The Tournament

Submitted Abstracts from the 2023 pilot Protein Engineering Tournament.

Thank you to all of our participating teams! Teams were requested, but not required, to submit abstracts on their methods in the *in silico* and *in vitro* rounds.

ALIGN TO INNOVATE

In silico evolution of an a-amylase using multiobjective optimisation

Sergi Roda[1,2], Martin Floor[1*]

[1]*Department of Life Sciences, Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain*

[2]*Nostrum Biodiscovery, 08034, Barcelona, Spain*

*Corresponding author: mfloor@bsc.es

Key words: *in silico* enzyme engineering, molecular modeling, genetic algorithm

The EAPM team performed various ranges of predictions for the zero-shot round and contributed with a list of mutants for the *in vitro* round by utilizing an in-house implementation of a genetic algorithm for multi objective optimisation.

Methods

In silico prediction of aminotransferase activity against a set of substrates using DLKcat

To estimate the enzyme activity of the aminotransferase variants against the three reported substrates (S-Phenylethylamine, (4-Chlorophenyl)phenylmethanamine, and 1,3-Diphenyl-propane-1-amine), we used a deep-learning approach that can predict the turnover number of an enzyme against a specific substrate[1] (https://github.com/SysBioChalmers/DLKcat). The model uses an enzyme sequence and the substrate smiles as input to predict the turnover rate of the enzyme to catalyze the given substrate. In the article on this model, they claimed that it could also be used in enzyme mutants, so we wanted to test if this fast tool can be used in an independent benchmark.

In silico prediction of expression of a set of xylanases using a consensus of machine learning predictors

We used both SoluProt[2] (https://loschmidt.chemi.muni.cz/soluprot/) and MPEPE[3] (https://github.com/BRITian/MPEPE) to estimate whether the enzyme will be well expressed (1), badly expressed (0.5) or not even expressed (0). Both models give a value between 0 (not expressed) and 1 (expressed) to classify expressable and non-expressable proteins. SoluProt uses the protein sequence, but MPEPE uses the coding sequence. Since multiple coding sequences can translate a protein sequence, we back-translated the protein sequences in three replicas using the frequency of codon usage in *Escherichia coli* and then predicted the expression with MPEPE. Subsequently, we performed a weighted mean of predictions from both software. If the value was between 0 and ⅓, the sequence was labeled as 0, 0.5 if the value was higher than ⅓ and equal to or smaller than ⅔, and 1 if the value was higher than ⅔.

Design of a list of multiple mutants to enhance the activity of an α-amylase using a genetic algorithm

We used a structure-based approach to develop evolved variants of the studied α-amylase. First, we took the crystal structure with PDB ID 1BAG and reconstructed the WT enzyme by adding the seven mutations that separate both sequences. We used the experimental set of single mutants that enhanced the enzyme activity without compromising its stability and expressibility to construct a library of mutations for a multiobjective design approach. The multiobjective design approach is a genetic algorithm that iteratively creates sets of variants (either by mutation or by recombination of the list of variants from the previous iteration) to then evaluate their fitness through a set of metrics obtained by short calculations (here Rosetta optimizations) and selects those variants enhancing all the objectives of interest simultaneously (Pareto front). We employed the substrate binding energy and the system's overall total energy as explicit objectives for this optimization. Then, we chose those mutants with the best pareto ranks but with a maximum of 9 shared mutations among them to diversify the mutants represented while selecting the most improved variants.

Availability

The used code is still under development, but we can provide it to the public once experiments have provided validation.

References

1. Li, F. *et al.* Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nat. Catal.* **5**, 662–672 (2022).

2. Hon, J. *et al.* SoluProt: prediction of soluble protein expression in Escherichia coli. *Bioinformatics* **37**, 23–28 (2021).

3. Ding, Z. *et al.* MPEPE, a predictive approach to improve protein expression in based on deep learning. *Comput. Struct. Biotechnol. J.* **20**, 1142–1153 (2022).

