Microdroplet screening rapidly profiles a biocatalyst to enable its AI-assisted engineering

Maximilian Gantz[1], Simon V. Mathis[2], Friederike E. H. Nintzel[1], Paul J. Zurek[1], Tanja Knaus[3], Elie Patel[1], Daniel Boros[1], Friedrich-Maximilian Weberling[1], Matthew R. A. Kenneth[1], Oskar J. Klein[1], Elliot J. Medcalf[1], Jacob Moss[2], Michael Herger[1], Tomasz S. Kaminski[1,4], Francesco G. Mutti[3], Pietro Lio[2], Florian Hollfelder[1*]

[1]Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge, CB2 1GA, UK

[2]Department of Computer Science, University of Cambridge, 15 JJ Thomson Avenue Cambridge CB3 0FD, UK

[3]Van´ t Hoff Institute for Molecular Sciences, HIMS-Biocat, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, Netherlands

[4]Department of Molecular Biology, Institute of Biochemistry, Faculty of Biology, University of Warsaw, Miecznikowa 1, 02-096 Warsaw, Poland

*Correspondence: fh111@cam.ac.uk

Abstract

Engineering enzyme biocatalysts for higher efficiency is key to enabling sustainable, 'green' production processes for the chemical and pharmaceutical industry. This challenge can be tackled from two angles: by directed evolution, based on labor-intensive experimental testing of enzyme variant libraries, or by computational methods, where sequence-function data are used to predict biocatalyst improvements. Here, we combine both approaches into a two-week workflow, where ultra-high throughput screening of a library of imine reductases (IREDs) in microfluidic devices provides not only selected 'hits', but also long-read sequence data linked to fitness scores of >17 thousand enzyme variants. We demonstrate engineering of an IRED for chiral amine synthesis by mapping functional information in one go, ready to be used for interpretation and extrapolation by protein engineers with the help of machine learning (ML). We calculate position-dependent mutability and combinability scores of mutations and comprehensively illuminate a complex interplay of mutations driven by synergistic, often positively epistatic effects. Interpreted by easy-to-use regression and tree-based ML algorithms designed to suit the evaluation of random whole-gene mutagenesis data, 3-fold improved 'hits' obtained from experimental screening are extrapolated further to give up to 23-fold improvements in catalytic rate after testing only a handful of designed mutants. Our campaign is paradigmatic for future enzyme engineering that will rely on access to large sequence-function maps as profiles of the way a biocatalyst responds to mutation. These maps will chart the way to improved function by exploiting the synergy of rapid experimental screening combined with ML evaluation and extrapolation.

Main

Protein engineering has advanced enormously over the last decades, successfully employing a range of experimental[1] and computational approaches[2-8]. Purely computational approaches - avoiding the laborious iterative cycles of mutagenesis and experimental testing in directed evolution - reliably lead to stabilised proteins, even new folds, and also to binders with improved or altered functions[9]. The engineering of enzymes, however, remains a challenge, because - in addition to stability, folding and binding - the sub-Ångstrom precision of the catalytic machinery (including possible conformational changes as bonds are made and broken) place multiple biophysical and chemical demands on an engineered enzyme. The combinatorial space of possible amino acid variations during an engineering task is vast[1], while relatively few mutations satisfy the complex requirements for catalysis. Beyond identification of improving single mutations, their combination in constructive evolutionary trajectories can be hampered by their non-additivity, i.e. negative epistatic interactions between amino acids[10,11] pose a major practical challenge in enzyme engineering.

At the same time, the industrial transition to a green bioeconomy urgently requires new enzymes for biocatalysis, to provide sustainable routes to valuable chemicals: atom efficient, with chiral control and with a more benign environmental footprint than current processes based on chemical catalysis[12-14]. Biocatalysts can be sourced from naturally occurring enzymes with promiscuous activities[15] that accept non-natural substrates. However, these activities usually need to be further engineered experimentally in order to fulfil industrial demands on activity and stability[12]. The success of directed evolution depends heavily on the throughput of experimental screens: the more randomly generated mutants can be experimentally tested, the more likely the identification of an improved catalyst becomes. This process is thus labour and resource intensive and, despite improvements in screening throughput, the outcome of directed evolution remains unpredictable, making it a hit-and-miss approach[16]. Nevertheless, biocatalysts that substantially improve the economic viability and process sustainability of chemical synthesis have been identified in this way. This has been exemplified in the production of high-value intermediates for the pharmaceutical industry such as chiral amines, by using carbon-nitrogen bond forming enzymes such as transaminases[17] or imine reductases (IREDs)[18,19].

Here, we showcase the engineering of a biocatalytic IRED by synergising experimental and computational approaches using rapid sequence-function mapping based on ultrahigh throughput screening and machine learning (ML). First, we perform one round screening in water-in-oil emulsion microdroplets (formed in microfluidic devices)[20], where a scale-down to picolitre volumes keeps costs low (~10^{-5} cents per assay of a library member) and ramps up the speed of screening (taking only 2 weeks for the entire workflow shown in **Figure 1A**)[20,21]. This gives us access to individual improved variants (synonymous to traditional directed

evolution), but more importantly we harvest vast amounts of quantitative sequence-function data to profile the enzyme and gain insights into the importance of individual residues for catalysis, but also mutability and combinability. These experiments are designed to reveal each enzyme's idiosyncratic persona in response to mutagenesis. Basing a strategy on a comprehensive mutagenic profile that is unavailable from low throughput approaches changes the engineering paradigm from focusing on effects of individual residues to a network analysis covering the entire protein structure. The assay-labelled data generated here is selected for catalysis rather than structure[22,23] and harbours information that pure *in silico* strategies for enzyme engineering with AI[24,25] cannot access. The direction of the selection pressure towards more efficient catalysis enables us to rapidly extrapolate beyond the improvements encoded in the screened library by nearly one further order of magnitude, as demonstrated by a 23-fold improvement in $k_{cat,}$ showcasing how combining microfluidic sequence-function mapping and ML-enabled *in silico* extrapolation accelerates enzyme engineering substantially.

Droplet microfluidics rapidly provides a sequence-function map of an imine reductase via deep mutational scanning

Here we address the enzyme engineering challenge of improving the activity of a bacterial IRED from *Streptosporangium roseum* (*Sr*IRED) catalyzing the formation of carbon-nitrogen bonds via reductive amination. Picolitre sized droplets act as test tube equivalents for ultrahigh throughput screening (**Figure 1B, Supplementary Figure 1**) and contain single Poisson-distributed *E. coli* cells expressing members of an error-prone PCR (epPCR)-generated DNA library compartmentalized with lysis agent and substrates, to initiate the enzymatic reaction upon droplet formation in a microfluidic device. The reaction is stopped in a second microfluidic step by picoinjection of the dye WST-1 (to monitor the enzyme cofactor concomitant turnover from NADPH to $NADP^+$, **Figure 1C**). The resulting colorimetric readout distinguishes active library members and provides the basis for sorting by absorbance-activated droplet sorting (AADS) at 200 Hz[26] (see **Extended Data Figure 1** for a quantification of sorting efficiency by enrichment and **Supplementary Figure 2** showing selective enrichment of IRED activity).

