AI-ACCELERATED BIOCATALYST ENGINEERING BY RAPID MICROFLUIDIC SEQUENCE-FUNCTION MAPPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Engineering biocatalysts is central for sustainable chemical synthesis, but hampered by a lack of sequence-function data which is costly and slow to obtain. We introduce a new microfluidic workflow, 'Droplet lrDMS', which allows us to screen tens of thousands of enzyme variants within two weeks, a scale, speed and cost not feasible with plate screening or robotic workflows. Using this workflow, we generate large-scale sequence-function data of an imine reductase and rationally engineer improved variants with an up to 11-fold improvement in catalytic efficiency (k_{cat}/K_M) vs wild type. With machine learning, we further enhance catalytic efficiency up to 16-fold vs wild type, 4-fold better than the best variant in the dataset, by combining rational engineering and predictions from the AI model. The improvement is driven by a 24-fold improvement of catalytic rate (k_{cat}) over wild type significantly higher than rate improvements observed in an AI-informed campaign with a similar enzyme. Our study demonstrates the potential of droplet lrDMS sequence-function data to accelerate directed evolution by AI-informed biocatalyst engineering.

1 INTRODUCTION

Biocatalysts will enable the transition to a green chemical industry by providing a sustainable and atom efficient route to valuable chemicals (Buller et al., 2023; Sheldon, 2023; France et al., 2023). For example, both the economic viability and environmental footprint of the synthesis of chiral amines, important high-value intermediates in the pharmaceutical industry, can be improved using imine reductases (IREDs) (Kumar et al., 2021; Schober et al., 2019).

Biocatalysts are usually sourced from naturally occurring enzymes, but often have insufficient activity towards the non-natural substrates under the desired reaction conditions. To engineer biocatalysts for industrial application, directed evolution – an iterative cycle of mutagenesis and experimental testing to screen for variants with the desired properties – is the method of choice (Arnold, 2019). However, functional enzymes are rare in the vastness of sequence space (Keefe & Szostak, 2001) and trajectories to higher fitness can be hampered by the non-additivity of mutations (Miton & Tokuriki, 2016). Consequently, directed evolution is labor, time and resource intensive with unpredictable outcomes, making it a hit-and-miss approach (Buller et al., 2023; Truppo, 2017).

Artificial intelligence (AI)-based methods are promising alternative avenues for faster and in silico engineering of enzymes (Lu et al., 2022; Ma et al., 2021; Büchler et al., 2022). Zero shot approaches are often aimed at producing more stable and foldable enzymes while approaches aiming for improving catalytic activity are bottlenecked by the availability of assay-labelled data.

The success of directed evolution depends on the throughput of experimental screens: the more randomly generated mutants can be experimentally tested, the more likely is it to identify an improved catalyst. However, this experimental process is expensive when carried out at scale. A massive scale-down of screening volumes (and thus reagent costs) is possible when water-in-oil emulsion droplets (formed in microfluidic devices) with picoliter volumes are used instead of multiwell plates. Droplet microfluidics offers a screening capacity of millions of enzyme variants a day and is cheap (with an assay cost as low as 10^{-5} cents per library member (Agresti et al., 2010) and versatile (Gantz et al., 2023; Debon et al., 2019; Schnettler et al., 2022).

By combining droplet microfluidics with next-generation and nanopore sequencing (Zurek et al., 2020), we generate sequence-function data at ultrahigh-throughput for engineering an exemplary IRED. This data can then be used to perform rational and AI-enabled in silico biocatalyst engineer-

ing to rapidly extrapolate beyond the improvements encoded in the library, accelerating biocatalyst engineering by maximizing the improvement of each round of directed evolution. In summary, our key contributions are:

1. We introduce a new microfluidic workflow (droplet lrDMS) to rapidly generate quantitative sequence-function data at large scale. Specifically, droplet lrDMS allows us to quantify the function of 10^4 randomly selected single and higher-order variants within 2 weeks, at a scale, speed and cost not feasible with plate screening or robotic workflows.
2. We introduce an analytical framework to use the unique structure of droplet lrDMS data to successfully perform rational engineering on an imine reductase, achieving up to an 11-fold improvement in catalytic efficiency ($k_{\text{cat}}/K_{\text{M}}$) over wild type.
3. We utilize machine learning to extrapolate from the observed single mutants to all single and double mutants. We further test the value of these extrapolations for protein engineering by assaying the top 5 predicted single and double mutants, identifying variants with better catalytic efficiency than the best variant in the dataset. By combining AI extrapolation with rational engineering, we achieve an up to 16-fold catalytic efficiency improvement, driven by a 24-fold improvement in k_{cat}.

2 RESULTS

Droplet lrDMS workflow We develop a data-generation workflow for enzyme engineering, *droplet long-read deep mutational scanning*, for screening the catalytic activity of randomly generated mutants (Figure 1A). The workflow starts by generating a library of multiple 10^4 random mutants via error-prone PCR (ePCR). Each mutant gene is tagged via a unique molecular identifier (UMI) and the library is sequenced with long-read consensus sequencing (Zurek et al., 2020) to create an ID-to-sequence dictionary. Another portion of the amplified library is combined with the substrates into picoliter sized droplets and incubated. Sensor dye is injected, which couples the amount of leftover substrate to optical absorbance, into each droplet. This allows to rapidly sort the incubated library for activity above a defined absorbance threshold (Gielen et al., 2016). We then sequence the UMI-ID of the *active* bin, as well as the input library, and count the occurrences of each unique UMI-ID. The ratio of counts of each UMI-ID, and therefore variant, at the sorted output versus the input gives an enrichment score that is linked to catalytic activity. By normalising the enrichment ratios to the wild type enrichment, we obtain quantitative fitness scores of each variant relative to the wild type. Droplet lrDMS can be executed within 2 weeks to generate 10^4 unique sequence-function mappings.

Generation of sequence-function data for an IRED To test this new workflow we aim to improve the activity of an IRED as exemplary enzyme engineering challenge. For our IRED starting point (SrIRED, (Lenz et al., 2018)) we used droplet lrDMS to obtain 10905 individual variants (single & higher order mutations) with assigned fitness scores and additional 6238 variants which were classified as entirely inactive. To assess whether lrDMS fitness captures quantitative information about catalytic activity, we picked 14 random variants, 7 from the input library and 7 from the output library after sorting, and measured conversion in a lysate assay in plates. The logarithmic activity in plates correlates well with the fitness score (Pearson $r = 0.92$) suggesting that lrDMS fitness captures quantitative catalytic activity information of variants with both, smaller and larger activity than wild type (Figure 1B).