Generation of artificial alpha amylases with ZymCTRL

Geraldene Munsamy[1], Phil Lorenz[1*], Noelia Ferruz[2†]

[1]*Basecamp Research Ltd., London, United Kingdom*

[2]*Institute of Molecular Biology of Barcelona, IBMB-CSIC, Barcelona, Spain*

*,†Corresponding authors: *phil@basecamp-research.com; †noelia.ferruz@ibmb.csic.es

Methods

We fine-tuned ZymCTRL[1]in the set of sequences with values equal to or higher than 0.9 in expression, 0.9 in stability temperature, and 1 in activity, leading to around 772 sequences. The model was fine-tuned with a learning rate of 0.8e-06 for 300 epochs. After training, we generated 21834 sequences. We sought to find any correlation between the sequences in the experimental dataset and ZymCTRL perplexity values, we however did not find any correlation with activity, thermostability, or expression rates. We noticed, however, that the most active, stable, and well-expressed sequences tended to have perplexity values below 1.069. Following this, we selected generated sequences with values below that threshold, leading to 6799 sequences. 379 of these sequences were present in the experimental dataset and were removed, along with 117 sequences that did not have the canonical length of 425 amino acids. The final generated dataset contained 6303 sequences. We computed several metrics on these sequences, in particular, ESM-1v values[2], ProteinMPNN log-likelihoods[3], AlphaFold plDDTs[4], and Rosetta SAP score[5]. Due to most sequences being single-mutants of the wild type, we did not observe a great variability for these metrics, with the exception of dSAP values. We then ranked the sequences by their dSAP value and selected the top 200 for submission.

Availability

The model is freely available at: https://huggingface.co/AI4PD/ZymCTRL

References

1. Munsamy, G., Lindner, S., Lorenz, P. & Ferruz, N. ZymCTRL: a conditional language model for the controllable generation of artificial enzymes. in *Machine Learning for Structural Biology Workshop. NeurIPS 2022*.

2. Meier, J. *et al. Language models enable zero-shot prediction of the effects of mutations on protein function*. http://biorxiv.org/lookup/doi/10.1101/2021.07.09.450648 (2021) doi:10.1101/2021.07.09.450648.

3. Dauparas, J. *et al.* Robust deep learning based protein sequence design using ProteinMPNN. Preprint at https://doi.org/10.1101/2022.06.03.494563 (2022).

4. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

5. Verkuil, R. *et al. Language models generalize beyond natural proteins*. http://biorxiv.org/lookup/doi/10.1101/2022.12.21.521521 (2022) doi:10.1101/2022.12.21.521521.

Enhancing alpha-amylase activity using natural sequence models with bias by experimental data and biophysics hypotheses.

Hansen Spinner[1], Steffanie Paul[1], Nathan Rollins[2], Chris Sander[3,4], Debora Marks[1,4]

[1]Harvard Medical School - Department of Systems Biology ; [2]Seismic Therapeutic ; [3]Dana Farber Cancer Research Institute ; [4]Broad Institute of Harvard and MIT

Alpha-amylases are valuable enzymes used to produce food, biofuel, paper coating, and applied in detergents, desizing, and destaling agents. Each application has unique pH, temperature, and solvent conditions. For this competition, we aimed to design alpha-amylases that express in B subtilis and have high ability to liquefy cornstarch at mild temperature (50C) and low pH (4.5/5/5.5). We took two approaches, both of which build upon the capability of natural sequence models to predict [1] and enhance evolutionary properties of proteins [2,3].

In approach #1, we made use of machine learning models trained on natural sequences, and AlignBio's deep mutational scans of B. subtilis alpha amylase. We trained a EVcouplings model on 1000s of natural amylase sequences and compared to mutation effects on expression, activity, and activity after temperature stress. Mutations ranked top by the model had fully native-like properties. However, the correlations are modest (spearman r = 0.48,0.37,0.35 respectively) and mutations that enhance activity in the assay conditions were often not high scoring under the model. We designed sequences by maximizing EVcouplings probability via simulated annealing, with and without biasing by the assay data. We tested two bias strategies: a 'mask' to avoid substitutions deleterious in the assay, and a 'boost' for substitutions better than wildtype in the assay.

In approach #2, we selected 5 alpha-amylases from prior engineering efforts in literature. Each of these enzymes perform optimally at pH/T conditions unique from those of the competition. We employ EVcouplings to optimize the sequences, with and without biasing the model towards mutations with desirable biophysical properties. Namely: eliminating buried residues that adopt charge at target pH, optimizing the charge environment at the active site for target pH, and increasing the flexibility of 'lid regions' to encourage low temperature activity.