Identities and fitnesses of the selected variants were decoded by long-read deep mutational scanning (lrDMS) (**Figure 1A**) which provides two types of information and enables decoding sequence-function maps in proteins of any size: (i) The *identity* of the variants is encoded by first generating a 'dictionary' of 29,841 unique members present before selection (**Figure 1D; Supplementary Figures 3, 4, 5; Supplementary Note 1**) with high-quality, 99.5%-accurate long-read sequencing (over the whole length of the gene)[27] , after attaching unique molecular identifiers (UMIs) to their gene sequences. After selection, this dictionary allows to decode variant identity based on the short UMI only, saving on sequencing costs and effort (one pooled NextSeq run compared to >10 Oxford Nanopore flow cells). (ii) The *fitness* of the selected variants is computed by counting the occurrences of the UMIs (giving a 'fitness score') before

and after selection (**Supplementary Figure 3 and Supplementary Note 1**; for definition of fitness **Supplementary Figure 9**): a calibration curve (relating the conversion rates of variants with UMI counts; correlation coefficient r = 0.92) suggests that fitness scores capture quantitative fitness information of variants across a wide range of activities (**Figure 1E**).

We performed IrDMS of *Sr*IRED by sorting approximately 1.5% of the most active variants (of 6.9 million droplets screened at 200 Hz, 57.5-fold oversampling of the library), revealing 10,905 unique members to be active (with quantitative activities as UMI-based fitness scores; **Figure 1C, Supplementary Table 1**) and, conversely, 6,238 variants to be inactive (occurring in the input library but not in the output library, i.e. below the selection threshold; **Figure 1D).** This workflow thus unlocks quantification of the fitness for >10^4 randomly selected single and higher-order variants within 2 weeks: a scale, speed and cost not feasible with plate screening or robotic workflows[20].

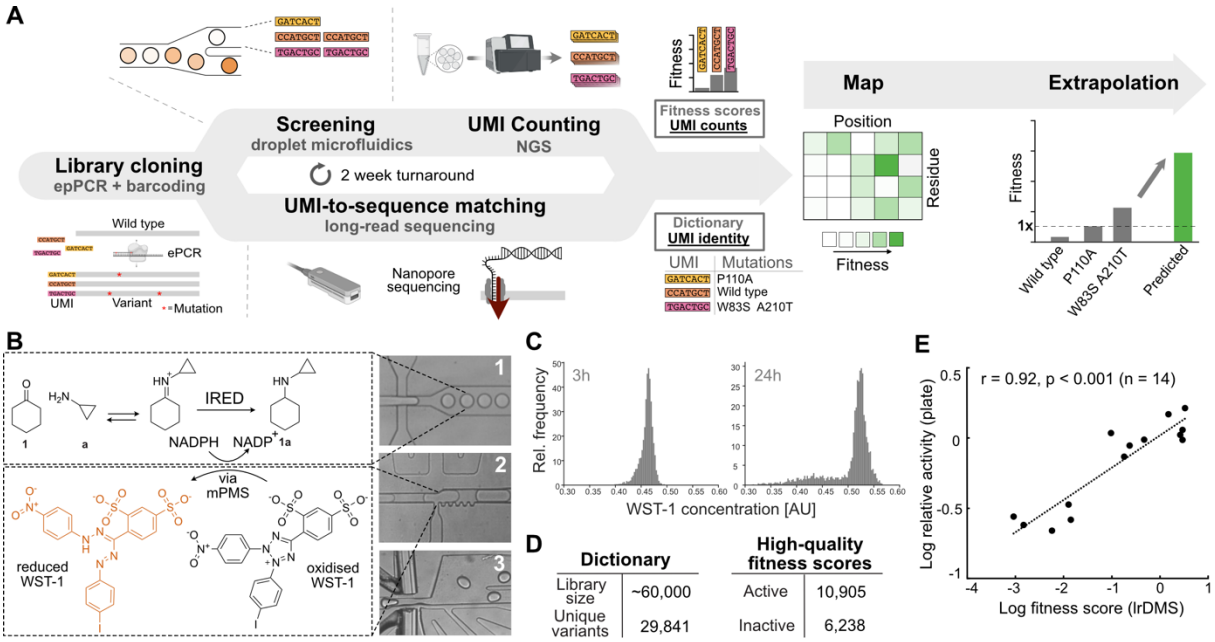


Figure 1: A microfluidics-enabled workflow for rapid large-scale fitness data generation and extrapolation **(A)** Long read deep mutational scanning workflow. A random mutagenesis library, barcoded with unique molecular identifiers (UMIs), is sequenced using nanopore sequencing to give a dictionary that identifies each library member by linking its UMI sequence to the corresponding variant sequence. The library is screened using a droplet-based uHTS (ultrahigh-throughput screening) assay in multiple replicates. UMIs of the sorted and unsorted library are sequenced using NGS and fitness scores are calculated from the enrichment of individual UMIs upon screening. The fitness information and the identity of the UMIs are combined to give an accurate and quantitative sequence-function map, which serves as the basis for rational and AI driven extrapolation. **(B)** Ultra-high throughput assay. Substrates (cyclohexanone **1** and cyclopropylamine **a**) are encapsulated with NADPH, *E. coli* expressing the IRED library and lysis agent. After incubation, mPMS and the formazane dye WST-1 are added by picoinjection, producing the colorimetric readout that is used as the criterion of selection by AADS reporting on activity. (**C**) Activity distribution of droplets containing the library after 3h and overnight incubation. The positive population (low absorbance due to consumption of NADPH) increases as the reaction progresses in individual droplets. **(D**) Number of dictionary members and fitness scores obtained (a total of 17,143 IrDMS scores are calculated from 29,841 unique dictionary members). (**E**) A linear correlation of fitness scores with colorimetric lysate assay data for 14

individual variants chosen randomly from the input library and the library after sorting suggests that the picoliter assay in droplets quantitatively reports on genuine activity levels. *E. coli* lysate for each variant was incubated with **1** (10 mM), **a** (20 mM) and NADPH (5 mM) for 3 h or overnight in Tris buffer (100 mM, pH 8.0). The colorimetric readout was produced by the addition of the formazan dye INT (a less water-soluble alternative to WST-1). A fitness metric for lysate activity was calculated as the logarithm of the ratio of the average variant endpoint to the average wild type endpoint. [*Abbreviations* – lrDMS: long read deep mutational scanning; epPCR: error-prone PCR; UMI: unique molecular identifier; NGS: next-generation sequencing; IRED: imine reductase; AADS: absorbance-activated droplet sorting; uHTS: ultrahigh-throughput screening]

Long-read deep mutational scanning reveals the mutational profile of an IRED

Experimental lrDMS derived fitness scores for 17,143 variants provide us with information on how mutations across the *Sr*IRED sequence and structure affect activity as single mutations (*mutability*) and whether they can be combined in higher-order variants (*combinability*). On average, mutational effects are detrimental to fitness, with a median fitness of -1.2 (0.3-fold wild type activity), and the total number of improving mutations is low with 0.9% of mutations showing significantly improved fitness (1 SD > wt). 6,283 sequences (37% of total quantified variants) are classified as inactive as they occur in the input library but never in the output after selection (**Figure 2A**). We calculated the per-position median fitness of all single variant fitness scores in the IRED sequence (on average 4.3 datapoints per position) as a proxy for mutability, i.e. the capacity of a site to be functional after mutation (**Figure 2C**). First shell and the proposed catalytic residues[28,29] show very low mutability with the exception of T241 that contacts the ketone substrate and is positioned at the bottom of the binding pocket, flanked by all three catalytic residues (**Figure 2D and E**). Positions with high mutability are defined as hotspots where clusters in specific regions of the IRED sequence (e.g. around position 40) or structure were observed. Mapping mutability onto a crystal structure of *Sr*IRED[29](**Figure 2E**) identifies hotspots especially at the N-terminus (**Extended Data Figure 3B**) and in helices and loops of the N-terminal domain **(Extended Data Figure 3C and D**). We also observed a pattern of repeating hotspots in the central helices at the dimer interface (**Extended Data Figure 3E and F**), consistent with previous IRED engineering campaigns that selected mutations in this region [19,30].

