CAMP: COMBINATORIAL ENGINEERING OF PROTEINS

Manvitha Ponnapati*
Massachusetts Institute of Technology, USA
manvitha@mit.edu

Sapna Sinha*
Yang Tan Collective,
Massachusetts Institute of Technology, USA
ssapna@mit.edu

Brian Lynch
Independent
bal87@cornell.edu

Edward S. Boyden†
Yang Tan Collective,
Massachusetts Institute of Technology &
Howard Hughes Medical Institute, USA
edboyden@mit.edu

Joseph Jacobson†
Massachusetts Institute of Technology, USA
jacobson@media.mit.edu

ABSTRACT

Protein recombination has long been a key method in protein engineering to diversify and optimize sequences. We enhance and evolve this approach by using a protein language model, where we found that when log likelihood in the language model is represented as a spline, abrupt transitions in the spline identify crossover sites for designing recombinant protein libraries. We use these sites to guide recombination of sequence blocks from evolutionarily related sequences using MCMC sampling. Language models also enable generation of novel recombinant blocks beyond traditional MSAs increasing diversity, while a direct preference optimization algorithm is used to fine-tune these blocks for reduced immunogenicity. This method integrates modern deep learning architectures with traditional protein engineering techniques to improve success rate of the libraries designed for wetlab verification.

1 INTRODUCTION

Engineering proteins with desired functionality is fundamental to modern biotechnology, impacting areas ranging from therapeutics to synthetic biology. In recent years, significant progress in protein engineering has been driven by advances in foundational models that capture protein structure and sequence relationships(Jumper et al. (2021b)Dauparas et al. (2022)Anishchenko et al. (2021)). These advancements have enabled protein engineers to design novel proteins that are far more diverse than any previously known sequences (Ruffolo et al. (2024)Hayes et al. (2024)), demonstrating that machine learning models can effectively represent the complex sequence–structure–function relationships of proteins.

Despite advances in protein structure prediction and inverse design, designing novel proteins efficiently across all protein families remains a challenge. Current protein engineering pipelines generate diverse protein sequence libraries by inverse sampling sequences from protein structures, scaffolding active sites, and leveraging structure prediction methods (Anishchenko et al. (2021)). Metrics like pLDDT, TM-score, and others are then used to filter and rank candidates for wet-lab verification(Jumper et al. (2021a)Zhang & Skolnick (2004)Bryant et al. (2022)). However, these metrics don't correlate well with protein function due to the degeneracy in protein structures, where similar structures can arise from diverse sequences with varying functionality (Alley et al. (2019)).

*Contributed equally to this work. Authors agreed ordering can be changed for their respective interests.
†Advised equally to this work.

Protein function is often encoded in its evolutionary history, as captured in multiple sequence alignments (MSAs) of closely related protein sequences. Traditional methods like SCHEMA (Mateljak et al. (2019)) utilize this information by identifying optimal cut sites in protein sequences and recombining blocks from closely related sequences to create novel variants. We show that abrupt transitions in average log likelihood scores correlate with optimal crossover sites identified by SCHEMA, without the need for structure information. We utilize these crossover sites to implement an MCMC sampler to recombine blocks from various parent proteins, creating a diverse chimeric protein library. We also propose a method to integrate block-based combinatorial sampling with direct preference optimization (DPO) for aligning ESM-MSA-1b to produce recombination blocks with reduced MHC Class I Epitopes. Reducing the immunogenicity of protein sequences is highly valuable across therapeutic pipelines. Multiple molecular mechanisms—such as TAP transport, proteolytic cleavage sites, MHC Class I binding, and T-cell recognition—can trigger immune responses Neefjes et al. (2011), making immunogenicity prediction a challenging in silico problem. The limited availability of labeled data further restricts the use of sequence-based diffusion models or inverse design methods Sánchez-Lengeling & Aspuru-Guzik (2018). However, reinforcement learning provides a promising path forward by enabling combinatorial reward functions that can guide sequence proposals toward specific objectives Olivecrona et al. (2017).