We eagerly await the results!

References

1. Hopf et al. Mutation effects predicted from sequence co-variation. (2017)

2. Russ et al. An evolution-based model for designing chorismate mutase enzymes. (2020)

3. Fram et al. Simultaneous enhancement of multiple functional properties using evolution-informed protein design. (2023)

Enzyme design using protein language models

Robert Schmirler [1, 2, 3 *], Michael Heinzinger [2, 3] & Burkhard Rost [2, 4, 5]

[1]*Innovation Center, BTS IR LU, AbbVie Deutschland GmbH \& Co. KG, Ludwigshafen, Germany;* [2]*Faculty of Informatics, TUM (Technical University of Munich), Munich, Germany;* [3]*TUM School of Computation, Information and Technology (CIT), TUM Graduate School, Garching, Germany;* [4]*Institute for Advanced Study (TUM-IAS), TUM, Garching, Germany;* [5]*TUM School of Life Sciences Weihenstephan (WZW), TUM, Freising, Germany*

*Corresponding author: schmirler@rostlab.org

Abstract

In computational biology, large language models have learned to extract meaningful features from raw, unlabeled sequence data. The values of the last hidden layers from these pre-trained protein language models (pLMs) [1]–[4], dubbed the embeddings, are used as exclusive input to subsequent state-of-the-art (SOTA) prediction methods which tend to reach or outperform methods using evolutionary information, or hand-crafted features. Embedding-based predictions seem particularly successful when experimental data are very limited [5]. We use pLMs for predictions in the in-silico round and employ them for selection of optimized variants in the in-vitro round.

In silico round

We use ProtT5 [1], a encoder decoder pLM which is based on google's T5 architecture. Instead of using frozen pretrained embeddings we finetune the model. This means we modify the models embedding to better capture information that is needed for the specific prediction task at hand. For this we take the model encoder and put a simple regression prediction head on top. Due to the large model size (1.2B parameters in the encoder) we use LoRA [6] a form of parameter efficient finetuning, which makes training way less resource demanding.
We normalize the property values for Alkaline phosphatase (to standard deviation of 1 and mean of 0). We use the values for Alpha-amylase as is but removed all dataset 1 rows for training. We create random train (95%) and validation splits (5%). We train separate predictors for each property and protein. We then use these property specific predictors to make predictions on the test set.

In vitro round

We again train three property specific fine tuned ProtT5 models for Alpha-Amylase (dataset 2). These oracles are used to select the most promising variants. To get candidates for the oracles to evaluate, we use two methods. First we use the pLM based efficient evolution [7] deploying both ESM models and ProtT5 to further optimize the best sequence from the training set (L001A-G009M N043M-Q068H-I100L-S142L-V226I-V254I-P424A). Efficient evolution gives us 27 possibly optimizing point mutations to introduce. We evaluate combinations (up to four) of those with our oracle and select the best outcomes. As an alternative approach we use EvoPlay [8] a reinforcement learning agent. We seed it on different sequences (WT, best sequence mentioned before, some intermediate sequences). All variants suggested by EvoPlay are also evaluated by the pLM oracles and best variants are selected.

References

1. A. Elnaggar *et al.*, “ProtTrans: Towards Cracking the Language of Life’s Code Through Self Supervised Learning,” *bioRxiv*, p.2020.07.12.199554, Jan.2021, doi:10.1101/2020.07.12.199554.

2. A. Elnaggar *et al.*, “Ankh ☥: Optimized protein language model unlocks general-purpose modelling,” *bioRxiv*, pp. 2023–01, 2023.

3. Z. Lin *et al.*, “Evolutionary-scale prediction of atomic-level protein structure with a language 8 model,” *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

4. B. Chen *et al.*, “xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein,” *bioRxiv*, pp. 2023–07, 2023.

5. M. Littmann, M. Heinzinger, C. Dallago, K. Weissenow, and B. Rost, “Protein embeddings and deep learning predict binding residues for various ligand classes,” *Sci. Rep.*, vol. 11, no. 1, p. 23916, 2021.

