Design and Evolution of Neuron-Specific Proteases

Han B. Spinner
Program in Biological and Biomedical Science
Harvard Medical School
hspinner@g.harvard.edu

Colin Hemez
Program in Biophysics
Harvard University
colinhemez@g.harvard.edu

Julia McCreary
Program in Chemical Biology
Harvard University
jmccreary@g.harvard.edu

David R. Liu
Merkin Institute of Transformative Technologies in Healthcare
Broad Institute of MIT and Harvard
drliu@fas.harvard.edu

Debora S. Marks
Department of Systems Biology
Harvard Medical School
debbie@hms.harvard.edu

Abstract

Directed evolution has remarkably advanced protein engineering. However, these experiments are typically seeded with a single sequence, and they are limited by the amount of sequence space they can explore. Here, we aim to develop a machine learning method that learns from the natural distribution of sequences to design diverse seed sequences. We use Botulinum Neurotoxin X (BoNT/X) as a proof of concept for this approach since there is published data on this evolution campaign, and there are many therapeutic applications of neuron-specific proteases. Additionally, BoNT/X is especially promising for this approach since related BoNT proteases have specific substrate specificity, limiting the utility of simply drawing from the natural sequences. We hypothesize that our machine learning model can learn the 'essence' of the protein family and generate diverse substrate binding domains. We built an alignment of 452 sequences around BoNT/X and show that models trained on this data can separate known beneficial and deleterious mutations. Next, we will use these models to generate sequences and perform new evolution experiments. Finally, we will evaluate the impact of starting with a diverse set of seed sequences versus only one seed sequence. This work will not only create new proteases that can be used for therapeutic indications, but also puts forth a new approach for machine-learning-guided evolution experiments.

1 Introduction

Designing novel proteins for any desired function is a long-standing goal of bioengineering. Towards this end, there have been many advances in both high-throughput assays as well as computational methods, and there is fruitful grounds for integrating these two research areas.

There are very few examples of using unsupervised machine learning to accelerate protein engineering. For instance, training unsupervised models on natural sequences has created chorismate mutase enzymes with enhanced natural functions [1], has deimmunized protein biotherapeutics [2], and has generated nanobody libraries [3]. And on the other hand, most wetlab experiments do not include machine learning at all. To build sequences with a desired phenotype, biologists have developed

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

numerous high-throughput laboratory assays [4] spanning the gamut from classic directed evolution [5, 6] to continuous evolution platforms such as Phage Assisted Continuous Evolution (PACE)[7] and OrthoRep [8]. Some of the most challenging engineering objectives involve changing the enzyme specificity or functional conditions, such as protease cleavage specificity [9] or increased thermotolerance for plastic-eating enzymes [10, 11]. However, these experiments are labor-intensive and as such can only access a limited amount of sequence space. No method exists that can learn from natural sequence data to design custom, functionally diverse proteins.

Our innovation: Model-guided protein evolution
Here, we aim to learn the distribution of sequences that exist in nature to generate new highly developable sequences with which to begin continuous evolution experiments. In this way, we will increase the starting diversity of these experiments and traverse a greater sequence landscape.

Using BoNT/X as a proof of principle
As a proof of principle for this goal, we will diversify the substrate specificity of Botulinum Neurotoxin X (BoNT/X) with PACE. PACE is a type of directed evolution that connects phage survival to a desired phenotype in a protein of interest. Starting with a single seed sequence, hundreds of generations of phage infecting bacteria can be used to continuously evolve the protein of interest. BoNT proteases have advantageous biology because their heavy chain naturally trafficks them to neurons, and they have already shown impact in the clinic as a therapeutic medicine[12]. BoNT/X is particularly interesting because of its natural substrate flexibility and initial work proving it can be evolved to target novel substrates [9]. This research follows previously published work where BoNT/X was evolved to cleave two novel substrates in a step-wise manner that required 20+ passages [9].