2 Results

2.1 Using Protein Language Models to Identify Crossover Sites for Recombination

SCHEMA Mateljak et al. (2019) identifies optimal recombination crossover sites by selecting positions that minimize structural disruption. A contact between residues is considered disrupted if the amino acids at positions i and j are not found together in any of the parent sequences used for designing the chimeric libraries. We compared the performance of the crossover sites determined using our spline-based approach to those from SCHEMA, measuring both contact disruption and mutation distance from the most similar parent sequence. We used the protein structure with PDB ID 1g68, and the parent sequences for chimeric design are the same as those used in Mateljak et al. (2019). We generated 2000 chimeric sequences using the parent sequences and random choice of chimeric blocks. Figure 1 shows the disruption factor and mutation distance trade off of randomly generared chimeric sequences using the cut sites from SCHEMA and our approach. Figure 2 shows that the diversity of the sequences generated can be improved while preserving the disruption factor by sampling the chimeric blocks from protein language models.

2.2 Reinforcement Learning for Immune-Evasive Protein Design

Our DPO aligned MSATransformer Rao et al. (2021) eliminates predicted MHC Class I binding epitope regions for alleles HLA-A02:01 in channelrhodopsin compared to the baseline pre-trained model as demonstrated by Figure 3b. We evaluated the performance of the DPO aligned version by generating 100 samples at a randomly chosen predicted epitope region using netMHCPan4.1 Hoof et al. (2009) and masking the corresponding residues to generate new samples. Figure 3c shows the reduction in the number of total binders in the DPO finetuned version of the MSATransformer compared to the pre-trained version of it.

3 Methods

3.1 Leveraging ESM2 embedding to find optimal crossover sites.

In SCHEMA, Arnold et al. (Mateljak et al. (2019)) approached the optimal crossover sites for creating recombinant proteins by minimizing the disruption to the structure contact map. Here, we propose that optimal crossover sites are encoded in the evolutionary history of a protein and can be identified using pre-trained protein language models (PLMs) like ESM2. We use the log-likelihood scores computed by the ESM2 model to assign a per-residue score. Specifically, we obtain the log-likelihood of each amino acid in the input sequence conditioned on the rest of the sequence. This produces a noisy signal, which we then smooth using a spline fit, and we identify local maxima as candidate crossover sites.