Mutability and combinability as a tool for rational IRED engineering. Droplet lrDMS fitness scores provide us with a profile that quantifies how mutations across the SrIRED sequence and structure affect activity and how they can be combined to achieve high activity in higher-order variants. As a proxy for mutability, i.e. the capacity of a site to be functional after mutation, we calculated the median fitness of all single variant fitness scores for each position of the IRED (Figure 2A, Equation (1)). Positions with high mutability are defined as hotspots for IRED engineering. Additionally, we mapped mutability onto the IRED crystal structure, a domain-swapped dimer with two identical active sites with one reaction partner and the NADPH cofactor bound. First shell (residues contacting the substrate directly) and the proposed catalytic resides (Sharma et al., 2018; Lenz et al., 2018) show very low mutability except for T241 which contacts the ketone substrate and is positioned in the binding pocket flanked by all three catalytic residues (Figure 2C). Hotspots are especially abundant in regions flanking the cofactor binding pocket (around positions 38 and 69) and in the dimer interface (e.g. at position 203). Upon characterization of 12 single point mutations with a fitness > 0

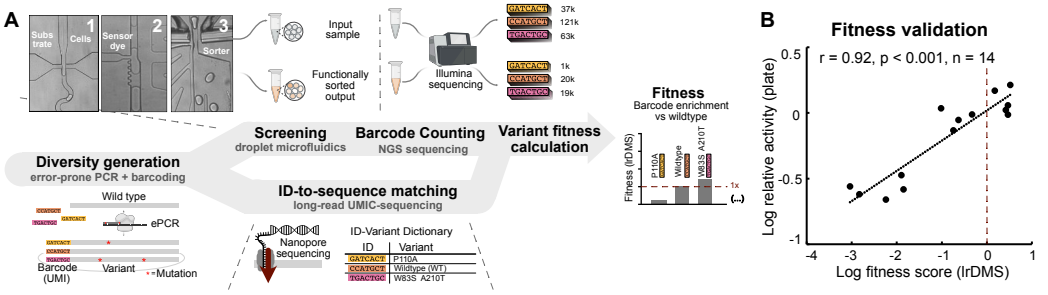

Figure 1: **Droplet lrDMS: A microfluidic workflow for rapid large-scale sequence function data generation of enzymes.** (**A**) The droplet lrDMS workflow is explained in the main text and allows generating a sequence-function mapping for 10^4 randomly generated variants around a wild type in 2 weeks. (**B**) Correlation of lrDMS fitness scores with plate-based colorimetric lysate assay data relative to the wild type for 14 random variants from the input library and the library after sorting. The high correlation of lrDMS fitness with plate based lysate activity indicates that lrDMS data can give quantitative fitness information about variants.

from hotspots and other positions and the variant with the highest fitness in the dataset (the double mutant V6L V67I), we concluded that the 4-fold improvement in catalytic efficiency ($k_{\text{cat}}/K_{\text{M}}$) of T241A represents the highest improvement present in the droplet lrDMS library.

Hotspots are useful for the identification of highly active point mutants, but combining them into improved higher-order variants is often hampered by unpredictable, negative non-linear effects (negative epistasis). Positive synergistic effects, on the other hand, are highly desirable but much less common and equally hard to predict. To inform the engineering of higher-order variants we define a *combinability* metric (Equation (2)) as the occurrence of mutations in higher-order variants which (1) improving over wild type and (2) are not negatively epistatic. According to this definition, we find T241A a global combinability hotspot (Figure 2B).

We next used this information to extrapolate to higher-order variants that have not been observed in the dataset by combining the combinability hotspot T241A with a set of variants with different background. We chose P203A, which lies at the dimer interface and has high mutability and combinability. Individually P203A's improvements in k_{cat} (rate under substrate saturation) are offset by an increased K_{M} (a proxy of the affinity to the substrate). We also chose K38E and D69H from hotspots flanking the active site. Individually, they each have modest improvements in k_{cat}, with position 38 showing higher combinability than position 69. Finally, we chose the variant with the second-highest fitness in the first shell, L174V, which features a high improvement in k_{cat} which is however offset by an increase in K_{M}. Three of four combined variants show a high synergistic improvement in $k_{\text{cat}}/K_{\text{M}}$, as compared to what would be expected from a purely additive effect, with the highest improvement 11-fold for T241A P203A. We hypothesise that the unsuccessful combination T241A L174V may be related to similar redundant mechanisms of improvement of L174 and T241 that can be rationalised by their close proximity in the active site. In summary, the rational engineering results show that droplet lrDMS-based combinability information combined with fitness information on individual variants can be used to rationally engineer highly efficient higher-order variants.

Single mutant AI-engineering. To assess whether AI modelling can successfully extrapolate to the single-mutant space, we follow Hsu et al. (2022) and build an ensemble of a ridge regression model, which is augmented with an evolutionary density from a large language model, to predict fitness scores for unobserved single mutants. As language model score we use the sequence pseudo-log-likelihood (PLL) of ESM2 (Rives et al., 2021). For the ridge-regression we swap the one-hot encoding of amino acids for the first 19 principal components of the AAINDEX database (Kawashima & Kanehisa, 2000), a database of physicochemical properties of amino acids. As compared to the one-hot encoding, this chemically informed encoding better captures the similarity of amino acids. Using the lrDMS data, we train the model on the observed single mutants and evaluate retrospectively on a hold-out test set, achieving Spearman $\rho = 0.41$ and Top100 normalized discounted cumulative gain (NDCG) of $\text{NDCG}_{100} = 0.21$ (details in Appendix A.3). While these scores are modest, they are not uncommon for noisy biological data (Hsu et al., 2022; Notin et al., 2022) and significantly improve over the zero-shot performance of ESM2-PLL ($\rho = 0.33, \text{NDCG}_{100} = 0.14$) alone on our dataset (c.f. Appendix A.3).

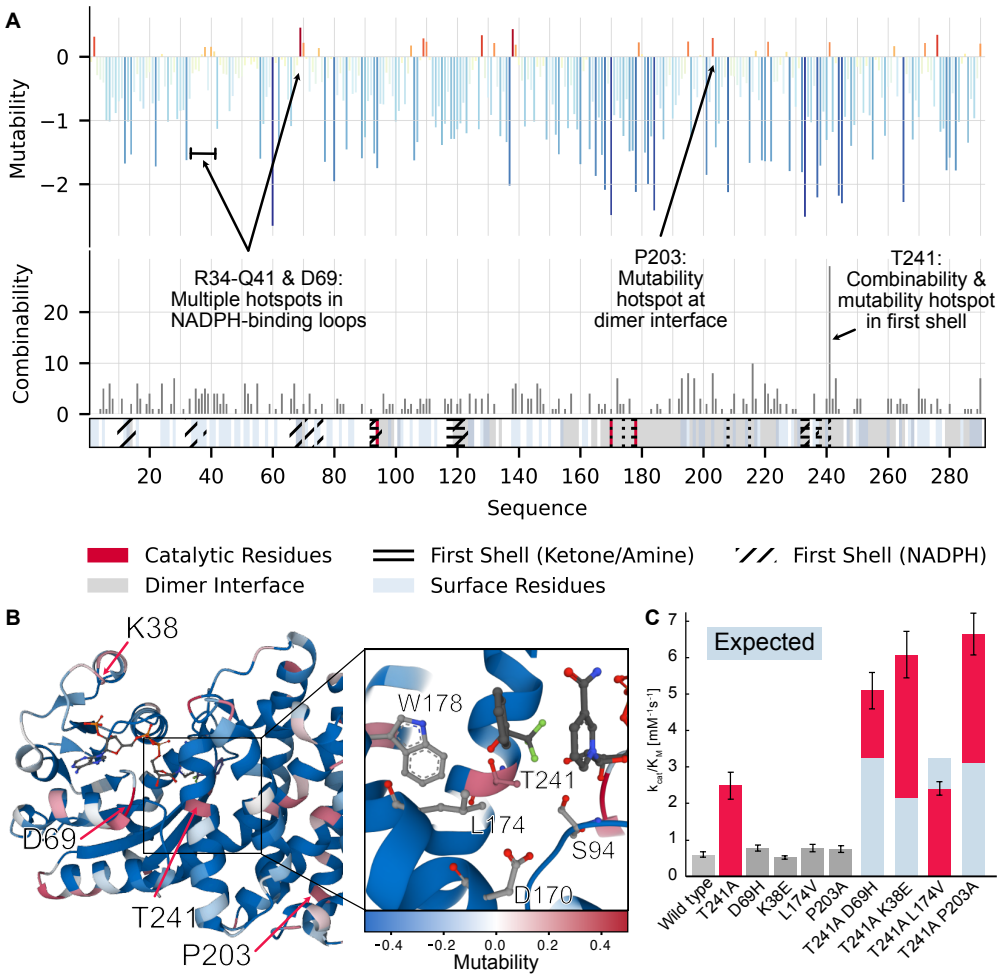

Figure 2: **A mutagenic profile of an IRED enables rational engineering.** (**A**) Median fitness per position (mutability) and the number of productive higher order combinations produced per position (combinability) along the IRED sequence. Proposed catalytic resides, first shell (around the ketone and amine binding pocket), surface and cofactor binding residues along with residues at the dimer interface are marked as indicated. (**B**) A map of mutability onto the IRED structure (5OCM) reveals hotspots e.g. around K38, D69 and at multiple positions central domain, e.g. at position P203. (**C**) Combining T241A with mutations at hotspots K38E, D69H, and P203A yields highly improving synergistic improvements (up to 11-fold $k_{\text{cat}}/K_{\text{M}}$).