Can we use machine learning to design and evolve neuron-specific proteases against any target? And, can these models lead us into sequence space that is impossible to access through PACE alone?

Seeding PACE with sequences drawn from models trained on natural diversity may overcome several of the system's inherent limitations. Firstly, these experiments are very time intensive: they can take 6 months to 1 year from identifying the target to having a validated molecule of interest. Secondly, each evolution campaign only explores a few sequential steps in sequence space around a starting sequence, often resulting in (a) sequences being trapped in local minima and (b) not fully exploring epistatic interactions with multiple mutations co-occurring. Thirdly, the task of changing substrate specificity poses some unique additional challenges. Because changing the enzyme's specificity is quite difficult, there are intermediate 'stepping stone' substrates that the protease gets evolved to cleave. In this way, rather than starting with WT and evolving against a new substrate directly, the WT must go through several sequential cycles of PACE each with increasingly different substrates [9]. All of these limits together create the perfect proof of principle application of integrating ML into continuous evolution workflows. By using a range of designed sequences as starting points, we increase the chances that they will avoid local minima traps and that sequences will accumulate more potentially epistatic mutations. Theoretical work on sequence evolution demonstrates that less fit amino acid substitutions can break sequences out of local minima to force more exploration [13]. We hypothesize the diverse starting library will lead to an expedited evolution and obviate the need for intermediate substrates.

Using deep evolutionary-informed ML models to design a starting library for continuous evolution experiments, we will generate sequences from the natural distribution in order to change enzyme specificity. The natural distribution contains proteases that can cleave a range of targets and the model will subsequently learn the 'essence' of this protein family, agnostic to the specificity. Therefore, we will use machine learning to create a highly evolvable starting set of sequences. We hypothesize that more initial diversity will lead to more fit enzymes and that PACE will hone this library into new candidate proteases.

2 Methods

Our approach is to build a multiple sequence alignment seeded with the BoNT/X light chain wild-type sequence (UniProt ID: P0DPK1), train and validate several state-of-the-art machine learning models for sequence data, use the best model to generate new sequences as launching points for the PACE campaigns, and run PACE campaigns on a variety of targets.

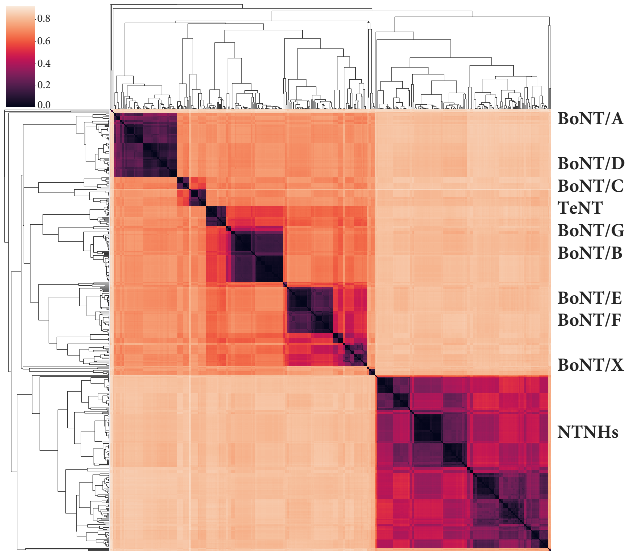

Figure 1: **Heatmap of sequences in alignment.** Orange indicates more diverse, less similar sequences and purple indicates less diverse, more similar sequences.

2.1 Align

We built an evolutionary alignment from querying both the UniRef100 [14] and MGnify [15] databases and compiled them into a single file and aligned the sequences using ClustalOmega [16]. In total, we gathered 452 sequences that included all other BoNT serotypes A-G as well as corresponding nontoxic-nonhemagglutinin (NTNH) anti-toxin proteins (Figure 1).