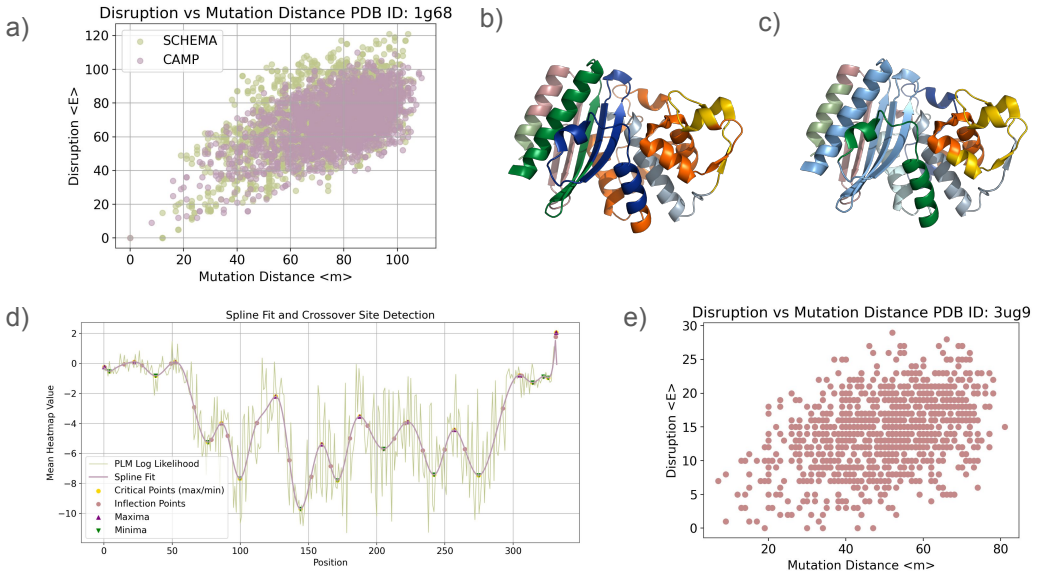

Figure 1: (a) Comparison of the disruption factor E and mutation distance m from the parent sequence using SCHEMA and our spline-based CAMP approach for chimeric sequences generated by chosing random block at each site. (b) Crossover sites determined by SCHEMA RASPP for PDB ID 1G68 Mateljak et al. (2019). (c) Crossover points determined using the spline method for the same protein. (d) Spline fit to the log-likelihood scores from ESM2 for channelrhodopsin (PDB ID: 3UG9). (e) Disruption and mutation distance trade-off for PDB 3UG9.

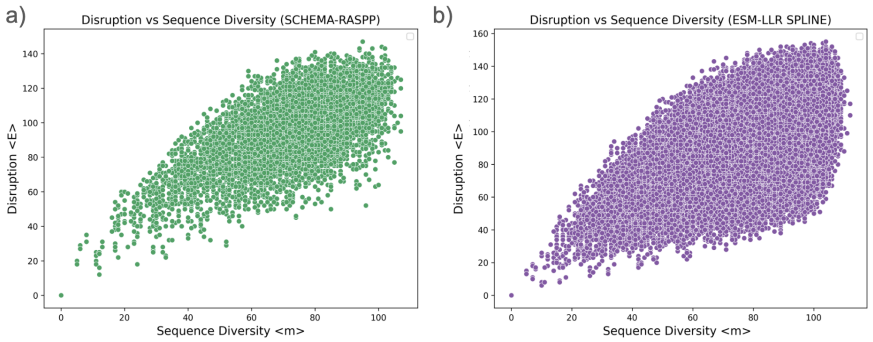

Figure 2: (a) Comparison of the disruption factor E and mutation distance from the parent sequence using SCHEMA. MCMC sampling was used to generate a diverse set of chimeras using the blocks generated from the SCHEMA-RASPP Mateljak et al. (2019) method with sequences selected from an MSA. (b) The same plot of disruption factor E and mutation distance, with the chimeric blocks selected from our log likelihood based method. this method produces more samples in the region of high sequence diversity and low disruption.

For each residue position j, we compute the average log-likelihood across all possible amino acids using the model's output distribution:

$$\ell_j = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \log p_j(a)$$

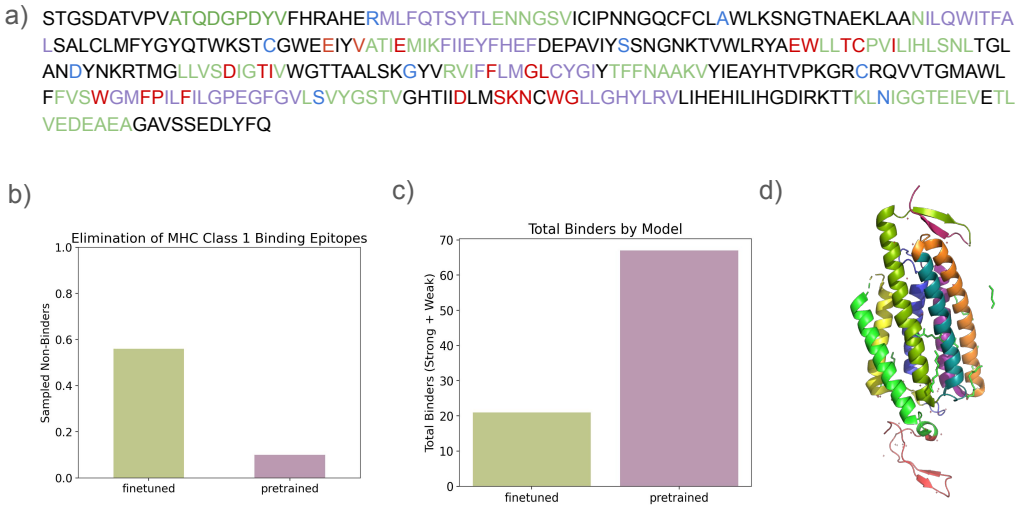

Figure 3: (a) Annotated sequence of channelrhodopsin C1C2 (PDB ID: 3UG9). Residues in red indicate positions identified as critical for opsin function. Crossover points are highlighted in blue. Residues in purple represent 9-mer peptides that strongly bind to the MHC Class I allele HLA-A02:01, as predicted by the percentile rank from netMHCPan. Weak binders are shown in green (b) Fraction of masked 9-mer MHC Class I epitopes that were sampled as non-binders (c) Number of total binders in the samples generated at the masked chimeric block regions of channelrhodopsin C1C2 (d) Channelrhodopsin structure (PDB ID: 3UG9)

where $\mathcal{A}$ is the amino acid alphabet (typically 20 standard amino acids), and $p_j(a)$ is the probability assigned by the model to amino acid a at position j. These log likelihood scores are defined at discreet points and are somewhat noisy, we fit a continouous spline to generate a smoothed importance function $S : [1, L] \to \mathbb{R}$. We then can choose crossover sites c that are local maxima of $S(j)$:

$$S'(j) = 0 \text{ and } S''(j) < 0$$

This gives us crossover sites that leverage evolutionary encoded information in an ESM2 model. We used these crossover sites to generate the chimeras in Figures 1 and 2.

3.1.1 MCMC SAMPLER FOR GENERATING CHIMERIC PROTEINS BY USING ESM2 AS AN ORACLE

After identifying the crossover sites, we applied a Markov Chain Monte Carlo (MCMC) approach to sample recombinant chimeric proteins by using MSATransformer. We utilized the MSATransformer to ensure that generated chimeric blocks are derived from the context of closely related evolutionary sequences in the MSA. Chimeric blocks were sampled from multiple sequence alignment (MSA) of the target protein, with a score calculated as the sum of the negative log probabilities for each token at position t, given the rest of the sequence (Rives et al. (2021)). We generated equal-sized libraries using the sampler for the crossover sites identified by (Mateljak et al. (2019)). As shown in Figure 2, this method achieves greater sequence diversity with a reduced disruption factor E (Mateljak et al. (2019)) by utilizing protein language models. ESM2 (Rives et al. (2021)) perplexity was used as the acceptance criterion during sampling.

Algorithm 1 MCMC Sampler For Chimeric Protein Library Generation

Data: T (number of steps), N (number of rounds), initial sequence set $\{\mathbf{S}_0^1, \mathbf{S}_0^2, \ldots, \mathbf{S}_0^N\}$
Result: Final set of sequences after MCMC sampling and clustering
for $n \leftarrow 1$ **to** N **do**
  for $t \leftarrow 1$ **to** T **do**
    At each timestep t, choose crossover site C_i, randomly select one subsection to recombine. Replace with the corresponding subsection randomly sampled from the MSA to generate new sequence set $\{\mathbf{S}_t^1, \mathbf{S}_t^2, \ldots, \mathbf{S}_t^N\}$
  end
  Cluster the sequences $\{\mathbf{S}_T^1, \mathbf{S}_T^2, \ldots, \mathbf{S}_T^N\}$ based on similarity Select a cluster leader from each cluster as the seed for the next round $n + 1$
end

3.1.2 Aligning Protein Language Models with Direct Preference Optimization

Recent advances in reinforcement learning from human feedback (RLHF) for large language models (LLMs) have demonstrated that reinforcement learning techniques like Direct Preference Optimization (DPO) can effectively align outputs to preference datasets, allowing models to better align with human preferences. This principle has recently been extended to protein language models (PLMs), as shown in the ProteinDPO studyWidatalla et al. (2024). ProteinDPO adapts DPO to align PLMs toward design goals such as stability by training on preference datasets for specific properties, like stability changes due to single or double mutations in protein sequences. Remarkably, ProteinDPO showed that by aligning on these targeted preference datasets, models can generalize effectively to broader mutational landscapes, even beyond the observed mutation space.