To see whether the model can inform protein engineering, we clone the top 5 predicted single mutants and assay their functional characteristics (Figure 3A). We find that 2 of 5 predictions exceed $k_{\text{cat}}/K_{\text{M}}$ and 3 of 5 exceed k_{cat} of the wild type by more than a standard deviation. The top prediction, T241G, exceeds the catalytic efficiency of the best variant in the droplet lrDMS dataset by improving over wild type catalytic efficiency by 7-fold (Figure 3B). To understand whether this prediction was enabled by the droplet lrDMS data, we ablate the ESM component of the model and retrain it on the same data. Doing so does only change one of the top 5 predictions, with T241G remaining among the top 5. Conversely, to ablate the droplet lrDMS data we order all single mutants by the ESM score alone and find T241G ranked 457th out of 5510 possible single mutants. This suggests that the lrDMS data was crucial to align the model to what constitutes *better* for our task.

Double mutant AI-engineering. We next ask whether AI modelling can identify highly functional variants within the space of double mutants. We conjecture that, in addition to the language model score, combinability, lrDMS fitness of the individual mutants, structural and dynamic information are relevant to predict the interaction of two mutations. To this end we derive a set of structural and dynamic features from the wildtype structure (5OCM) (Appendix A.3). Based on these features, we train a simple gradient-boosted tree model (Chen & Guestrin, 2016) to predict the fitness of

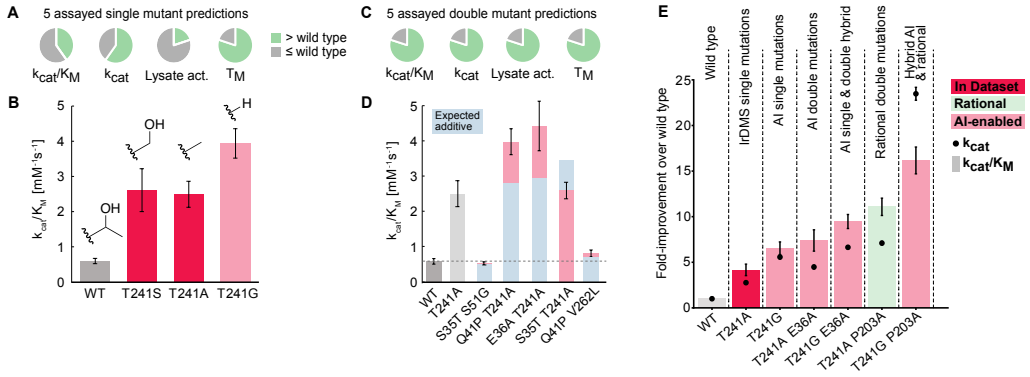

Figure 3: **AI modelling with lrDMS data. (A)** The fraction of the top 5 single mutant predictions that improve over the wild type in a given parameter by over one standard deviation are shown in green. **(B)** The catalytic efficiency of the top two single mutants in the lrDMS dataset and the best model prediction. **(C)** The top 5 predictions for the double mutant model are assayed as in (A). **(D)** The catalytic efficiencies of the 5 best predicted double mutants. **(E)** The catalytic efficiency and catalytic rates of the best single and double mutants in the dataset and the best model predictions.

unobserved double mutants. As with the single mutant model, we perform a retrospective evaluation of the double mutant model via cross-validation, achieving Spearman 0.38 and Top100 NDCG of 0.2 (details in Appendix A.3). Similar to the single mutant model, these scores are modest but again clearly improve over zero-shot ESM2-PLL ($\rho = 0.24, \text{NDCG}_{100} = 0.14$) alone, indicating that the models can learn from the droplet lrDMS data.

For prospective evaluation of the model's use for protein engineering, we evaluate the model on the over 1 Mio. possible double mutants constituted of single mutants observed in the lrDMS dataset and synthesize the top 5 predicted double mutants. The restriction on the double mutant space was chosen to ensure the model does not extrapolate too far from the lrDMS data given the limited experimental budget of 5 mutants. In practice one could extrapolate to all possible double mutants (or higher-order mutants with some model adjustments). In contrast to the zero-shot models, the model identifies T241A as a key mutation for building double mutants from the data and ranks many double mutants containing T241A among the top 100 predictions and 4 of the top 5 predictions contain T241A. This aligns with the rational analysis in Section 2, which identified T241A as a combinability hotspot. Of the top 5 predictions, 4 exceed the wild type in catalytic efficiency, catalytic rate, lysate activity and thermostability by more than a standard deviation (Figure 3C). The best prediction, T241A E36A leads to an 8-fold increase in catalytic efficiency over wild type (Figure 3D), which is better than the best variant in the droplet lrDMS dataset.The best variants were engineered by combining the results from the single mutant AI model with rational engineering and the double mutant AI model. Substitution of T241A with T241G from the single mutant model improves $k_{\text{cat}}/K_{\text{M}}$ by 10-fold for T241A E36A and 16-fold for T241G P203A (Figure 3E).

3 Discussion & Conclusion

Using ultra-fast data generation in microdroplets combined with AI-based extrapolation, we accelerate directed evolution by improving the catalytic efficiency of a biocatalyst beyond the limited improvement observed in the dataset. The improvement is driven by a 24-fold rate (k_{cat}) improvement, which is significantly higher than the rate improvements in AI-informed campaigns with a similar IRED (Ma et al., 2021). We identity the following features of our study as key for success: (i) The dataset was unbiased by position and included variants far off the first shell of interaction with the substrate, the focus of many enzyme engineering campaigns (Reetz, 2022). (ii) The dataset includes higher-order mutations which gives information on synergistic improvements and prevents negative epistatic effects. (iii) The dataset includes variants with neutral and negative effects. Neutral or deleterious mutations, such as K38E, deleterious in both $k_{\text{cat}}/K_{\text{M}}$ and lysate activity, give high synergistic improvements when combined with T241A. Consequently, protein engineering strategies that only carry improving mutations forward can miss out on highly improving, synergistic variants that are impossible to predict from single mutation data alone. (iv) As our learning curves indicate, fine-tuning zero shot models for catalysis requires large datasets. Obtaining dataset sizes $> 10^4$ would be extremely costly to achieve with traditional plate screening efforts and liquid handling robots.