2.2 Train

Using the compiled alignment, we trained EVmutation [17], and EVE [18]. EV Mutation is a very interpretable statistical model that can capture residue dependencies for all pairwise interactions across the length of the protein. EVE is a deep generative variational autoencoder that learns the underlying distribution of sequences in order to capture complex, higher-order dependencies. Tranception [19], on the other hand, does not rely on an alignment, and it is a pre-trained autoregressive transformer trained on the whole protein universe that has learned general principles of the protein language.

2.3 Validate

To validate each models' performance, we predicted the fitness of 12 known advantageous mutations from previously published work and 5 known deleterious mutations (H227, E228, H231, E266, C423) [20] to either alanine or serine. While most of the beneficial mutations were observed in the context of multi-mutation variants in experiments, we treated the impact of each mutation as independent to avoid known issues with depth-dependency of fitness predictions in unsupervised sequence models.

3 Results

To see if our fitness predictions match what is known, we compared our set of variant predictions with the full distributions of predicted fitness scores for all single amino acid substitutions (Figure 2). Both EVmutation and Tranception show a modest separation between predictions for beneficial and harmful mutations while EVE shows a larger separation. We can see that some lab-validated mutations are predicted to be very deleterious when scored as single mutants. As previously stated, in the experiments, these mutations do not exist on their own. Most of these advantageous mutations actually occur with other mutations present as well, and it is likely that some of these mutations work in an epistatic manner that cannot be observed when scoring them as single mutations. We evaluated significance with Welch's t-test [21] since the variance of beneficial and deleterious mutations are

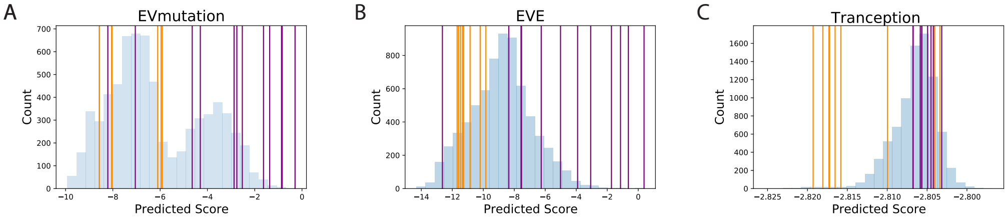

Figure 2: **Model Predictions.** For all panels, light blue is the distribution of model predictions for all possible single amino acid substitutions. Purple lines indicate previously published advantageous mutations and orange lines indicate deleterious mutations. (A) EVmutation predictions. (B) EVE predictions. (C) Tranception predictions.

Table 1: Summary of model performances

	Mean predicted score			
Model	Advantageous	Deleterious	Difference	P-value
EVmutation	-3.1125	-7.4051	-4.2927	0.0000895
EVE	-4.7903	-11.0292	-6.2390	0.0001346
Tranception	-2.8051	-2.8126	-0.007512	0.0052515

not equal, but each roughly follow a normal distribution. All three differences appear significant. Currently, EVE performs best with the experimental data and we can see a mostly clear separation of beneficial and harmful mutations (Figure 2b).

4 Future work

As we add more validation data, either in the form of sequences with full sets of mutations or new single mutants that are shown to work in the lab, we will continue evaluating model performance. Then, we will generate new sequences using the best-performing model. If EVE continues to perform the best, we can sample sequences from the latent space learned by the model [22], with some biological constraints like not mutation the parts of BoNT/X that interact with the heavy chain. We will verify biophysical properties of those sequences to ensure that qualities like hydrophobicity and isoelectric point are indistinguishable from the natural distribution as is done in other protein design work[3]. Because this protein family and the subsequent alignment is so diverse, we expect that the generated sequences will form a library of highly diverse sequences. We will aim to generate sequences that are 10-50 amino acids away from wild-type and that have a high mutation rate in the substrate-binding residues and in areas previously known to have an impact in substrate specificity. And importantly, we will enforce that our model does not generate mutations in the sites necessary to bind the neuron-trafficking heavy domain. These properties will be computationally verified before experiments begin.