We created an immunogenicity preference dataset using the opsin sequence (PBD ID 3UG9). We used MSATransformer combined with netMHCPan-4.1 Hoof et al. (2009) to generate preferred and dispreferred samples. We used netMHCPan4.1 to identify 9-mers in the opsin sequence which are binders to MHC class 1 for allele HLA-A02:01. We then used the MSATransformer to mask and sample one of the 9mer regions known to bind to MHC Class I molecules. The generated sample was added to preferred or dispreferred dataset depending upon the predictions from netMHCPan-4.1. We then fine tuned the MSATransformer on this preference dataset using DPO to align the model to our dataset. This method is outlined in Algorithm 2.

Residues in channelrhodopsin sequence which are key residues were identified from previous studies, particularly Karl et al., and held constant during sampling. These residues included the pore residues (His134, His265, Glu82, Glu83, Asn258, Glu90, Glu97, Glu101, and Gln56) and the retinal binding pocket motifs (Ser155, Thr159, Gly181, Asp156, Cys128, Asp253, Glu123, Lys93, and Gly163). Flexible loop regions at the N-terminus (A49–A83) and C-terminus (A319–A342) were also fixed, given AlphaFold2's limitations in accurately predicting these regions.

Algorithm 2 DPO Fine-tuning for ESM-MSA-1b with Preferences Dataset

Data: Pretrained model π_θ, preferences dataset $\mathcal{D} = \{(x_i^+, x_i^-)\}_{i=1}^N$, learning rate η, number of iterations T, 20% held-out mask locations
Result: Fine-tuned ESM-MSA-1b model π_{θ^*}
for $t \leftarrow 1$ **to** T **do**
  Sample a batch of preference pairs $\{(x_i^+, x_i^-)\}$ from $\mathcal{D}$ **for** *each preference pair* (x_i^+, x_i^-) *in the batch* **do**
    Compute log probabilities: $\log \pi_\theta(x_i^+)$ and $\log \pi_\theta(x_i^-)$ Compute preference score: $\Delta_i = \log \pi_\theta(x_i^+) - \log \pi_\theta(x_i^-)$ Compute loss: $L_i = -\log \sigma(\Delta_i)$ σ is the sigmoid function
  end
  Compute batch loss: $L_{\text{batch}} = \frac{1}{|\text{batch}|} \sum_i L_i$ Update model parameters: $\theta \leftarrow \theta - \eta \nabla_\theta L_{\text{batch}}$
end

3.2 DISCUSSION

CAMP demonstrates a powerful integration of protein language models (PLMs) with combinatorial engineering to guide protein recombination and optimization. By identifying optimal crossover sites using log likelihood derived splines, we replicate and extend the logic behind SCHEMA without relying on structural data. The ability to generate diverse, functionally plausible chimeras shows that PLMs can effectively encode useful evolutionary and structural information. The MCMC sampling approach further enhances library diversity while maintaining low disruption, offering a scalable method for generating robust protein variants.

A key strength of CAMP is its use of Direct Preference Optimization (DPO) to reduce immunogenicity across multiple epitope sites, aligning PLMs with therapeutic design goals. Our approach provides several advantages over existing methods: (1) it identifies optimal recombination sites without requiring structural information, (2) it leverages the evolutionary information encoded in PLMs to guide recombination, and (3) it allows for targeted optimization of specific properties like immunogenicity through preference alignment.