REFERENCES

Jeremy J Agresti, Eugene Antipov, Adam R Abate, Keunho Ahn, Amy C Rowat, Jean-Christophe Baret, Manuel Marquez, Alexander M Klibanov, Andrew D Griffiths, and David A Weitz. Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proceedings of the National Academy of Sciences*, 107(9):4004–4009, 2010. (Cited on page 1)

Frances H Arnold. Innovation by evolution: bringing new chemistry to life (nobel lecture). *Angewandte Chemie International Edition*, 58(41):14420–14426, 2019. (Cited on page 1)

Antonia Boca and Simon Mathis. Predicting protein variants with equivariant graph neural networks. *arXiv preprint arXiv:2306.12231*, 2023. (Cited on page 10)

Johannes Büchler, Sumire Honda Malca, David Patsch, Moritz Voss, Nicholas J Turner, Uwe T Bornscheuer, Oliver Allemann, Camille Le Chapelain, Alexandre Lumbroso, Olivier Loiseleur, et al. Algorithm-aided engineering of aliphatic halogenase welo5* for the asymmetric late-stage functionalization of soraphens. *Nature Communications*, 13(1):371, 2022. (Cited on page 1)

R Buller, S Lutz, RJ Kazlauskas, R Snajdrova, JC Moore, and UT Bornscheuer. From nature to industry: Harnessing enzymes for biocatalysis. *Science*, 382(6673):eadh8615, 2023. (Cited on page 1)

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016. (Cited on page 4, 10, 11)

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022. (Cited on page 10)

Aaron Debon, Moritz Pott, Richard Obexer, Anthony P Green, Lukas Friedrich, Andrew D Griffiths, and Donald Hilvert. Ultrahigh-throughput screening enables efficient single-round oxidase remodelling. *Nature Catalysis*, 2(9):740–747, 2019. (Cited on page 1)

Scott P France, Russell D Lewis, and Carlos A Martinez. The evolving nature of biocatalysis in pharmaceutical research and development. *JACS Au*, 3(3):715–735, 2023. (Cited on page 1)

Maximilian Gantz, Stefanie Neun, Elliot J Medcalf, Liisa D van Vliet, and Florian Hollfelder. Ultrahigh-throughput enzyme engineering and discovery in in vitro compartments. *Chemical Reviews*, 123(9):5571–5611, 2023. (Cited on page 1)

Fabrice Gielen, Raphaelle Hours, Stephane Emond, Martin Fischlechner, Ursula Schell, and Florian Hollfelder. Ultrahigh-throughput–directed enzyme evolution by absorbance-activated droplet sorting (aads). *Proceedings of the National Academy of Sciences*, 113(47):E7383–E7389, 2016. (Cited on page 2)

Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Learning protein fitness models from evolutionary and assay-labeled data. *Nature biotechnology*, 40(7):1114–1122, 2022. (Cited on page 3, 10, 11)

Shuichi Kawashima and Minoru Kanehisa. Aaindex: amino acid index database. *Nucleic acids research*, 28(1):374–374, 2000. (Cited on page 3, 10, 11)

Anthony D Keefe and Jack W Szostak. Functional proteins from a random-sequence library. *Nature*, 410(6829):715–718, 2001. (Cited on page 1)

Rajesh Kumar, Michael J Karmilowicz, Dylan Burke, Michael P Burns, Leslie A Clark, Christina G Connor, Eric Cordi, Nga M Do, Kevin M Doyle, Steve Hoagland, et al. Biocatalytic reductive amination from discovery to commercial manufacturing applied to abrocitinib jak1 inhibitor. *Nature Catalysis*, 4(9):775–782, 2021. (Cited on page 1)

Maike Lenz, Silvia Fademrecht, Mahima Sharma, Jürgen Pleiss, Gideon Grogan, and Bettina M Nestl. New imine-reducing enzymes from β-hydroxyacid dehydrogenases by single amino acid substitutions. *Protein Engineering, Design and Selection*, 31(4):109–120, 2018. (Cited on page 2)

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. (Cited on page 11)

Hongyuan Lu, Daniel J Diaz, Natalie J Czarnecki, Congzhi Zhu, Wantae Kim, Raghav Shroff, Daniel J Acosta, Bradley R Alexander, Hannah O Cole, Yan Zhang, et al. Machine learning-aided engineering of hydrolases for pet depolymerization. *Nature*, 604(7907):662–667, 2022. (Cited on page 1, 10)

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. (Cited on page 15, 16)

Eric J Ma, Elina Siirola, Charles Moore, Arkadij Kummer, Markus Stoeckli, Michael Faller, Caroline Bouquet, Fabian Eggimann, Mathieu Ligibel, Dan Huynh, et al. Machine-directed evolution of an imine reductase for activity and stereoselectivity. *ACS Catalysis*, 11(20):12433–12445, 2021. (Cited on page 1, 5)

Charlotte M Miton and Nobuhiko Tokuriki. How mutational epistasis impairs predictability in protein evolution and design. *Protein Science*, 25(7):1260–1272, 2016. (Cited on page 1)

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pp. 16990–17017. PMLR, 2022. (Cited on page 3)

Manfred Reetz. Making enzymes suitable for organic chemistry by rational protein design. *ChemBioChem*, 23(14):e202200049, 2022. (Cited on page 5)

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, 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):e2016239118, 2021. (Cited on page 3, 10)

J David Schnettler, Oskar James Klein, Tomasz S Kaminski, Pierre-Yves Colin, and Florian Hollfelder. Ultrahigh-throughput directed evolution of a metal-free α/β-hydrolase with a cys-his-asp triad into an efficient phosphotriesterase. *Journal of the American Chemical Society*, 145 (2):1083–1096, 2022. (Cited on page 1)

Markus Schober, Chris MacDermaid, Anne A Ollis, Sandy Chang, Diluar Khan, Joseph Hosford, Jonathan Latham, Leigh Anne F Ihnken, Murray JB Brown, Douglas Fuerst, et al. Chiral synthesis of lsd1 inhibitor gsk2879552 enabled by directed evolution of an imine reductase. *Nature Catalysis*, 2(10):909–915, 2019. (Cited on page 1)

Mahima Sharma, Juan Mangas-Sanchez, Scott P France, Godwin A Aleku, Sarah L Montgomery, Jeremy I Ramsden, Nicholas J Turner, and Gideon Grogan. A mechanism for reductive amination catalyzed by fungal reductive aminases. *ACS Catalysis*, 8(12):11534–11541, 2018. (Cited on page 2)

Roger Arthur Sheldon. The e factor at 30: a passion for pollution prevention. *Green Chemistry*, 25 (5):1704–1728, 2023. (Cited on page 1)

Matthew D Truppo. Biocatalysis in the pharmaceutical industry: the need for speed. *ACS Medicinal Chemistry Letters*, 8(5):476–480, 2017. (Cited on page 1)

Bingxin Zhou, Lirong Zheng, Banghao Wu, Kai Yi, Bozitao Zhong, Pietro Lio, and Liang Hong. Conditional protein denoising diffusion generates programmable endonucleases. *bioRxiv*, pp. 2023–08, 2023. (Cited on page 10)

Paul Jannis Zurek, Philipp Knyphausen, Katharina Neufeld, Ahir Pushpanath, and Florian Hollfelder. Umi-linked consensus sequencing enables phylogenetic analysis of directed evolution. *Nature Communications*, 11(1):6023, 2020. (Cited on page 1, 2)

A Appendix

A.1 Characterisation of variants of interest

Variants with > 0 fitness

	Fitness	k_{cat}[s^{-1}]	K_M[mM]	k_{cat}/K_M[$mM^{-1}s^{-1}$]
wt	0	2.1 ± 0.1	3.5 ± 0.3	0.6 ± 0.1
T241A	**1.8 ± 0.5**	**5.8 ± 0.2**	**2.3 ± 0.3**	**2.5 ± 0.4**
T241S	**0.7 ± 0.6**	**6.0 ± 0.4**	**2.3 ± 0.5**	**2.6 ± 0.6**
M120L	0.3 ± 0.4	2.1 ± 0.1	5.3 ± 0.7	0.4 ± 0.1
L174V	1.1 ± 1.4	10.7 ± 0.5	13.8 ± 1.5	0.8 ± 0.1
D69H	0.5 ± 2.4	4.1 ± 0.2	5.3 ± 0.6	0.8 ± 0.1
D69G	0.5 ± 0.5	3.3 ± 0.1	6.0 ± 0.4	0.5 ± 0.04
D236G	0.4 ± 0.6	2.0 ± 0.1	3.1 ± 0.4	0.7 ± 0.1
P203A	2.9 ± 1.3	5.0 ± 0.2	6.7 ± 0.8	0.7 ± 0.1
P203S	0.7 ± 0.5	2.5 ± 0.1	3.2 ± 0.2	0.8 ± 0.1
E146V	0.9 ± 0.6	1.7 ± 0.1	2.4 ± 0.3	0.7 ± 0.1
D267G	0.2 ± 0.5	2.7 ± 0.1	4.3 ± 0.3	0.6 ± 0.1
D5H	1.5 ± 2.3	1.9 ± 0.1	5.1 ± 0.9	0.4 ± 0.1
V6L V67I	4.8 ± 2.3	1.7 ± 0.03	2.2 ± 0.3	0.7 ± 0.1