To characterize the enzymatic properties of hits, Michaelis-Menten kinetics of 12 single variants with fitness scores >0 and the variant with the top fitness in the library (V6L V67I) were measured (**Extended Data Figure 4, Supplementary Figure 6, Supplementary Table 2**). Overall, single mutations at position T241 lead to the highest catalytic efficiency (4-fold improved k_{cat}/K_M along with a 3-fold improved k_{cat} in T241A and T241S). 8 of 12 variants showed an improved k_{cat} while only the T241 mutants and E146V showed improved K_M values. As in previous DMS datasets[31], along with catalysis, thermostability was improved for 6 (of 12) variants (up to a T_m of 50.0 °C for T241A *vs* 38.5 °C for wild type), reflecting the need for long

term integrity of the catalyst for the entire duration of the biotransformation (here up to 24h)[32]. Indeed, under conditions relevant for larger scale synthesis (i.e. catalytic amounts of NADPH and a cofactor-recycling system consisting of formate dehydrogenase from *Candida boidinii* (FDH-QRN) and formate [33,34]; **Extended Data Figure 5E**), product formation of T241A was markedly improved: T241A improved across most tested ketone and amine substrates, e.g. conversion to **1a** from 83% to 97% (after 4h) and to **4a** from 41% to 67% (after 24h) (with exception of **1b**; **Extended Data Figure 5F, Supplementary Figure 8**). By contrast, other mutants are strictly specialized to one co-substrate combination (K38E and D69H for **1a** **(Extended Data Figure 5D, Supplementary Figure 7**). These mutations are found outside the first shell, confirming the notion that substrate specificity can be engineered globally[35]. Generally, excellent correlations between specific activities (based on initial rates of NADPH depletion) and the conversions observed in biotransformations suggest that the measurement of product in our assay reliably elicited catalysts with superior biotransformation capacity.

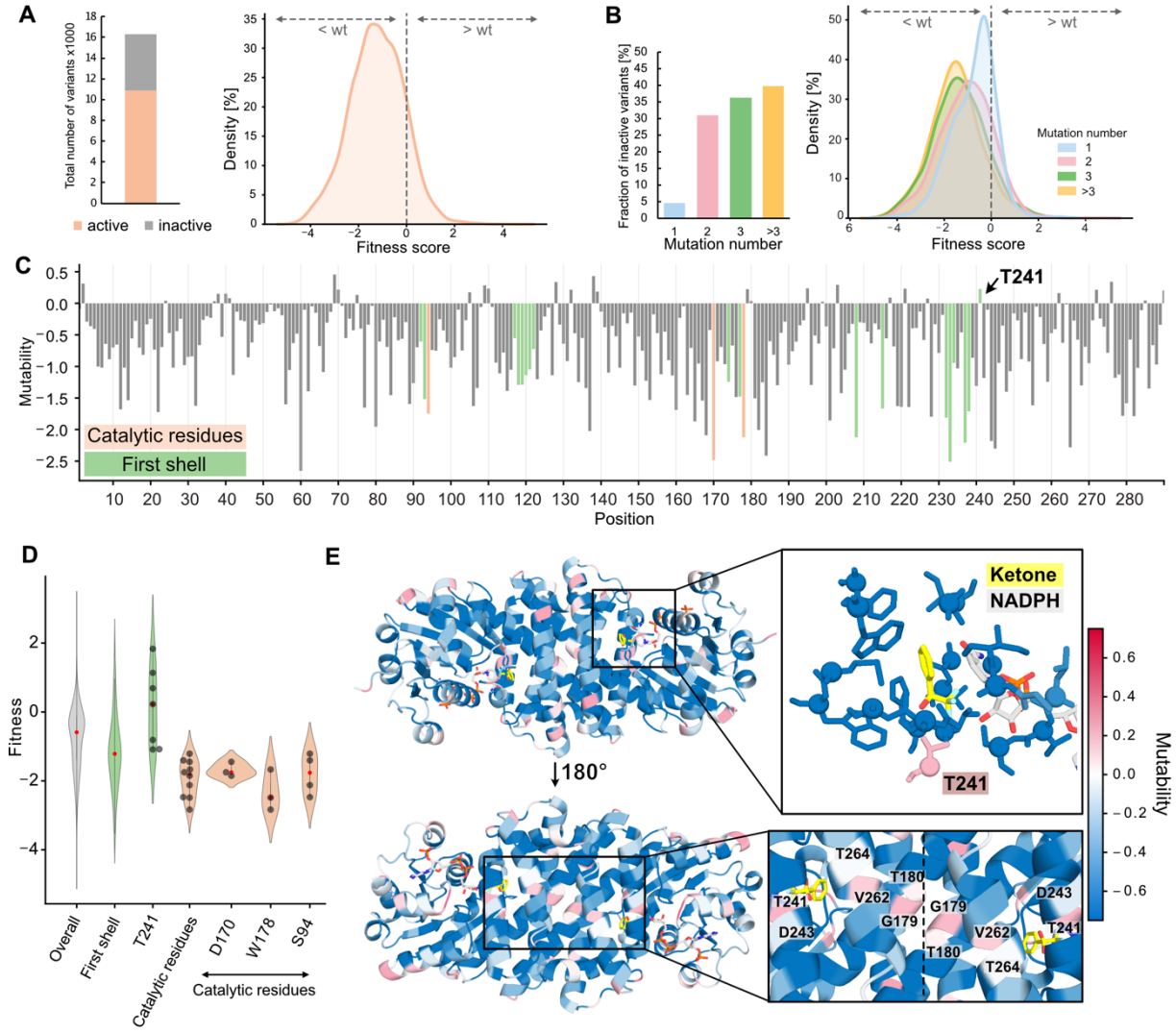


Figure 2: Distribution of mutational effects in an error-prone PCR library (**A**) 63% of all assessed variants show detectable activity. Distribution of fitness effects of all mutations with detectable activity shows, that most variants are less fit than the wild type enzyme, with a small fraction of mutations showing an improving effect (**B**)

Distribution of fitness effects for single (blue), double (pink), triple (green) and more than triple (yellow) mutations shows, that the mutational effect is dependent on the number of mutations introduced. The more mutations are introduced the more abundant are inactive variants (left) and the more negative the average fitness of functional variants. (**C**) Positional distribution of median fitness shows that there is a large difference in the mutability of different positions in the protein (orange: catalytic residues; green: first shell). (**D**) Violin plot showing the fitness of variants with different mutational backgrounds. Mutants of first shell and active site residues are on average less fit than the average mutations (-1.21 vs -1.82). However, there is one exemption with positive mutability and three highly fit mutations: T241. No mutation at any active site position (D170, W178, S94) shows positive activity. (**E**) Structural map on median fitness shows distribution of mutability on structure (*blue*: low mutability, *red*: high mutability). The first shell and the dimer interface are highlighted.

Access to the entire protein by lrDMS, rather than fragments obtained in Illumina sequencing, gives access to the joint effects of multiple mutations and their combinability (i.e. the potential of a mutation for constructive interaction; see **Supplementary Figure 9** for definition) which can be assumed to be a major contributor to evolvability (i.e. the capacity of a mutant to originate improving trajectories in evolution). Our lrDMS data reveals that, along with an increasing number of mutations per variant, a higher fraction becomes inactive and shows on average decreasing fitness (**Figure 2B**). Among the 11,958 available datapoints for higher-order variant, we used 6,188 with measurable activity and available single mutation data for all mutations in the higher-order variant to quantify epistasis by comparing their fitness to the expected additive fitness calculated from single-point mutation data alone **(Figure 3A, Supplementary Figure 9)**. As it can be assumed that there is no positive epistasis in the residual 4,130 inactive variants, the overall fraction of positive epistasis is lower than that for negative epistasis, in line with previous observations[36,37]. However, a large fraction of improved variants (67 of 84, 80%) are positively epistatic (**Figure 3A and B**) and 60 out of 67 (90%) of these contain single mutations that on their own are neutral or negative: they would have been lost in screening campaigns that iterate combinatorial screening of improving variants from single site saturation libraries.

