Justin Ashworth, Paul Murphy, Florian Richter, Gordon Lemmon, Jens Meiler, and David Baker. RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. *PLOS ONE*, 6(6):e20161, June 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0020161. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0020161>.
- John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Židek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL <https://www.nature.com/articles/s41586-021-03819-2>.
- Andrew Leaver-Fay, Michael Tyka, Steven M. Lewis, Oliver F. Lange, James Thompson, Ron Jacak, Kristian W. Kaufman, P. Douglas Renfrew, Colin A. Smith, Will Sheffler, Ian W. Davis, Seth Cooper, Adrien Treuille, Daniel J. Mandell, Florian Richter, Yih-En Andrew Ban, Sarel J. Fleishman, Jacob E. Corn, David E. Kim, Sergey Lyskov, Monica Berrondo, Stuart Mentzer, Zoran Popović, James J. Havranek, John Karanicolas, Rhiju Das, Jens Meiler, Tanja Kortemme, Jeffrey J. Gray, Brian Kuhlman, David Baker, and Philip Bradley. Chapter nineteen - Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. In Michael L. Johnson and Ludwig Brand, editors, *Methods in Enzymology*, volume 487 of *Computer Methods, Part C*, pages 545–574. Academic Press, January 2011. doi: 10.1016/B978-0-12-381270-4.00019-6. URL <https://www.sciencedirect.com/science/article/pii/B9780123812704000196>.
- Suyue Lyu, Shahin Sowlati-Hashjin, and Michael Garton. Variational autoencoder for design of synthetic viral vector serotypes. *Nature Machine Intelligence*, 6(2):147–160, February 2024. ISSN 2522-5839. doi: 10.1038/s42256-023-00787-2. URL <https://www.nature.com/articles/s42256-023-00787-2>.
- Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pages 1–8, January 2023. ISSN 1546-1696. doi: 10.1038/s41587-022-01618-2. URL <https://www.nature.com/articles/s41587-022-01618-2>.
- Donatas Repecka, Vyktas Jauniskis, Laurynas Karpus, Elzbieta Rembeza, Irmantas Rokaitis, Jan Zrimec, Simona Poviloniene, Audrius Laurynenas, Sandra Viknander, Wissam Abuajwa, Otto Savolainen, Rolandas Meskys, Martin K. M. Engqvist, and Aleksej Zelezniak. Expanding functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence*, 3(4):324–333, April 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00310-5. URL <https://www.nature.com/articles/s42256-021-00310-5>.

Michael J Stam and Christopher W Wood. DE-STRESS: A user-friendly web application for the evaluation of protein designs. *Protein Engineering, Design and Selection*, 34:gzab029, February 2021. ISSN 1741-0126. doi: 10.1093/protein/gzab029. URL <https://doi.org/10.1093/protein/gzab029>.