6. E. J. Hu *et al.*, “Lora: Low-rank adaptation of large language models,” *ArXiv Prepr. ArXiv210609685*, 2021.

7. B. L. Hie *et al.*, “Efficient evolution of human antibodies from general protein language models,” *Nat. Biotechnol.*, pp. 1–9, Apr. 2023, doi: 10.1038/s41587-023-01763-2. 18 [8] Y. Wang *et al.*, “Self-play reinforcement learning guides protein engineering,” *Nat. Mach. Intell.*, vol. 5, no. 8, pp. 845–860, Aug. 2023, doi: 10.1038/s42256-023-00691-9.

Computational Protein Engineering of Alpha-Amylase Variants

Team: Nimbus
Jason C. Klima, PhD[1]
[1]*Encodia, Inc., San Diego, CA, 92131*

In silico Round

Zero-shot Round

In the zero-shot Aminotransferase, Alpha-amylase, and Xylanase events, macromolecular models of the pre-transition state of each enzyme and substrate(s) were prepared and scored in PyRosetta[1] after a literature review of relevant mechanisms. Substrates were parameterized in PyRosetta, and enzyme backbone and sidechain coordinates were predicted using ColabFold-AlphaFold2[2,3] or ESMFold[4], or downloaded from the AlphaFold Protein Structure Database[3,5] with UniProt identification numbers. DiffDock[6] was used to dock substrates into enzymes, then the highest ranking docked poses were used as starting points for modeling in PyRosetta, which included optional local substrate docking refinement[7], constrained sidechain and backbone minimization in a Cartesian energy function into the pre-transition states, and scoring the final poses with a variety of metrics[1,8,9] over many simulation replicates. In each zero-shot event, to compute specific activity, thermostability, and expression of each mutated and wild-type sequence, metric values were normalized between 0 and 1 (where higher is better for each metric) to the global maximum and global minimum metric values across all sequences, and a linear combination of the normalized metric values were calculated with surmised weights. For the Xylanase zero-shot event, the histogram of the resulting expression scores was computed with three bins for classifying the expression scores of each mutated sequence into three classes: no expression, low expression, and good expression.

Supervised Round

In the supervised Alpha-Amylase, Imine reductase, β-glucosidase B, and Alkaline Phosphatase PafA events, macromolecular models of the pre-transition state of each enzyme and substrate(s) were prepared and scored in PyRosetta[1] similarly to the Zero-shot round. However, mutated sequences in the training data set that were missing expression, thermostability, and/or specific activity data were filtered out of the training set. For the supervised Imine reductase event, a protein BLAST[10] of the mutated sequences in the test set revealed that the sequences were closely related to glucose-1-dehydrogenase from *Bacillus subtilis*, and the substrates for quantifying fold-improvement over positive control (FIOP) of the mutated sequences in the test set were chosen as D-glucopyranose and NADH. For the supervised Alpha-amylase event, only the training data from dataset 2 was used since mutated sequences in the test set were only from dataset 2, and the computed mean metric values across replicates were normalized to that of the

wild-type sequence. For the supervised Imine reductase event, the mean metric values across replicates were not normalized to that of the positive control, since the sequence of the positive control was not provided. The expression, thermostability, and specific activity values for duplicates of mutated sequences in the training data set were averaged for training predictive models. For each supervised event, across all of samples in the training data, if a certain metric scored in PyRosetta (i.e., feature) correlated with the expression, thermostability, or specific activity values (i.e., labels) with a squared Pearson product-moment correlation coefficient above a certain threshold, then the features were independently fit to the labels using spline interpolation with piecewise polynomials. The piecewise polynomial fits per metric were then globally fit in a linear combination with weights optimized toward predicting the expression, thermostability, or specific activity values across all samples in the training data. The global linear combination fits for the expression, thermostability, or specific activity data types were then smoothed by fitting a polynomial function up to degree 6. Using this methodology, the expression, thermostability, and specific activity predictions for each of the mutated sequences in the test set were computed algebraically from the mean of the selected metrics derived from 3-dimensional structural modeling of the pre-transition states in PyRosetta.