Variants from rational engineering

	Fitness	Combinability score	k_{cat}[s^{-1}]	K_M[mM]	k_{cat}/K_M[$mM^{-1}s^{-1}$]
wt	0		2.1 ± 0.1	3.5 ± 0.3	0.6 ± 0.1
T241A	1.8 ± 0.5	29	5.8 ± 0.2	2.3 ± 0.3	2.5 ± 0.4
P203A	2.9 ± 1.3	8	5.0 ± 0.2	6.7 ± 0.8	0.7 ± 0.1
K38E	0.0 ± 0.7	5	3.2 ± 0.1	6.3 ± 0.6	0.5 ± 0.5
D69H	0.5 ± 2.4	0	4.1 ± 0.2	5.3 ± 0.6	0.8 ± 0.1
L174V	1.1 ± 1.4	2	10.7 ± 0.5	13.8 ± 1.5	0.8 ± 0.1
T241A P203A	-		15.0 ± 0.4	2.3 ± 0.2	6.6 ± 0.6
T241A D69H	-		12.1 ± 0.3	2.4 ± 0.2	5.1 ± 0.5
T241A K38E	-		11.3 ± 0.3	1.9 ± 0.2	6.1 ± 0.6
T241A L174V	-		12.7 ± 0.3	5.3 ± 0.4	2.4 ± 0.2

Variants suggested by the AI model

	k_{cat} [s^{-1}]	K_M [mM]	k_{cat}/K_M [$mM^{-1}s^{-1}$]
wt	2.1 ± 0.1	3.5 ± 0.3	0.6 ± 0.1
T241G	11.7 ± 0.4	3.0 ± 0.3	3.9 ± 0.4
M1P	3.6 ± 0.1	5.1 ± 0.5	0.7 ± 0.1
E146L	2.1 ± 0.1	2.9 ± 0.3	0.7 ± 0.1
P203G	2.8 ± 0.1	5.1 ± 0.6	0.5 ± 0.1
A251I	not soluble		
S35T S51G	1.9 ± 0.05	3.6 ± 0.3	0.6 ± 0.04
Q41P T241A	6.4 ± 0.2	1.6 ± 0.1	4.0 ± 0.4
E36A T241A	9.4 ± 0.4	2.1 ± 0.3	4.4 ± 0.7
S35T T241A	3.3 ± 0.1	1.3 ± 0.1	2.6 ± 0.2
Q41P V262L	3.6 ± 0.1	4.4 ± 0.5	0.8 ± 0.1
S35T	2.8 ± 0.04	3.4 ± 0.2	0.8 ± 0.04
S51G	2.1 ± 0.1	5.5 ± 0.5	0.4 ± 0.04
Q41P	4.2 ± 0.1	6.3 ± 0.3	0.7 ± 0.03
T241A	5.8 ± 0.2	2.3 ± 0.3	2.5 ± 0.4
E36A	3.7 ± 0.1	5.2 ± 0.3	0.7 ± 0.05
T241G K38E	10.2 ± 0.6	3.9 ± 0.7	2.6 ± 0.5
T241G D69H	28.6 ± 0.9	4.1 ± 0.4	6.9 ± 0.7
T241G L174V	22.7 ± 0.8	6.2 ± 0.6	3.6 ± 0.4
T241G P203A	49.3 ± 1.5	5.1 ± 0.4	9.7 ± 0.9
E36A T241G	14.0 ± 0.3	2.5 ± 0.2	5.7 ± 0.5

Figure 4: Michaelis Menten kinetics were conducted with variable concentrations of ketone substrate (0.05 mM - 35 mM) with a constant concentration of amine (150 mM) in Tris pH 8.0 with NADPH (0.5 mM) with variable enzyme concentrations (from 0.03 μM to 0.3 μM) depending on the rate. Each curve was measured in three replicates. Errors represent standard errors of the fit and three technical replicates with a 95% confidence interval.

A.2 Rational engineering – Mutability & Combinability

To perform rational engineering, we define two key metrics – mutability and combinability – that guide our search for rational designs. For clarity, we describe these in detail below.

We use (p, a) to denote the substitution of position p in the sequence of a wild type to amino acid a. A variant v is then defined as a collection of K substitutions $v = \{(p_k, a_k)\}_{k=1}^{K}$, where each position may only occur at most once. The mutation order $|v|$ is the number of mutations K in v. For example, the single mutant T241A would be denoted $\{(241, A)\}$. The double mutant T241A P203A would be $\{(203, A), (241A)\}$, and so on.

Given these definitions, we define mutability of a position p_0 as

$$\text{Mutatability}(p_0) = \text{Median}(\{f(v)|v = \{(p_0, a)\} \wedge |v| = 1\}), \tag{1}$$

where $f(v)$ is the logarithmic lrDMS fitness score vs wildtype. In words, the mutatability of position p_0 is the median fitness over all single mutants in the dataset that occur at position p_0.

Further, we define combinability of position p_0 as

$$\text{Combinability}(p_0) = \sum_{\{v|(p=p_0,\cdot)\in v \wedge |v|>1\}} |v| \cdot \mathbf{1}_{[f(v)>0]} \cdot \mathbf{1}_{\left[f(v)\geq\sum_{(p,a)\in v} f((p,a))\right]}, \tag{2}$$

with $\mathbf{1}$ the indicator function, which is 1 if the condition is fulfilled and 0 otherwise. I.e. the combinability of position p_0 is the weighted sum over all higher order mutants, which contain a mutation at p_0 and have positive and non-negatively epistatic fitness. The weighting is given by the mutation order $|v|$. Note that this definition can be extended naturally to a combinability score for individual mutations (p, a) and not just for positions p.

A.3 AI MODELLING

Since droplet lrDMS allows us to functionally annotate a local region of sequence space, we hypothesized that lrDMS can successfully kick-start an AI modelling campaign. We were specifically interested in three questions:

1. Can AI modelling extrapolate successfully to the single-mutant space? This is important as diversity in droplet lrDMS libraries are generated via ePCR, which can generate only about a third of single mutants, because mutating more than one nucleotide per codon is unlikely.
2. Can AI modelling identify highly functional variants with higher mutation order in a low N regime, so that we can use it as an input for protein engineering?
3. How well would comparative zero-shot models perform and how crucial is droplet lrDMS data to *align* the model to what constitutes *fitness* for our task.