Having established that positive epistasis plays a major role in building constructive trajectories in *Sr*IRED evolution we next investigated the sequence background of combinability. Among the 82 higher-order mutants shown to be highly active with no negative epistatic interactions, one highly combinable mutation with high potential for constructive interaction stands out: the global hotspot at position 241 (in the first shell, **Figure 2E, Figure 3C, and Supplementary Figure 10**; a further combinability hotspot is evaluated in **Supplementary Figure 11)** with contribution mostly from T241A, the mutation previously identified with the highest improvement in catalytic efficiency (4-fold) and 11°C improvement in T_M. The central role of T241 is evident: it shows combinable interactions with positions distributed broadly over the N-terminal and C-terminal domains (**Figure 3D** and **Extended Data Figure 6, Supplementary Figure 10, Supplementary Table 3**). Elevated protein stability may explain the high

combinability of T241A in cell-based assays as for example our droplet assay, by enhancing mutational robustness, extending the stability threshold of the protein and allowing incorporation of destabilizing mutations[38-40]. However, when characterizing positive epistatic mutations with T241A and the distant mutations K38E, D69D, or P203A (see next section), we also observed positive epistasis in catalytic parameters (k_{cat}/K_M as well as k_{cat}). Indeed, the destabilizing variant T193A (**Extended Data Figure 6**), for instance, is part of a larger cluster around a hinge region in the C-terminal domain indicating a shared mechanism of action (**Figure 3E and F**). This is consistent with models where combinability is not only influenced by mutational robustness but also protein dynamics where distant mutations can yield synergistic effects by optimizing the conformational landscape of the enzyme[41,42].

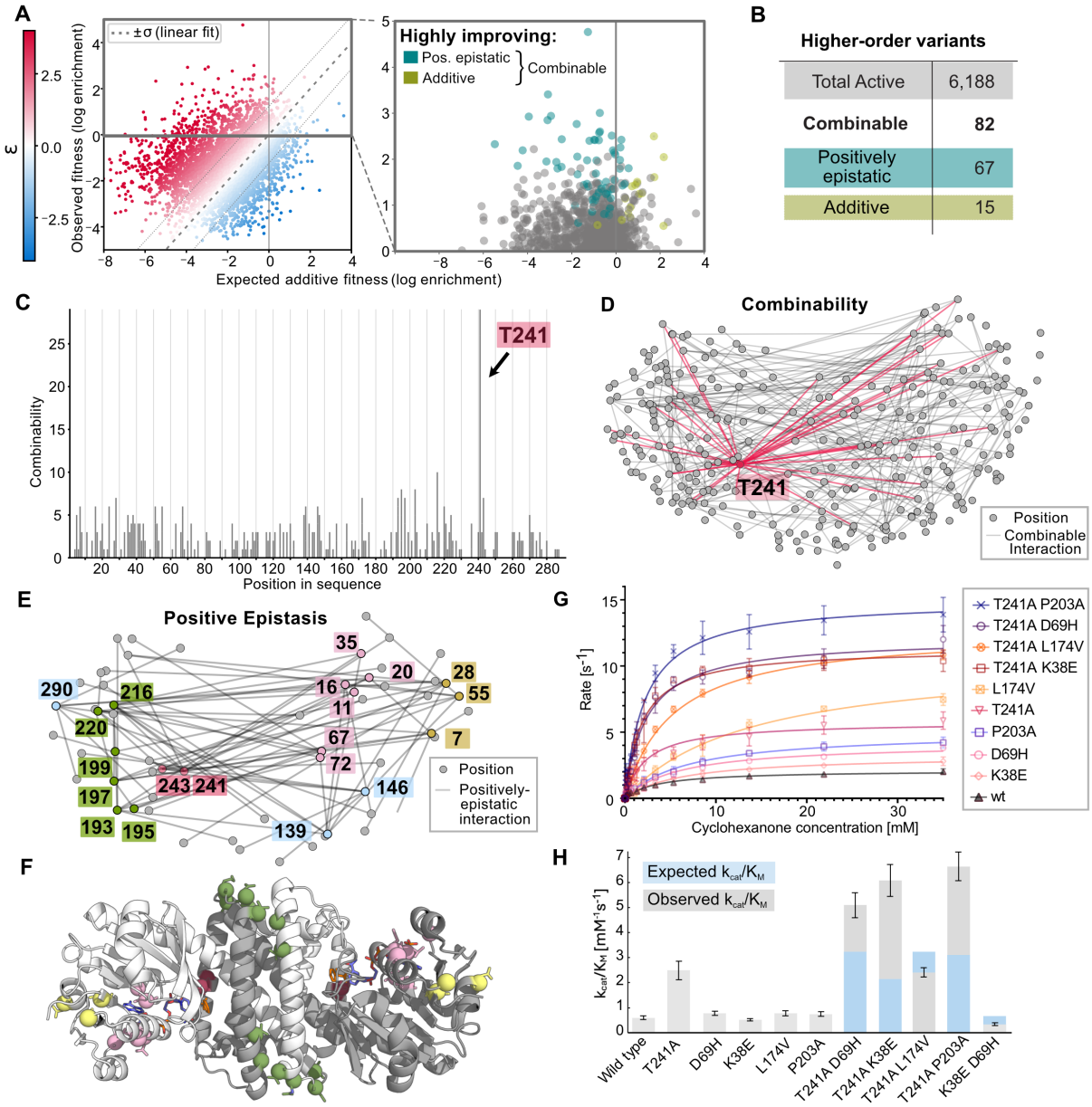


Figure 3: Combinability distribution informs rational engineering. (A) Scatter plot showing expected fitness for all higher-order variants from single mutation data (calculated via the product model) versus observed fitness. Positive epistasis is shown in red, negative epistasis in blue. On the right, variants with larger than wild type fitness

(observed fitness > 0) are shown. Positively epistatic (1SD > additive) and additive (1 SD > negative epistasis) variants with high fitness (1 SD > wt) are shown in red and purple, respectively. **(B)** Of 6188 higher-order mutations in this dataset (2-5 mutations with single mutation data for all mutations), 84 are highly active (1SD > wt) and 82 are highly active and combinable. Most of these variants are positively epistatic (67) and only 15 are additive. **(C)** Positional frequency in highly active and combinable variants. T241 is the global hotspots of combinability, followed by V216. **(D)** A combinability graph shows combinable interactions of mutations at distinct positions. Nodes represent positions in the IRED sequence that are grouped by distance in IRED structure. Connections represent cooccurrence in a higher-order variant the corresponding fitness of which is larger or equal to the combined effect of single mutations: combinable. The high node degree of T241 (edges in red) indicates high potential for combinability and positive epistasis with all parts of the protein. **(E)** A simplified graph only including positive epistatic interactions grouped by distance of nodes in the structure shows structural grouping of positive epistatic hotspots. Positions with larger than 2 positive epistatic interactions are colored (light blue, green, purple, pink and yellow). **(F)** Epistatic hotspots are marked on the *Sr*IRED crystal structure (5ocm). **(G)** Michaelis Menten kinetics with single variants chosen for rational engineering with T241A and their combinations showing high performance of combined variants apart from combinations with L174V and indicates general combinability of T241A. **(H)** Comparing expected k_{cat}/K_M of single variants (blue) to the observed k_{cat}/K_M (grey) of rationally engineered higher order variants shows positive epistasis for all rationally engineered variants apart from T241A L174V. Standard errors of the fit and three technical replicates are represented with a 95% confidence interval.