In conclusion, CAMP serves a bridge between traditional recombination techniques and modern PLMs. As PLMs continue to advance, methods like CAMP that leverage their implicit understanding of protein structure and function will become increasingly valuable tools in the protein engineer's toolkit. Looking ahead, CAMP can be expanded to include other design constraints, like solubility or enzymatic activity, making it a flexible and efficient tool for protein engineering.

3.3 ACKNOWLEDGEMENTS

S.S. acknowledge the Schmidt Science Fellows for their generous support through the postdoctoral fellowship. S.S also thanks the Schmidt Future's Virtual Institute for Scientific Software (VISS program) and Johns Hopkins University's Scientific Software Engineering Center (JHU SSEC) for computational support and valuable discussions. M.P. and J.J. acknowledge Eleven Eleven Foundation and NecSys at MIT Media Lab for their support and compute resources. ESB acknowledges HHMI, Lisa Yang, NIH R01DA029639, NIH R01MH122971, NSF 1848029, and John Doerr.

REFERENCES

Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16(12):1315–1322, December 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0598-1. URL https://www.nature.com/articles/s41592-019-0598-1.

Ivan Anishchenko, Samuel J. Pellock, Tamuka M. Chidyausiku, Theresa A. Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K. Bera, Frank DiMaio, Lauren Carter, Cameron M. Chow, Gaetano T. Montelione, and David Baker. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, December 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-04184-w. URL https://www.nature.com/articles/s41586-021-04184-w.

Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. Improved prediction of protein-protein interactions using alphafold2. *Nature Communications*, 13(1):1265, March 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-28865-w. URL https://www.nature.com/articles/s41467-022-28865-w.

Fengyuan Dai, Yuliang Fan, Jin Su, Chentong Wang, Chenchen Han, Xibin Zhou, Jianming Liu, Hui Qian, Shunzhi Wang, and Anping Zeng. Toward de novo protein design from natural language. *bioRxiv*, pp. 2024–08, 2024. URL https://www.biorxiv.org/content/10.1101/2024.08.01.606258.abstract.

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. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.add2187. URL https://www.science.org/doi/10.1126/science.add2187.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model, July 2024. URL https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1.

David G. Hernández-Mejía, Iván Aurelio Páez-Gutiérrez, Valerie Dorsant Ardón, Nathalie Camacho Ramírez, Melissa Mosquera, Paola Andrea Cendales, and Bernardo Armando Camacho. Distributions of the hla-a, hla-b, hla-c, hla-drb1, and hla-dqb1 alleles and haplotype frequencies of 1763 stem cell donors in the colombian bone marrow registry typed by next-generation sequencing. *Frontiers in Immunology*, 13:1057657, January 2023. ISSN 1664-3224. doi: 10.3389/fimmu.2022.1057657. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9869256/.

Ilka Hoof, Bjoern Peters, John Sidney, Lasse Eggers Pedersen, Alessandro Sette, Ole Lund, Søren Buus, and Morten Nielsen. Netmhcpan, a method for mhc class i binding prediction beyond humans. *Immunogenetics*, 61(1):1–13, January 2009. ISSN 1432-1211. doi: 10.1007/s00251-008-0341-z. URL https://doi.org/10.1007/s00251-008-0341-z.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, and Anna Potapenko. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021a. URL https://www.nature.com/articles/s41586-021-03819-2).

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, 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 2021b. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2.

Ivan Mateljak, Austin Rice, Kevin Yang, Thierry Tron, and Miguel Alcalde. The generation of thermostable fungal laccase chimeras by schema-raspp structure-guided recombination in vivo. *ACS Synthetic Biology*, 8(4):833–843, April 2019. ISSN 2161-5063, 2161-5063. doi: 10.1021/acssynbio.8b00509. URL https://pubs.acs.org/doi/10.1021/acssynbio.8b00509.