We will design and build 100 novel proteins in the lab. We will run PACE campaigns on each sequence against five novel protein substrates. In this way, each starting sequence will evolve in five different directions resulting in 500 PACE experiments. Additionally, we will explore different experimental approaches and pool all 100 sequences against each substrate to create a library-seeded PACE experiment. And as a control, we will also evolve the wild-type BoNT/X against each of the 5 novel substrates. We will take samples at various time points from each of these evolution experiments and sequence the proteins as they evolve. Each evolved protease will be evaluated, as previously described[9], for cleavage activity on the novel substrate as well as the canonical wild-type substrate. These new proteases will be successful if they are able to cleave the new substrates and no longer cleave the original substrate. Additionally, total diversity and number of unique mutations will be evaluated in each of these PACE experiments to see if the starting library impacts the evolutionary trajectory of the populations.

5 Discussion

The research described in this paper implements machine learning into a new biological field and we hypothesize it will carve a faster path to designing novel enzymes. By starting PACE at many diverse seeds, we hope to capture multi-mutational interactions that the experiment would otherwise miss. Not only will this research create novel proteases that will have impact in therapeutics and biotechnology, but this new method also demonstrates the power of designing machine-learning-guided starting points for evolution experiments.

Acknowledgments and Disclosure of Funding

We thank members of the Marks laboratory and Liu Group for many great discussions. Nicole N. Thadani and Noor Youssef made valuable comments on the manuscript. Han B. Spinner is supported by the Fujifilm Fellowship. David R. Liu is funded by US National Institutes of Health (NIH) grants R01EB027793, R01EB031172, R35GM118062, and the Howard Hughes Medical Institute. Debora S. Marks holds a Ben Barres Early Career Award by the Chan Zuckerberg Initiative as part of the Neurodegeneration Challenge Network, CZI2018-191853, as well as NIH TR01 grant (R01CA260415).

References

[1] William P. Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, and Rama Ranganathan. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369, 2020. ISSN 10959203. doi: 10.1126/science.aba3304.

[2] Benjamin Schubert, Charlotta Schärfe, Pierre Dönnes, Thomas Hopf, Debora Marks, and Oliver Kohlbacher. Population-specific design of de-immunized protein biotherapeutics. *PLoS Computational Biology*, 14, 2018. ISSN 15537358. doi: 10.1371/journal.pcbi.1005983.

[3] Jung Eun Shin, Adam J. Riesselman, Aaron W. Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C. Kruse, and Debora S. Marks. Protein design and variant prediction using autoregressive generative models. *Nature Communications*, 12, 2021. ISSN 20411723. doi: 10.1038/s41467-021-22732-w.

[4] Michael S. Packer and David R. Liu. Methods for the directed evolution of proteins, 2015. ISSN 14710064.

[5] K. Chen and F. H. Arnold. Tuning the activity of an enzyme for unusual environments: Sequential random mutagenesis of subtilisin e for catalysis in dimethylformamide. *Proceedings of the National Academy of Sciences of the United States of America*, 90, 1993. ISSN 00278424. doi: 10.1073/pnas.90.12.5618.

[6] Frances H. Arnold. Design by directed evolution. *Accounts of Chemical Research*, 31, 1998. ISSN 00014842. doi: 10.1021/ar960017f.

[7] Kevin M. Esvelt, Jacob C. Carlson, and David R. Liu. A system for the continuous directed evolution of biomolecules. *Nature*, 472, 2011. ISSN 00280836. doi: 10.1038/nature09929.

[8] Gordon Rix, Ella J. Watkins-Dulaney, Patrick J. Almhjell, Christina E. Boville, Frances H. Arnold, and Chang C. Liu. Scalable continuous evolution for the generation of diverse enzyme variants encompassing promiscuous activities. *Nature Communications*, 11, 2020. ISSN 20411723. doi: 10.1038/s41467-020-19539-6.