Rational and AI-informed extrapolation yields highly improved positively epistatic variants

Directed evolution relies on sampling and testing the vast combinatorial diversity of sequence space by iterative fixation of mutations, thereby building trajectories to higher fitness. Rational engineering aims to avoid these sometimes long and laborious screening campaigns but is hampered by negative epistasis and often misses out on large leaps in activity provided by positive epistasis even when information on improving single mutations is available[10]. Hypothesizing that lrDMS data can provide a window into future rounds of directed evolution we use combinability and mutability information from a single lrDMS round for rational and AI-informed extrapolation to higher-order variants. These higher-order variants have not been observed in the dataset and would otherwise only be accessible by conducting multiple rounds of directed evolution.

(i) Rational engineering: lrDMS-derived combinability data provides a handbook on how to bypass negative and harness positive epistasis during rational engineering for the entire length of typical proteins, where common NGS sequencing fails to resolve long-range interactions (>200 AA). Building on these data, we implemented our combinability hotspot T241A with improving mutations from hotspots distributed all over the protein. We chose P203A, K38E and D69H from hotspots at the dimer interface and at loops interacting with the cofactor, and L174V, the second-best mutation in the first shell (**Figure 3G and H, Extended Data Figure 6**). Three out of four variants show positive epistasis with the highest improvement being 11-fold for T241A P203A. We hypothesize that the unsuccessful combination T241A L174V may

be related to similar, redundant mechanisms of improvement of L174V and T241A that can be rationalized by their proximity in the active site. The non-triviality of this task is illustrated when combining mutations not previously identified as combinability hotspots such as K38E and D69H. The double mutation K38E D69H shows strong negative epistasis with a 2-fold drop in k_{cat}/K_M compared to wild type, showing the relevance of our lrDMS combinability data for successful rational engineering of higher-order variants (**Figure 3H**).

(ii) AI-informed engineering for single mutants: As not all amino acid changes are accessible by single nucleotide changes (in the case of *Sr*IRED only 30% of all single amino acid mutations; **Extended Data Figure 2**), error-prone PCR libraries are biased to a limited chemical space. We use machine learning to extrapolate to the full chemical space. Specifically, we build a ridge regression model based on physiochemical amino acid embeddings, computed from a principal component analysis of the AAIndex database[43,44], and augment it with the ESM2[45] pseudo-log-likelihood ratio of a given mutant sequence vs wildtype, as proposed in [46] (**Figure 4A**). We train this model on the observed single mutants from the lrDMS campaign and extrapolate fitness values for all unobserved single mutations (**Supplementary Figure 12**). We selected the 5 top-ranked predictions P203G, T241G, M1P, A251I and E146L (omitting predictions at the same positions to avoid position bias) for experimental characterization (**Figure 4C and D**, **Extended data figure 7A)**. Of these 5 variants, 1 showed an improved k_{cat}/K_M, 3 an improved k_{cat}, 1 improved activity in lysate and 2 out of 5 variants show an improved melting temperature (improvement > 1SD from wild type). The best predicted variant, T241G, shows a significantly higher improvement in k_{cat}/K_M than all characterized variants in the lrDMS dataset (7-fold). Notably, T241G also increases T_M to 45.8°C, up from 38.5°C for wild type but below the 50.0°C for T241A. The learned and extrapolated systematic decrease in side-chain volume at position 241 (threonine in the wild type, alanine and serine in the dataset, glycine in the AI prediction), illustrates the potential of AI-based extrapolation to provide improved variants that could not have been found in the experimental screening data. For T241G to be included in the dataset, two consecutive nucleotide substitutions would have been necessary to reach G (GGN) from T (ACG), a highly unlikely event in epPCR mutagenesis. Learning curves and predictions without ESM (**Extended data figure 8**) show that the success of the model can be primarily attributed to the experimental dataset: Mutant rankings without using any information from ESM2 still include 3/5 mutations in the top 5, including T241G (**Supplementary Figure 12**). Conversely, a mutant ranking based on ESM2 alone does not find T241G until rank 457/5510 and only proposes top 5 variants with up to 2-fold improved k_{cat}/K_M in experimental characterisation, albeit with no variant smaller than wild type performance (**Supplementary Figure 12 and 14)**. While these results indicate that information from ESM2 can serve as important feature, or act

as a filter to eliminate poorly performing variants, they also show it cannot substitute experimental data in the context of the relevant reaction and highlight how the synergy of experimental data and data-driven techniques unlocks improvements unattainable by either technique in isolation.

(iii) AI-informed engineering for double mutants: Predictive models for higher-order variants must take the combinability of mutations into account to circumvent negative and leverage positive epistasis and are thus dependent on training data that captures long-range interactions of mutations. We developed a tree-based model to automate rational engineering of higher order mutants (**Figure 4B**), focusing on double mutants. The model is trained on lrDMS fitness scores and has access to information on the protein structure (distance between mutated residues, distance to active site residues, dimer interface and first shell, flexibility information computed from an elastic network model[47]). Like in the single mutant model, evolutionary information is incorporated via the protein language model ESM2. Predictions were ranked (**Supplementary Figure 12**) and the 5 best variants were tested experimentally (**Figure 4E and F**, **Extended Data Figure 7C and D)**. We found that 4 out of the 5 tested variants showed considerably higher k_{cat}/K_M, lysate activity and T_M than the wild type (7-fold improvement in k_{cat}/K_M for E36A-T241A). When comparing k_{cat}/K_M values of the double mutants with respective single mutations, we observed positive epistasis in 2 out of 5 variants, additivity in 2 out of 5 variants, and negative epistasis in 1 variant, showing that the model was able to predict combinability in 4 out of 5 cases (**Figure 4F, Supplementary Figure 13**). The top predictions emerging from this model appear to focus on a loop involved in binding the NADPH phosphate group (**Figure 4J),** consistent with a reorientation of the cofactor that, in combination with mutation T241A, enables a more favorable substrate binding mode that leads to the observed synergistic effects. The successful prediction of multi-residue cooperativity indicates the ability of the model to learn complex structural patterns as the basis for higher-order mutants. This can be primarily attributed to the experimental data. Learning curves (**Extended data figure 9**) show that large amounts of data are needed (jump in performance at 3,072 datapoints) and that addition of ESM on top of assay features has minor influence on model performance (as evidenced by ablating ESM features, **Supplementary Figure 12**). Furthermore, a SHAP analysis[48] (**Extended data figure 10**) identifies assay labeled data as the core feature of the prediction, an observation which is also supported by direct analysis of the learned decision trees. Finally, experimental characterisation of the top 5 purely ESM predicted double mutants shows up to 3-fold improved variants, not rivalling the 7-fold improvement observed when assay labelled data is included (**Supplementary Figure 12 and 14)**.

(iv) Combination of AI and rational engineering: Access to even higher activity improvements was possible by combining information about the combinability hotspot T241 from lrDMS (that led to rational engineering of T241A P203A, so far the best variant) with the AI-predicted extrapolation from T241A to T241G, which significantly boosted the single mutant activity. Hypothesizing that the combinability of T241A would be transferable, we replaced T241A with G in all double mutations from rational engineering and the best AI double mutant. In 4 out of 5 cases k_{cat}/K_M levels were clearly improved (**Figure 4G and H**). With a 16-fold improved k_{cat}/K_M, driven by a 23-fold improvement in k_{cat}, T241G P203A overcomes all previous activity improvements and is the best variant generated in this study. This illustrates how combinability information with the single mutation AI model can boost improvements to levels exceeding both AI and rational engineering alone (**Figure 4I**).