Jacques Neefjes, Marjolein L. Jongsma, P. Paul, and O. Bakke. Towards a systems understanding of MHC class i and MHC class ii antigen presentation. *Nature Reviews Immunology*, 11(12): 823–836, 2011. doi: 10.1038/nri3084.

Martin Olivecrona, T Blaschke, O Engkvist, and H Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48, 2017.

Roshan M. Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8844–8856. PMLR, July 2021. URL https://proceedings.mlr.press/v139/rao21a.html.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, April 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2016239118. URL https://pnas.org/doi/full/10.1073/pnas.2016239118.

Jeffrey A. Ruffolo, Stephen Nayfach, Joseph Gallagher, Aadyot Bhatnagar, Joel Beazer, Riffat Hussain, Jordan Russ, Jennifer Yip, Emily Hill, Martin Pacesa, Alexander J. Meeske, Peter Cameron, and Ali Madani. Design of highly functional genome editors by modeling the universe of CRISPR-Cas sequences, April 2024. URL https://www.biorxiv.org/content/10.1101/2024.04.22.590591v1.

Ben Sánchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *ACS Central Science*, 4(2):268–276, 2018.

Talal Widatalla, Rafael Rafailov, and Brian Hie. Aligning protein generative models with experimental fitness via direct preference optimization. *bioRxiv*, 2024. doi: 10.1101/2024.05.20.595026. URL https://www.biorxiv.org/content/early/2024/05/21/2024.05.20.595026.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, December 2004. ISSN 0887-3585, 1097-0134. doi: 10.1002/prot.20264. URL https://onlinelibrary.wiley.com/doi/10.1002/prot.20264.

A Appendix

A.1 Choice of the Alleles

Since experimental data on immunogenicity is difficult to obtain, we aim to reduce epitope recognition for common alleles, using these as a proxy in in silico models to estimate immunogenic potential and guide our design process. Common distribution of alleles was obtained from Hernández-Mejía et al. (2023)

A.2 Application to other proteins

We generated a multiple sequence alignment (MSA) for esmGFP Dai et al. (2024) and applied our method to propose new sequences, using only the available sequences similar to esmGFP, to assess whether we could successfully recapitulate its sequence.

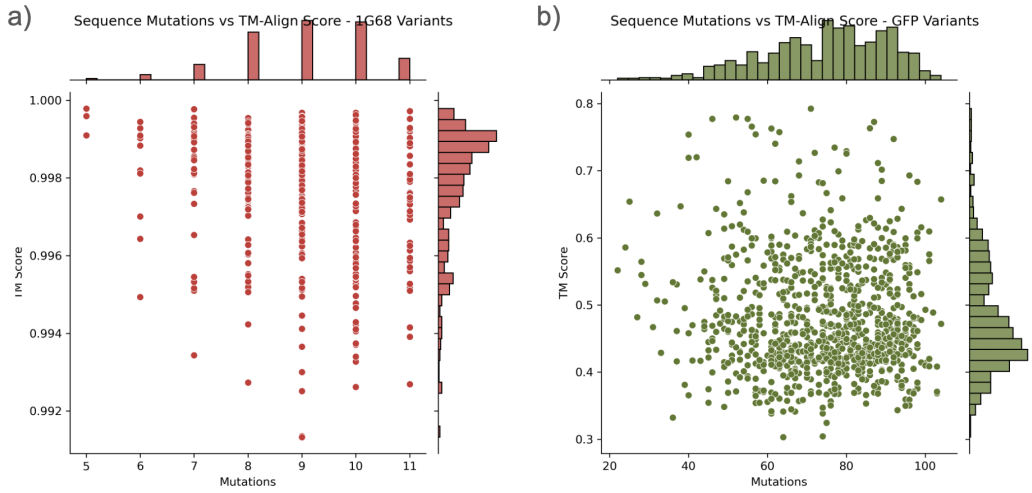

Figure 4: (a) Beta Lactamase sequences aligned to the original strucutre after strucutre prediction (b) Sequences generated from esmGFP's MSA using our recombination strategy