[9] Travis R Blum, Hao Liu, Michael S Packer, Xiaozhe Xiong, Pyung-Gang Lee, Sicai Zhang, Michelle Richter, George Minasov, Karla J F Satchell, Min Dong, and David R Liu. Phage-assisted evolution of botulinum neurotoxin proteases with reprogrammed specificity, 2021. URL `https://www.science.org`.

[10] Elizabeth L. Bell, Ross Smithson, Siobhan Kilbride, Jake Foster, Florence J. Hardy, Saranarayanan Ramachandran, Aleksander A. Tedstone, Sarah J. Haigh, Arthur A. Garforth, Philip J.R. Day, Colin Levy, Michael P. Shaver, and Anthony P. Green. Directed evolution of an efficient and thermostable pet depolymerase. *Nature Catalysis*, 5, 2022. ISSN 25201158. doi: 10.1038/s41929-022-00821-3.

[11] Stefan Brott, Lara Pfaff, Josephine Schuricht, Jan Niklas Schwarz, Dominique Böttcher, Christoffel P.S. Badenhorst, Ren Wei, and Uwe T. Bornscheuer. Engineering and evaluation of thermostable ispetase variants for pet degradation. *Engineering in Life Sciences*, 22, 2022. ISSN 16182863. doi: 10.1002/elsc.202100105.

[12] Marco Pirazzini, Ornella Rossetto, Roberto Eleopra, and Cesare Montecucco. Botulinum neurotoxins: Biology, pharmacology, and toxicology, 4 2017. ISSN 15210081.

[13] Noor Youssef, Edward Susko, Andrew J. Roger, and Joseph P. Bielawski. Evolution of amino acid propensities under stability-mediated epistasis. *Molecular Biology and Evolution*, 39, 2022. ISSN 15371719. doi: 10.1093/molbev/msac030.

[14] Baris E. Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H. Wu. Uniref: Comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23, 2007. ISSN 13674803. doi: 10.1093/bioinformatics/btm098.

[15] Alex L. Mitchell, Alexandre Almeida, Martin Beracochea, Miguel Boland, Josephine Burgin, Guy Cochrane, Michael R. Crusoe, Varsha Kale, Simon C. Potter, Lorna J. Richardson, Ekaterina Sakharova, Maxim Scheremetjew, Anton Korobeynikov, Alex Shlemov, Olga Kunyavskaya, Alla Lapidus, and Robert D. Finn. Mgnify: The microbiome analysis resource in 2020. *Nucleic Acids Research*, 48, 2020. ISSN 13624962. doi: 10.1093/nar/gkz1035.

[16] Fabian Sievers, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D. Thompson, and Desmond G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, 7, 2011. ISSN 17444292. doi: 10.1038/msb.2011.75.

[17] Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P.I. Schärfe, Michael Springer, Chris Sander, and Debora S. Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35:128–135, 2 2017. ISSN 15461696. doi: 10.1038/nbt.3769.

[18] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K. Min, Kelly Brock, Yarin Gal, and Debora S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599:91–95, 11 2021. ISSN 14764687. doi: 10.1038/s41586-021-04043-8.

[19] Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan Gomez, Debora S. Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. 5 2022. URL `http://arxiv.org/abs/2205.13760`.

[20] Geoffrey Masuyer, Sicai Zhang, Sulyman Barkho, Yi Shen, Linda Henriksson, Sara Košenina, Min Dong, and Pål Stenmark. Structural characterisation of the catalytic domain of botulinum neurotoxin x - high activity and unique substrate specificity. *Scientific Reports*, 8, 12 2018. ISSN 20452322. doi: 10.1038/s41598-018-22842-4.

[21] John H. McDonald. *Handbook of Biological Statistics, 3th edition*. 2014.

[22] Alex Hawkins-Hooker, Florence Depardieu, Sebastien Baur, Guillaume Couairon, Arthur Chen, and David Bikard. Generating functional protein variants with variational autoencoders. *PLoS Computational Biology*, 17, 2021. ISSN 15537358. doi: 10.1371/JOURNAL.PCBI.1008736.