(v) Application in biotransformations: While is it crucial to monitor kinetic parameters during enzyme evolution to dissect and steer mutagenesis and selection parameters, the ultimate challenge for biocatalysts is their application under conditions relevant for larger scale synthesis. Thus, the six best-performing variants from rational and AI-informed engineering were tested besides the variants T241A and D69H as well as wild type in biotransformations for the reductive amination of cyclohexanone **1** with cyclopropylamine **a** as amine donor (**Extended Data Figure 5F and Supplementary Figure 15**). In agreement with the enhanced catalytic efficiency, all variants showed increased catalytic performance in biotransformations. Variants T241G, T241G P203A and E36A T241G performed best and yielded more than 40% conversion after 30 min. In contrast, the wild type gave approximately half of the conversion at the same time. To evaluate how stereoselectivity changed during evolution, the same variants were assayed for the reductive amination of the chiral ketone (*R*)-3-methylcyclohexanone (**5**) to install a second stereocenter using cyclopropylamine (**a**) as the amine donor. While the wild type enzyme yielded a diastereomeric ratio of 30:70, T241A as well as all six rational and AI-engineered variants exhibited an improved stereoselectivity ranging from 15:85 to 10:90 d.r (**Extended Data Figure 5G and Supplementary Figure 16**), indicating that the mutation of the residue T241 with an alanine or a glycine is primarily responsible for the increased stereoselectivity in the reductive amination of **5** with **a**. In accordance with catalytic efficiency measurements, the best-performing variant in terms of conversion and stereoselectivity was the variant T241G P203A, engineered via a rational and AI-informed hybrid approach (90% (10:90 d.r.) conversion for T241G P203A, up from 47% (30:70 d.r.) for the wild type and 71% (15:85 d.r.) for the best performing mutant in the dataset, T241A).

In summary, we show that by deriving mutability and combinability information via droplet microfluidics and lrDMS, we define the mutational profile of a biocatalyst: improved hits

(T241A, 4-fold improved k_{cat}/K_M) discovered in experimental screens are amenable to rapid extrapolation accelerated by AI (based on the data harvested in the screen) to give biocatalysts with substantially improved catalytic properties (T241G P203A,16-fold improved k_{cat}/K_M).

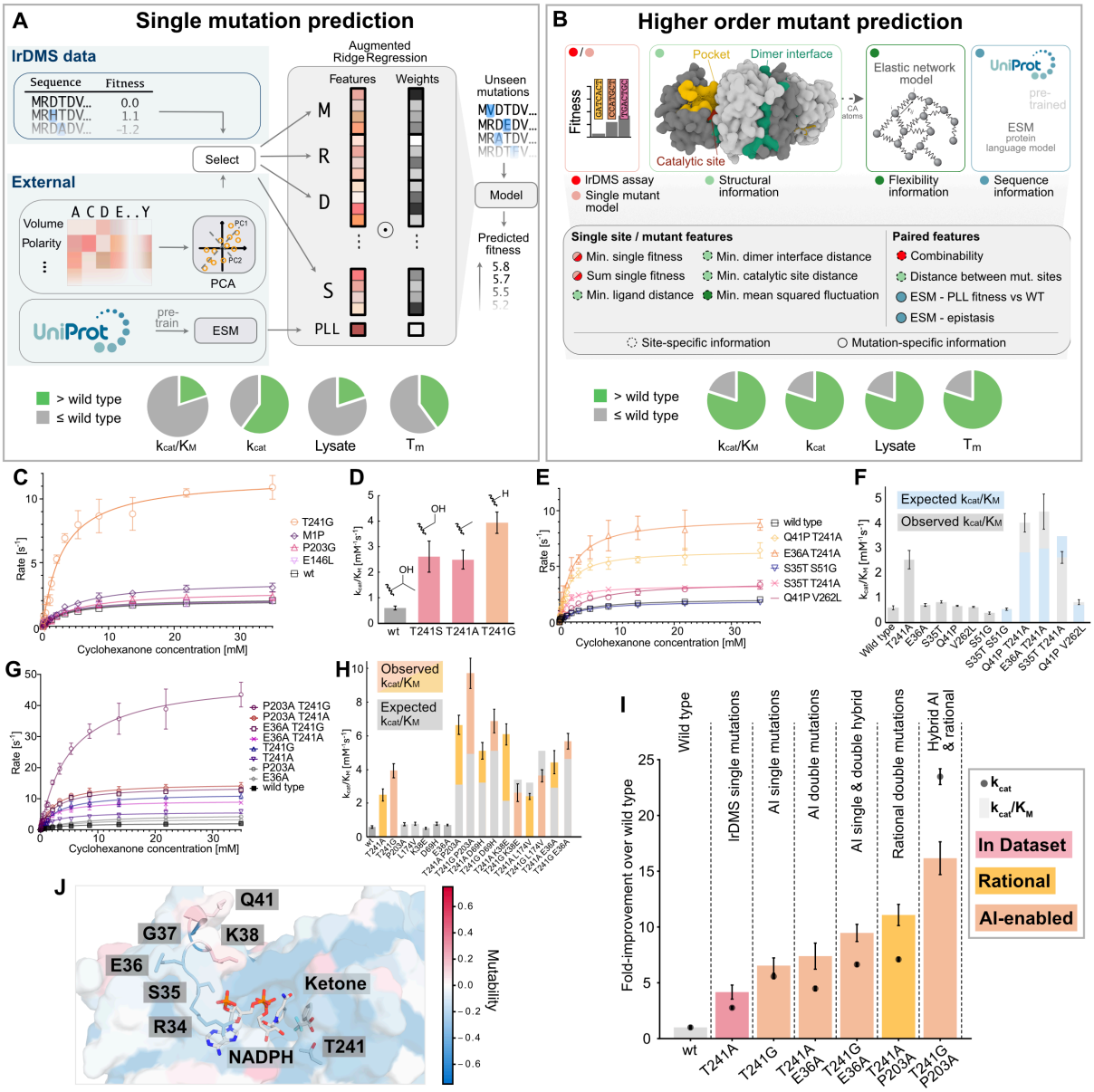


Figure 4: AI and rational engineering enable extrapolation to highly active variants (A) Single mutant model and its performance in predicting improved mutations. IrDMS data and external parameters (ESM pseudo-log likelihood vs wildtype (PLL) and amino acid structure embeddings) are used in an augmented ridge regression model to predict the fitness of single point mutations. The 5 best single point mutations are characterized experimentally (pie chart). 2 out of 5 characterized mutations show an improved k_{cat}/K_M (> 1 SD from wild type), 3 out of 5 an improved kcat, 1 out of 5 improved lysate activity and 4 out of 5 show an improved melting temperature. (**B**) The double mutant model is based on assay-derived, structural, flexibility and sequence information. Gradient boosted decision trees trained on the IrDMS data provide a predicted fitness estimate for unobserved double-mutants, which is used to rank and select mutants to test. Among the top 5 characterised mutants from the model 4 out of 5 show improved k_{cat}, k_{cat}/K_M and lysate activity. (**C**) Michaelis Menten curves of single point variants predicted by AI showing slight improvements for M1P and E146L and a 6.6-fold improvement for T241G. Standard errors of the fit and three technical replicates are represented with a 95% confidence interval. (**D**) k_{cat}/K_M values of