In vitro Round

For the *in vitro* Alpha-Amylase event, the predictive model developed in the supervised Alpha-amylase event was re-trained to enable predictions outside of the range of the original training set. To design novel Alpha-amylase mutants, initially 611 mutants were generated by taking all compatible combinations of the 37 variants with the highest specific activity from the training set in the supervised Alpha-amylase event. Furthermore, 1,620 mutants were generated by taking all compatible combinations of between 5-10 variants using the 38th variant with the highest specific activity through the remaining variants with specific activity ≥ 2.0. Concomitantly, 54 structured insertions between 6−45 residues were *de novo* designed on wild-type Alpha-amylase between residues 273 and 275 using RFDiffusion[11] and ProteinMPNN[12] in ColabDesign[2] using residues 130, 141, 142, and 212 as hotspots such that structured insertions may interact with the substrate. The 54 structured insertions were further designed to interact with the substrate using FastDesign[13] in PyRosetta, and insertions that were clashing with the substrate were filtered out. Each of the aforementioned combinations of mutations on wild-type backbone were optionally combined with each of the structured insertion designs. Due to limited computational resources in modeling and scoring all sequences in PyRosetta, the re-trained predictive model was used on the structured insertions and the combinations of mutations independently, and the specific activity scores were multiplied to predict combined specific activity scores used to initially rank designs for PyRosetta simulations. As designs were modeled and scored in PyRosetta, the final poses were scored and metrics used to predict the final specific activity scores using the re-trained predictive model. After modeling and scoring tens of thousands of designs in

PyRosetta, designs were ranked by predicted specific activity, and the 200 designs with the highest predicted specific activity were inspected in PyMOL[14] software and submitted for *in vitro* experimentation.

References

1. S. Chaudhury, S. Lyskov & J. J. Gray, PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, 26(5), 689-691 (2010).
2. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022 Jun;19(6):679-682. doi: 10.1038/s41592-022-01488-1. Epub 2022 May 30. PMID: 35637307; PMCID: PMC9184281.
3. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583-589. doi: 10.1038/s41586-021-03819-2. Epub 2021 Jul 15. PMID: 34265844; PMCID: PMC8371605.
4. Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, Alexander Rives. Evolutionary-scale prediction of atomic level protein structure with a language model. *Science* 379, 6637 (2023).
5. Varadi, M., *et al*. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* (2021).
6. Corso, G., Stärk, H., Jing, B., Barzilay, R., & Jaakkola, T. (2022). Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv* preprint arXiv:2210.01776.
7. Lemmon G, Meiler J. Rosetta Ligand docking with flexible XML protocols. Methods Mol Biol. 2012;819:143-55. doi: 10.1007/978-1-61779-465-0_10. PMID: 22183535; PMCID: PMC3749076.
8. Wang JYJ, Khmelinskaia A, Sheffler W, Miranda MC, Antanasijevic A, Borst AJ, Torres SV, Shu C, Hsia Y, Nattermann U, Ellis D, Walkey C, Ahlrichs M, Chan S, Kang A, Nguyen H, Sydeman C, Sankaran B, Wu M, Bera AK, Carter L, Fiala B, Murphy M, Baker D, Ward AB, King NP. Improving the secretion of designed protein assemblies through negative design of cryptic transmembrane domains. Proc Natl Acad Sci U S A. 2023 Mar 14;120(11):e2214556120. doi: 10.1073/pnas.2214556120. Epub 2023 Mar 8. PMID: 36888664; PMCID: PMC10089191.
9. Lauer TM, Agrawal NJ, Chennamsetty N, Egodage K, Helk B, Trout BL. Developability index: a rapid in silico tool for the screening of antibody aggregation propensity. J Pharm Sci. 2012 Jan;101(1):102-15. doi: 10.1002/jps.22758. Epub 2011 Sep 20. PMID: 21935950.