To answer these questions, we built two simple models: a single mutant model ridge regression model inspired by Hsu et al. (2022) and a double mutant gradient boosted tree model Chen & Guestrin (2016) based on a few curated features we deemed important for predicting the fitness of double mutants, most notably the input from lrDMS via combinability and single mutant fitness scores. Although there are more complex models available, we opted for these simpler and interpretable models to gain insights into the model components and the amount of training data required. We acknowledge that performance could potentially be enhanced by utilizing more complex models, such as those leveraging structural data through micro-environments (Lu et al., 2022) or incorporating assay labeled data similar to language models, which have been shown to provide complementary information (Boca & Mathis, 2023). Additionally, models that remodel significant portions of the protein, like inverse folding or diffusion models (Dauparas et al., 2022; Zhou et al., 2023), could also be explored. However, preliminary experiments indicate that more complex models are prone to overfitting on more noisy (albeit lrDMS data and finding the best ways of using more complex models with lrDMS data is an exciting avenue for future research. Given the focus of this study on understanding the value of droplet lrDMS data, we chose to employ simple, robust models.

We then evaluated these single and double mutant models in a retrospective manner, by performing cross-validation in multiple replicas on the observed lrDMS data, and in a prospective manner, by using the model to predict the fitness of unobserved single and double mutants and assaying the top 5 predictions in the lab. We also perform ablation and learning curve studies to understand the importance of the different components of the model and the amount of data required to train the model. The results of these studies are summarised in the main text in Figure 3 and Figure 6 and Figure 7. Below and in Figure 5 we proceed to describe the models and the results in more detail.

A.3.1 SINGLE MUTANT MODEL

The single mutant model is a ridge regression model, which is augmented with an evolutionary density from a large language model, to predict fitness scores for unobserved single mutants. As language model score we use the sequence pseudo-log-likelihood (PLL) of ESM2 (Rives et al., 2021). For the ridge-regression we swap the one-hot encoding of amino acids for the first 19 principal components of the AAINDEX database (Kawashima & Kanehisa, 2000), a database of physicochemical properties of amino acids. The 19 principal components capture over 98% of the variance in the AAINDEX database and are a chemically informed encoding of amino acids. As compared to the one-hot encoding, this chemically informed encoding better captures the similarity of amino acids. This chemical similarity between amino acids is crucial to help the model understand the effects of

unseen mutations at positions: The top prediction T241G for example substitutes glycine (G) instead of alanine (A) as in T241A, which is chemically similar but smaller than alanine (c.f. Figure 3B, where the side-chains of three highly performing T241 variants are shown). A schematic of the model is shown in Figure 5A.

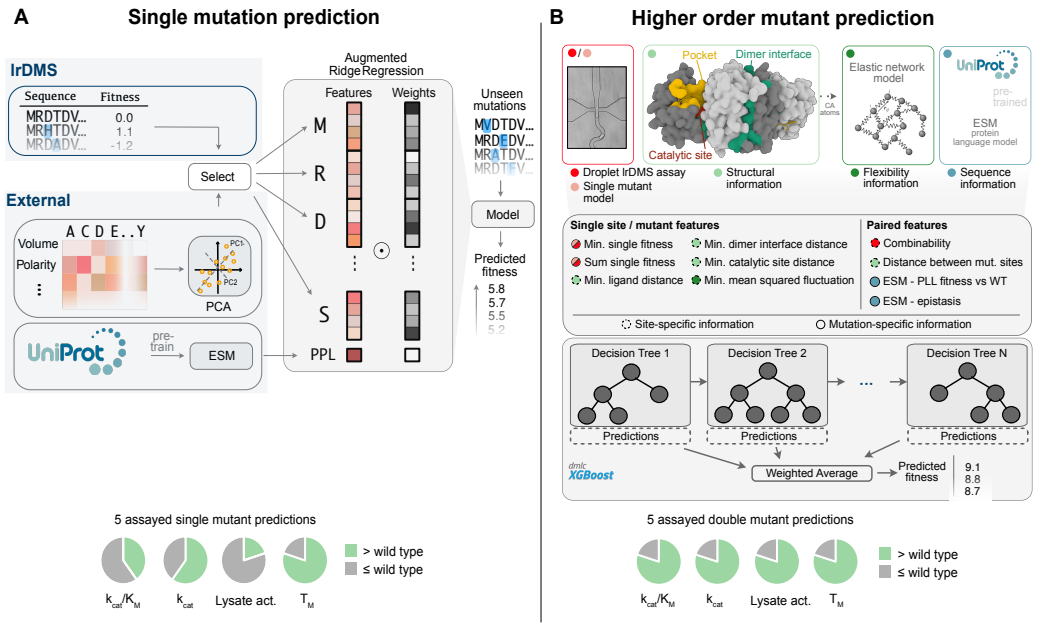

Figure 5: **Single and double mutant model schematics. (A)** Schematic of the single mutant model. The model is trained on the observed single mutants and uses the ESM2 (Lin et al., 2023) pseudo-log-likelihood (PLL) and the first 19 principal components of the AAINDEX database (Kawashima & Kanehisa, 2000) as features in a simple ridge regression (Hsu et al., 2022). The model is used to rank single mutants and the top 5 predictions are assayed. The pie charts show the fraction of these top 5 predictions that improve over the wild type in a given parameter by over one standard deviation in the lab. **(B)** Schematic of the double mutant model. The model combines lrDMS assay data (red) with structural (light green), dynamics (dark green) and sequence features (blue) to predict the fitness of unobserved double mutants. The features are passed to a simple gradient-boosted tree model (Chen & Guestrin, 2016), which is trained on the observed double mutants. After using the model to rank around 1 Mio. double mutants, the top 5 predictions are assayed experimentally for important catalytic properties as in (A).

A.3.2 DOUBLE MUTANT MODEL

For the double mutant model, we decided to use a gradient boosted tree model (xgboost, Chen & Guestrin (2016)) on a few selected features. The features can be grouped into four categories:

1. **Assay derived features**: Here we include the (1) fitness of single of the least fit mutation in the double mutant, as well the (2) additive fitness of both single mutations. This data can be derived directly from the dataset, or extrapolated from a single mutant model when the single mutations in question were not observed. Finally, we also include combinability information (3).
2. **Structural information**: To give the model a rough idea of the structural context of a given double mutation we provide (1) the minimum distance of the singles that comprise the double mutant to the NADPH binding pocket, (2) the minimum distance of said singles to the dimer interface and (3) the catalytic site residues. We also include (4) the pairwise distance between the C_α atoms of the two mutants.
3. **Coarse grained dynamics information**: We use normal mode analysis on an elastic network model of the protein structure to derive the mean squared expected fluctuation, a

value that mimics b-factors, of each position. We use provide the mean squared fluctuation of both single mutants (min & max) as a feature.

4. **Language model information**: Finally, we include sequence and evolutionary information through (1) the pseudo-log likelihood (PLL) of the ESM2 language model as well as an ESM2 'epistasis' feature (2) which predicts the difference between ESM2 PLL for the double mutant versus the sum of the PLL predictions for the single mutants.

We investigated the relative importance of these features for our top 10 double mutant predictions (Figure 9 & Figure 10) as well as for all double mutant predictions Figure 11. As can be seen, the assay derived features play a key role in the predictions as a whole, and particularly for the predictions at the top end.

A.3.3 Retrospective analysis: Learning curves and ablations

To understand the scale of data that is needed to inform a model about *function* in our IRED engineering task we perform learning curves, which evaluate the model's performance on hold-out data for increasing training set sizes.

For the single mutant model we only have about $1'200$, so we perform 10 replicas of 3-fold cross validation (30 replicas in total) to reduce batch effects at the evaluation due to small evaluation datasets. For the double mutant model we perform 5 replicas of 6-fold cross validation. The learning curves are described in more detail in the captions of Figure 6 and Figure 7.

A conclusion from the learning curves is that the single mutation model needs to see a mutation (other than WT) at each position of the sequence at least once to get information about how the local chemistry at that position influences the final readout. Real improvements are then only obtained from an order of magnitude of 300 onwards – which corresponds to the sequence length of our IRED. We could also show that ESM has a minor influence on the prediction of top hits with the model mostly relying on lrDMS data for its prediction.