T241A mutants (grey: wild type, pink: in dataset, orange: AI-predicted) with side-chain structures highlighting the decrease in steric bulk from threonine to serine, alanine and glycine (**E**) Michaelis Menten curves of variants derived from the combinability-based model showing improving parameters for 4 out of 5 variants (**F**) The k_{cat}/K_M of the top 5 predictions of the higher-order mutant model and corresponding single point mutants (Expected k_{cat}/K_M calculated from single mutation data is shown in blue, observed parameters in grey). 2 out of 5 variants show positive epistasis, one negative epistasis and two additivity. The best AI-predicted double mutant shows a 7.4-fold improved k_{cat}/K_M, better than the best predicted AI single mutant T241G with a 6.6-fold improvement. (**G**) Michaelis Menten plots and (**H**) k_{cat}/K_M values for hybrid AI and rational engineering-based strategy where T241G (orange) is transplanted into rationally engineered and AI-predicted variants with T241A (pink). Replacing T241A with T241G has a positive effect in 3 out of 4 cases. A 16-fold improved k_{cat}/K_M paired with strong positive epistasis was observed with T241G P203A. The negative epistasis with L174V is conserved but paired with a jump in improvement from 4-fold to 6-fold. Replacing T241A with T241G switches the positive epistatic effect with K38E into a strong negative epistatic effect underlining the influence small residue changes can have on non-linear effects on activity. Replacing T241A with T241G has a similar effect on AI-generated double mutants transforming a 7.4-fold improvement with E36A T241A into a 9.5-fold improvement with E36A T241G. (**I**) k_{cat} (points) and k_{cat}/K_M (bars) data for top variants derived from lrDMS data (T241A), the AI single mutation model (T241G), the AI double mutation model (T241A E36A), rational engineering based on both AI models (T241G E36A, rational engineering based on single mutation and combinability data (T241A P203A) and a hybrid approach based on AI single mutation and combinability data (T241G P203A). (**J**) Structure of the loop containing residues predicted by the AI model. K38 has been shown to be positively epistatic with T241A in rational engineering, E36A and Q41P show a similar effect.

Discussion

Generating sequence-function maps in merely two weeks enables a radically different approach to enzyme engineering. Instead of simply isolating one improved mutant in traditional directed evolution, large datasets are generated that profile the idiosyncrasies of a protein in a way that eludes enzyme engineering based on single site-directed mutagenesis. We obtain a transparent picture of the mutational potential of a protein and its ability to adapt to an improved function. The complex interplay of biophysical properties and biochemical reactivity that has made enzymes harder to engineer than binders is illustrated by insights into (i) hotspots vs regions that should not be touched by engineering (from single mutations), described by mutability, (ii) cooperative vs detrimental epistatic interactions across the entire protein (from higher-order mutations), described by combinability and (iii) a sense of the tolerable mutational burden (for maintaining stability while introducing catalytically beneficial mutations). Previously, these complex interactions were seen as hard to dissect, but algorithmic help from ML allowed for the extraction of useful, interpretable information from sequence-function datasets, as shown by enhancing the activity of our biocatalyst beyond the improvements achieved by hits isolated from the experimental screen.

Several key features of our approach will make this a general method for enzyme engineering: (i) The use of droplet microfluidics is crucial to assay large libraries to produce datasets with > 10^4 variants, a feat which would be prohibitively expensive and time intensive in plate

screening or robotic workflows[49-51]. The capital expenditure and the costs of reagents limit feasibility of obtaining similar throughput by these well-established methods. Droplet microfluidics is fast, affordable (even for public sector labs) and scalable, with the throughput depending on the detection method (AADS ~1 Hz[52]; FADS: >5 kHz or FACS with double emulsion droplets: >10 kHz)[53]. While miniaturizing assays for use in droplet format requires an optical readout, assays for all 7 EC classes are already available[20], and coupled assays[26,54,55] or turnover detection with product sensors[56] cover reactions for which chromo- or fluorogenic substrates are lacking.

(ii) Targeted mutagenesis (based on closeness to the active site[57] or phylogeny[58] is an option, but the entire enzyme is better profiled by epPCR which is cheap, compared to designed libraries that need to be synthesized, and hypothesis-free. Mutations are unbiased by position, the order of mutation or role in the mechanism and the catalytic improvements observed here were brought about by mutations that would not have been targets of a design-led approach. Indeed, our observed 23-fold rate (k_{cat}) improvement was an order of magnitude higher than the rate improvements (1.3-fold) achieved by a previous, targeted, AI-informed IRED engineering campaign based on single mutation data[49]. Natural evolution is also based on mutations in random positions, so this approach is essentially biomimetic and empirical, making hypotheses superfluous.

(iii) The crux of enzyme engineering is the complexity of epistatic interactions[10] and these are not fully captured by short-read Illumina-based approaches[59] (limited to 600 bps (i.e. ~200 AA) with remedial stitching read approaches poorly scalable[60,61], unless the proteins are unusually small. Most bacterial and eucaryotic proteins are longer (on average 320-472 amino acids[62]), so capturing long-range epistasis on this scale requires the lrDMS sequencing workflow outlined here. A scalable approach of this sort has been employed successfully for a binding binding domain[63], but not for an enzyme. Elucidating long-range epistatic interactions requires sequence-function datasets containing full-length gene sequences (see **Supplementary Table 4** for an overview of enzyme DMS studies highlighting the limited availability of quantitative sequence-function datasets over the whole gene length), while ignoring this crucial information will make datasets biased and less useful for AI interpretation.

(iv) Finally, high hit-rates are crucial in typical enzyme engineering campaigns to reduce time and cost of experimentation. The high success rate of predictions of catalytic improvements in our work (>80% for k_{cat}, **Extended Data Figures 6 and 7**) means that very few mutants must be experimentally tested. This compares favorably with ML-based engineering approaches in the low or zero input data regime with hit rates of 2-3%[22,64] and design approaches where testing of thousands of mutants was necessary to identify three functional catalysts[65]. We thus unlock significant time and cost savings for enzyme engineering by identification of significantly

improved hits among only a handful of designed sequences after only a single round of screening in microdroplets.

Beyond achieving specific improvements of this IRED, the ability to quickly and cheaply generate datasets that enable AI-based extrapolation for hard-to-predict phenotypes like enzymatic activity addresses a key unmet need for AI assisted protein engineering. Structure prediction by AlphaFold[66] and design of protein backbones[9] are now feasible and routinely used and benefitted from the availability of large-scale structure datasets (e.g. the pdb), while protein function can still not be predicted with similar ease for lack of training data. For data-driven models, the input datasets (and the way they were generated) matter significantly for the outcome. Previously enzyme stability was enhanced and appeared to be the main driver, e.g. in zero-shot AI approaches based e.g. on PDB data [22,23]. However, catalysis is a function of biophysical properties *and* biochemical reactivity. Therefore, particularly with respect to the latter, a dataset based on an assessment of catalysis (i.e. turnover assay-labelled[67]) is crucial for a comprehensive description of an enzyme. When all assay labels relate to function, ML interpretation should reveal a clearer 'response' to trigger improvements in all properties, including biocatalytic turnover.
Starting with a catalytic selection in our system, it will be interesting to follow up with another microscale assay system, HT-MEK, that is able to analyse fewer variants (low thousands rather than 10-100-fold more in microfluidic droplets) but provides disaggregated, high accuracy biophysical characterisation rather than a summary 'fitness' assessment in droplet screening. Using these systems in succession may gradually increase the functional information in datasets leading to further improved outcomes of AI extrapolations and potentially allows better insights into the mechanistic contributions.
The combination of a high throughput assay and AI is important as unbiased approaches are data hungry: most published AI models utilizing experimental data use focused libraires as the smaller sequence space of possible variants allows achieving higher data depth, which leads to improved model extrapolation. Indeed, analysing the performance of our models with different amounts of data shows that large amounts of data are needed (**Extended Data Figure 8 and 9**) and that even larger amounts of data would continue to substantially improve the models, which can only be achieved by the scale that droplet microfluidics offers.

It is tempting to speculate whether the characterisation of individual proteins via sequence-function datasets will provide a growing empirical body that allows further meta-analysis – perhaps revealing 'types' of enzymes, where complex unifying characteristics are revealed by ML analysis. This large-scale molecular information would in turn trigger testable hypotheses that should reinvigorate mechanistic studies and, apart from being useful, bring us closer to a

fundamental understanding of mechanisms of enzyme catalysis and evolution in a language that describe enzyme catalysis in its enormous complexity that currently eludes us.