10. Mark Johnson, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, Thomas L. Madden, NCBI BLAST: a better web interface. Nucleic Acids Research, Volume 36, Issue suppl_2, 1 July 2008, pages W5–W9.
11. Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, Ahern W, Borst AJ, Ragotte RJ, Milles LF, Wicky BIM, Hanikel N, Pellock SJ, Courbet A, Sheffler W, Wang J, Venkatesh P, Sappington I, Torres SV, Lauko A, De Bortoli V, Mathieu E, Ovchinnikov S, Barzilay R, Jaakkola TS, DiMaio F, Baek M, Baker D. De novo design of protein structure and function with RFdiffusion. Nature. 2023 Aug;620(7976):1089-1100. doi: 10.1038/s41586-023-06415-8. Epub 2023 Jul 11. PMID: 37433327; PMCID: PMC10468394.
12. Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, Wicky BIM, Courbet A, de Haas RJ, Bethel N, Leung PJY, Huddy TF, Pellock S, Tischer D, Chan F, Koepnick B, Nguyen H, Kang A, Sankaran B, Bera AK, King NP, Baker D. Robust deep learning-based protein sequence design using ProteinMPNN. Science. 2022 Oct 7;378(6615):49-56. doi: 10.1126/science.add2187. Epub 2022 Sep 15. PMID: 36108050; PMCID: PMC9997061.
13. Maguire JB, Haddox HK, Strickland D, Halabiya SF, Coventry B, Griffin JR, Pulavarti SVSRK, Cummins M, Thieker DF, Klavins E, Szyperski T, DiMaio F, Baker D, Kuhlman B. Perturbing the energy landscape for improved packing during computational protein design. Proteins. 2021 Apr;89(4):436-449. doi: 10.1002/prot.26030. Epub 2020 Dec 11. PMID: 33249652; PMCID: PMC8299543.
14. PyMOL, The PyMOL Molecular Graphics System, Version 2.5.0, Schrödinger, LLC.

Baseline Submission: Greedy Recombination of Beneficial Mutations and Model-Guided Protein Sequence Space Sampling

Hassan Kane[1*], Yasser Ibrahim[1], Adil Yusuf[1*]

[1]*Medium Biosciences, Boston, Massachusetts*

*Corresponding authors: *founders@medium.bio

This writeup introduces a baseline strategy for the greedy recombination of beneficial mutations and a model-guided approach to protein sequence space sampling. This method was used to provide a list of 30 amino acid sequences that maximize enzyme activity while maintaining at least 90% of the parent sequence's stability and expression. The goal of this method is to provide a simplet, yet competitive baseline.

Dataset

We only used the Alpha Amylase dataset. The dataset contains sequence-assay scores for alpha amylase variants expressed using bacillus subtilis for activity against RBB corn starch substrate. Stability and expression properties are also measured. The dataset contains sequence-assay scores for two experimental conditions: the first with 8075 data points focusing on single point mutants (7575 with associated data), and the second with 1897 data points, encompassing mutations ranging from 1 to 11 per sequence.

Methods:

Our methodology encompassed three distinct approaches with 10 sequences selected for each approach

- Purely Greedy Approach: This involved programmatically identifying mutations frequently found in sequences that exceeded the parent enzyme's activity, with stability and expression maintained at a minimum of 90%. From this list, we removed any mutations found in sequences that had all three properties decreased. We rank-ordered these sequences by activity, looked at the number of mutations for top performing sequences and shuffled the top-performing mutants at this edit distance weighted by their frequency in the top sequences and selected some at random.
- Model-Guided Greedy Approach: Here, we focused on mutations that improved all three properties and removed those detrimental to them. An exhaustive recombination list was generated. Using one-hot[1], ESM1V[2], ESM1B[3] and Georgiev Embeddings[4] as sequence feature vectors, we benchmarked machine learning models with the selection based on the lowest mean squared error via 5-fold cross-validation of all three properties. The

model then selected the most promising mutants.

- Markov Chain Monte Carlo (MCMC)[5]: We employed one-hot vectors along with the best-performing model (random forest) for each property. MCMC sampling of the sequence space was conducted, maintaining a trust radius of 8 and a temperature of 0.1, with 20,000 samples taken over 50 steps.

References

1. Jing, Xiaoyang, et al. "Amino acid encoding methods for protein sequences: a comprehensive review and assessment." *IEEE/ACM transactions on computational biology and bioinformatics* 17.6 (2019): 1918-1931.
2. Meier, Joshua, et al. "Language models enable zero-shot prediction of the effects of mutations on protein function." *Advances in Neural Information Processing Systems* 34 (2021): 29287-29303.
3. Rives, Alexander, et al. "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences." *Proceedings of the National Academy of Sciences* 118.15 (2021): e2016239118.
4. Georgiev, Alexander G. "Interpretable numerical descriptors of amino acid space." *Journal of Computational Biology* 16.5 (2009): 703-723.
5. Biswas, Surojit, et al. "Low-N protein engineering with data-efficient deep learning." *Nature methods* 18.4 (2021): 389-396.