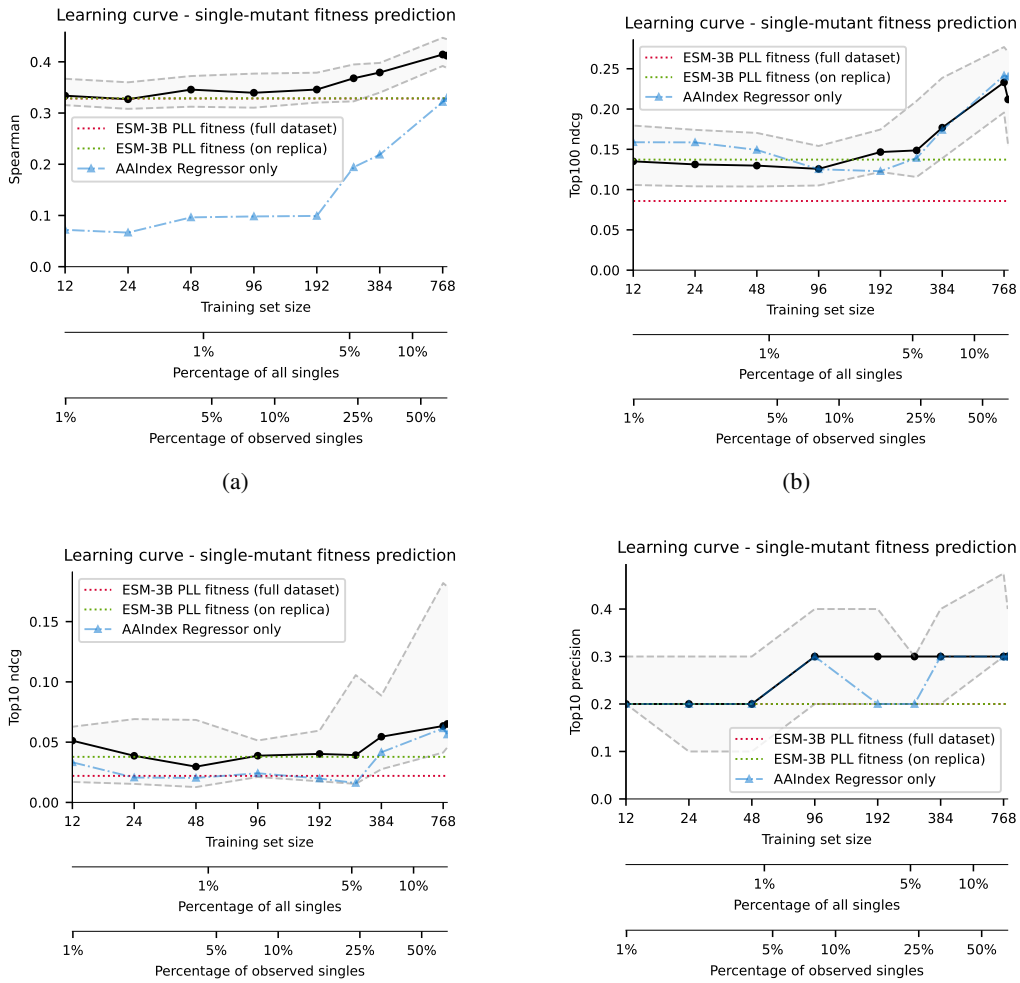

Figure 6: Evaluation of the effect of dataset size on training outcome. Learning curves representing training of the model with different amounts of data. We produced 30 replica (10 replicas of 3-fold cross validation). For each replica, we successively increase the dataset size that we train the model on from 0% to the full 67%of the singles data (the remaining 33% are the holdout of that replica and what we evaluate on). At each point we compute success metrics of the model (a-d) and then compute the median and quantiles across 30 replica. The black line corresponds to the median performance across 30 replicas (3 fold cross validation with 10 replicas each). The top / bottom dashed gray line corresponds to the 75% and 25% quantiles respectively. The dotted red line shows the "zero-shot" performance that one would get on the entire single mutant dataset when using the 3B parameter ESM2 model and using pseudo log likelihood (PLL) as fitness. The dotted green line shows the "zero-shot" median performance that the ESM-PLL strategy would have across the 30 replicas (this is different from the red line because of batch effects: each replica has a distinct 33% of the data that is used to evaluate performance, rather than the entire 100% of the data that are used to obtain the dotted red line). The blue line is an ablation of ESM derived features, which are the most expensive to compute, and demonstrate that our model can be performant even on a table-top computer without GPU, where one would omit the ESM features. It also demonstrates that the key information is carried by the lrDMS data itself, rather than the language model.

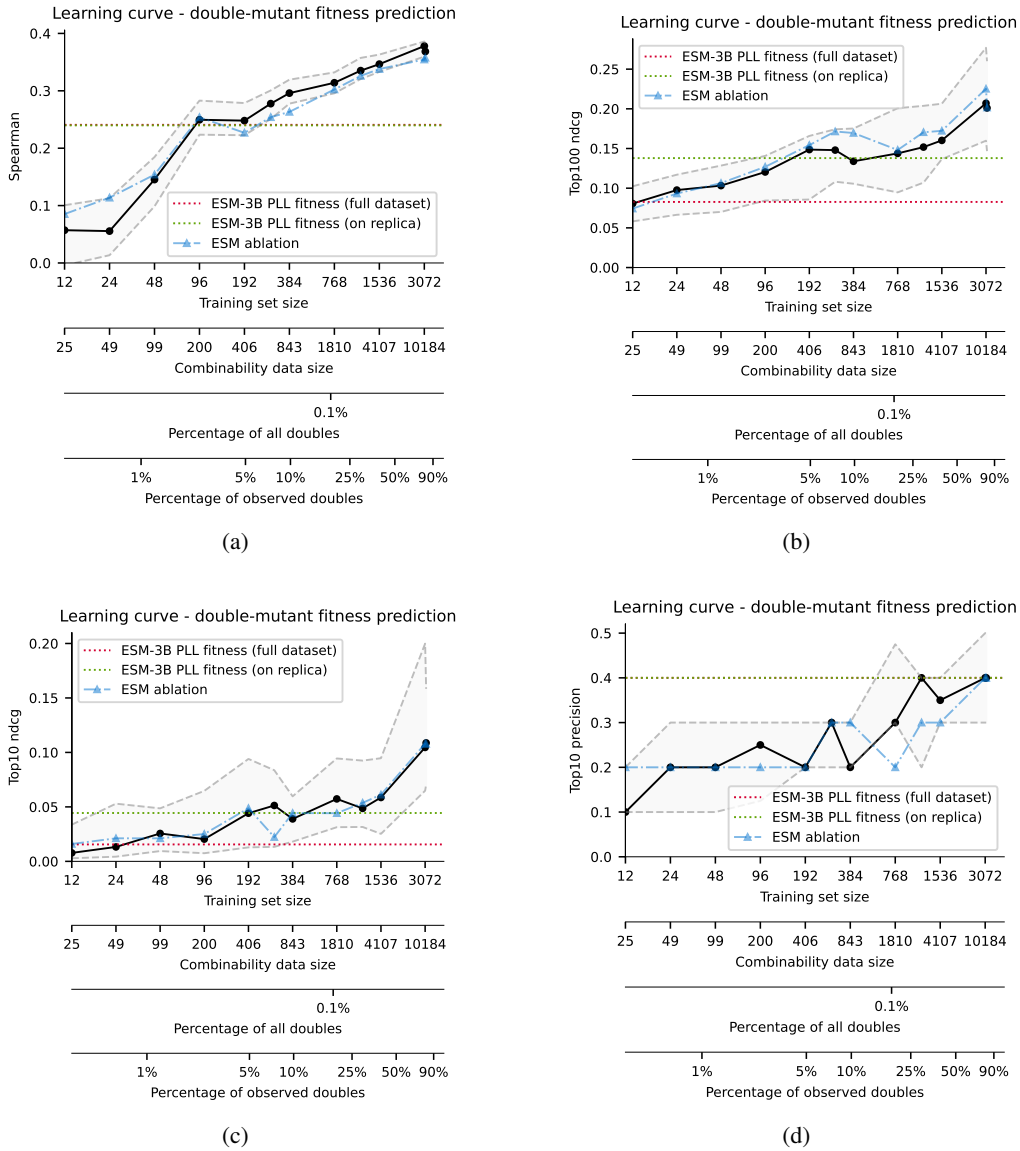

Figure 7: Learning curves for the double mutant model. The learning curves were calculated in the same way as learning curves for the single mutation model (Figure 6), but we used 6 replicas of 5-fold cross validation (30 replicas in total) instead to explore extrapolation at higher dataset sizes. The black line corresponds to the median performance across 30 replicas. The top / bottom dashed gray line corresponds to the 75% and 25% quantiles respectively. The dotted red line shows the "zero-shot" performance that one would get on the entire single mutant dataset when using the 3B parameter ESM2 model and using pseudo log likelihood (PLL) as fitness. The dotted green line shows the "zero-shot" median performance that the ESM-PLL strategy would have across the 30 replicas (this is different from the red line because of batch effects. The blue line is an ablation of ESM derived features, which are the most expensive to compute, and demonstrate that our model can be performant even on a table-top computer without GPU, where one would omit the ESM features. It also demonstrates that the key information is carried by the lrDMS data itself, rather than the language model.

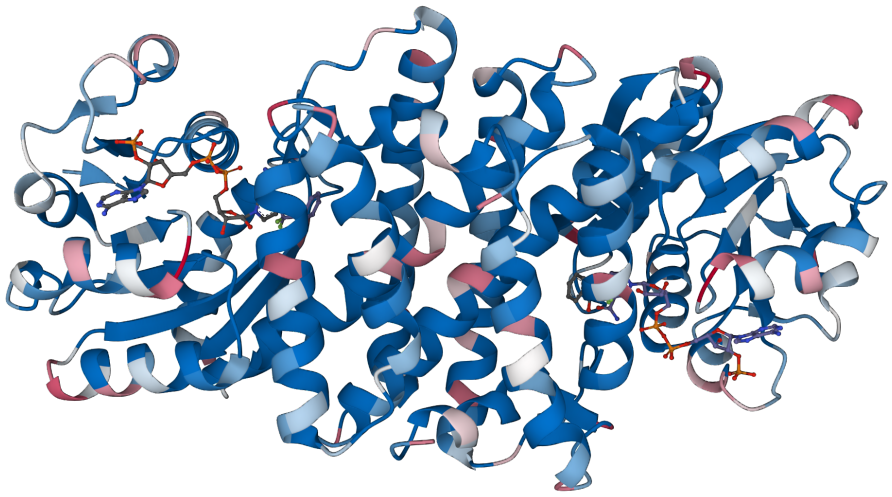

Figure 8: Mutability plotted across the entire SrIRED (high mutability in red, low mutability in blue), which is a homodimer (composed of two identical chains) with two catalytic sites to the left and right of the figure. The bound ligands (NADPH and ketone) are shown in sticks.

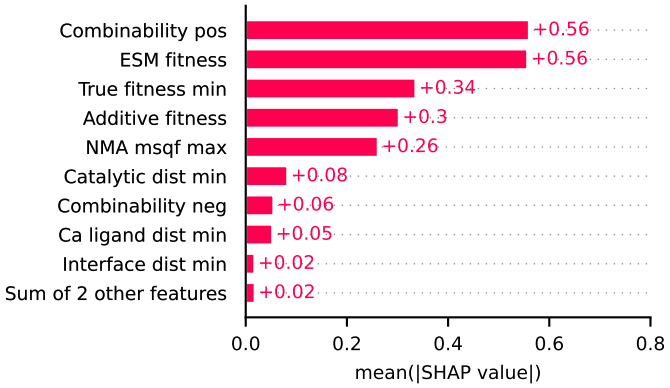

Figure 9: Mean absolute SHAP values (Lundberg & Lee, 2017) as feature importance for the top 10 double mutant predictions.

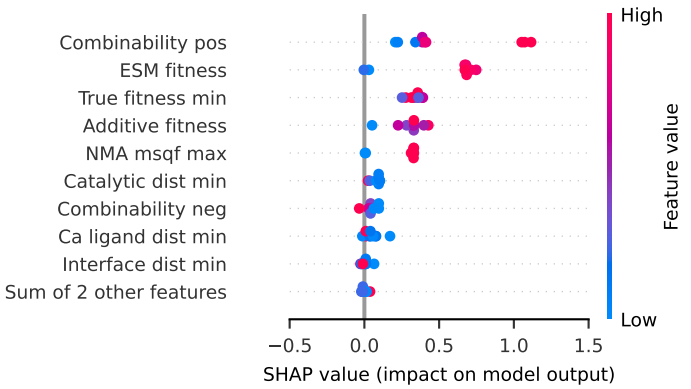

Figure 10: SHAP (Lundberg & Lee, 2017) values as a proxy for feature importance for the top 10 double mutant predictions.

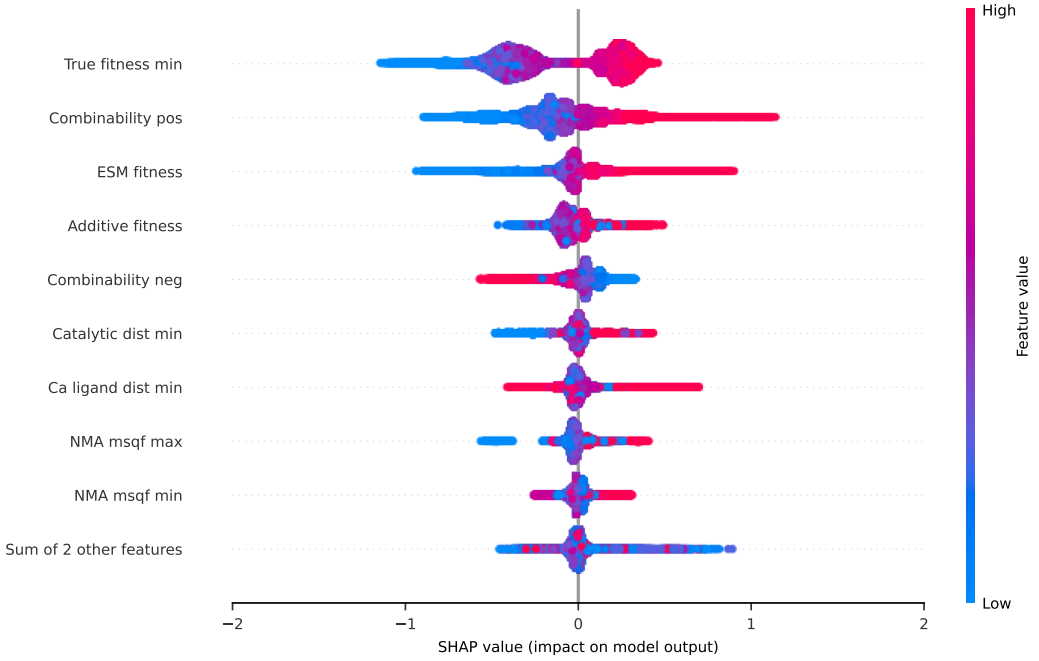

Figure 11: SHAP Lundberg & Lee (2017) values as a proxy for feature importance for all double mutant predictions.