Data and code availability

Raw sequencing data will be made available via the European Nucleotide Archive. Fitness scores will be attached as a txt file to the supplementary data. Machine learning models and custom python code for analysis are available via https://github.com/Hollfelder-Lab/lrDMS-IRED.

Acknowledgments

M.G. holds a Trinity College Benn W Levy studentship, S.V.M. received funding from the Vice-Chancellor's award from Cambridge Trust and from the UKRI Centre for Doctoral Training in Application of Artificial Intelligence to the Study of Environmental Risks (EP/S022961/1). The European Union Marie-Curie networks 'Oligomed' (956070) and ES-Cat (722610) provided support for F.E.H.N. and P.Z. respectively. F.M.W. holds scholarships of the Cambridge Trust, Stiftung der deutschen Wirtschaft (sdw) and Evonik Stiftung. E.J. M. holds scholarships of the BBSRC (BB/M011194/1) and O.J.K of the EPSRC (EP/R513180/1). F.G.M. and T.K thank NWO Sector Plan for Chemistry and Physics and ERC StG (638271) for funding. F.H. was an ERC Advanced Grant holder (695669) and acknowledges support from the ACS Green Roundtable.

References

1 Arnold, F. H. Innovation by evolution: bringing new chemistry to life (Nobel Lecture). *Angew. Chem. Int. Ed* **58**, 14420-14426 (2019).
2 *RosettaCommons - the hub for Rosetta modeling software*, www.rosettacommons.org
3 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021).
4 Goldenzweig, A. *et al.* Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol Cell* **63**, 337-346 (2016).
5 Musil, M. *et al.* FireProt 2.0: web-based platform for the fully automated design of thermostable proteins. *Brief Bioinform* **25** (2023).
6 Hon, J. *et al.* SoluProt: prediction of soluble protein expression in Escherichia coli. *Bioinformatics* **37**, 23-28 (2021).
7 Shroff, R. *et al.* Discovery of Novel Gain-of-Function Mutations Guided by Structure-Based Deep Learning. *ACS Synth Biol* **9**, 2927-2935 (2020).
8 Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49-56 (2022).
9 Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089-1100 (2023).
10 Buda, K., Miton, C. M., Fan, X. C. & Tokuriki, N. Molecular determinants of protein evolvability. *Trends Biochem Sci* **48**, 751-760 (2023).
11 Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein science* **25**, 1204-1218 (2016).

12 Buller, R. *et al.* From nature to industry: Harnessing enzymes for biocatalysis. *Science* **382**, 8615 (2023).
13 Sheldon, R. A. The E factor at 30: a passion for pollution prevention. *Green Chemistry* **25**, 1704-1728 (2023).
14 France, S. P., Lewis, R. D. & Martinez, C. A. The evolving nature of biocatalysis in pharmaceutical research and development. *JACS Au* **3**, 715-735 (2023).
15 O'Brien, P. J. & Herschlag, D. Catalytic promiscuity and the evolution of new enzymatic activities. *Chemistry & biology* **6**, R91-R105 (1999).
16 Truppo, M. D. Biocatalysis in the pharmaceutical industry: the need for speed. *ACS Medicinal Chemistry Letters* **8**, 476-480 (2017).
17 Savile, C. K. *et al.* Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science* **329**, 305-309 (2010).
18 Kumar, R. *et al.* Biocatalytic reductive amination from discovery to commercial manufacturing applied to abrocitinib JAK1 inhibitor. *Nature Catalysis* **4**, 775-782 (2021).
19 Schober, M. *et al.* Chiral synthesis of LSD1 inhibitor GSK2879552 enabled by directed evolution of an imine reductase. *Nature Catalysis* **2**, 909-915 (2019).
20 Gantz, M., Neun, S., Medcalf, E. J., van Vliet, L. D. & Hollfelder, F. Ultrahigh-throughput enzyme engineering and discovery in in vitro compartments. *Chemical Reviews* **123**, 5571-5611 (2023).
21 Agresti, J. J. *et al.* Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proceedings of the National Academy of Sciences (USA)* **107**, 4004-4009 (2010).
22 Sumida, K. H. *et al.* Improving protein expression, stability, and function with ProteinMPNN. *Journal of the American Chemical Society* **146**, 2054-2061 (2024).
23 Lu, H. *et al.* Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **604**, 662-667 (2022).
24 Notin, P. *et al.* ProteinGym: large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems* **36** (2024).
25 Yang, J., Li, F.-Z. & Arnold, F. H. Opportunities and challenges for machine learning-assisted enzyme engineering. *ACS Central Science* (2024).
26 Gielen, F. *et al.* Ultrahigh-throughput-directed enzyme evolution by absorbance-activated droplet sorting (AADS). *Proc Natl Acad Sci U S A* **113**, E7383-E7389 (2016).
27 Zurek, P. J., Knyphausen, P., Neufeld, K., Pushpanath, A. & Hollfelder, F. UMI-linked consensus sequencing enables phylogenetic analysis of directed evolution. *Nat Commun* **11**, 6023 (2020).
28 Sharma, M. *et al.* A Mechanism for Reductive Amination Catalyzed by Fungal Reductive Aminases. *ACS Catalysis* **8**, 11534-11541 (2018).
29 Lenz, M. *et al.* New imine-reducing enzymes from beta-hydroxyacid dehydrogenases by single amino acid substitutions. *Protein Eng Des Sel* **31**, 109-120 (2018).
30 Gilio, A. K. *et al.* A Reductive Aminase Switches to Imine Reductase Mode for a Bulky Amine Substrate. *ACS Catal* **13**, 1669-1677 (2023).
31 Chen, L. *et al.* Learning protein fitness landscapes with deep mutational scanning data from multiple sources. *Cell Syst* **14**, 706-721 e705 (2023).
32 Bommarius, A. S. & Paye, M. F. Stabilizing biocatalysts. *Chemical Society Reviews* **42**, 6534-6565 (2013).
33 Ihara, M., Kawano, Y., Urano, M. & Okabe, A. Light driven CO2 fixation by using cyanobacterial photosystem I and NADPH-dependent formate dehydrogenase. *PLoS One* **8**, e71581 (2013).
34 Knaus, T., Corrado, M. L. & Mutti, F. G. One-Pot Biocatalytic Synthesis of Primary, Secondary, and Tertiary Amines with Two Stereocenters from α, β-Unsaturated Ketones Using Alkyl-Ammonium Formate. *ACS catalysis* **12**, 14459-14475 (2022).
35 Wrenbeck, E. E., Azouz, L. R. & Whitehead, T. A. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nat Commun* **8**, 15695 (2017).

36 Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current biology* **24**, 2643-2651 (2014).
37 Miton, C. M. & Tokuriki, N. How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci* **25**, 1260-1272 (2016).
38 Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111-114 (2006).
39 Bloom, J. D. L., Sy. T.; Otey, C.R.; Arnold, F.R. Protein stability promotes evolvability. *Proc Natl Acad Sci U S A* **103**, 5869–5874 (2006).
40 Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Current opinion in structural biology* **19**, 596-604 (2009).
41 Campbell, E. *et al.* The role of protein dynamics in the evolution of new enzyme function. *Nature chemical biology* **12**, 944-950 (2016).
42 Tokuriki, N. & Tawfik, D. S. Protein dynamism and evolvability. *Science* **324**, 203-207 (2009).
43 Kawashima, S. *et al.* AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* **36**, D202-205 (2008).
44 Gelman, S., Fahlberg, S. A., Heinzelman, P., Romero, P. A. & Gitter, A. Neural networks to learn protein sequence-function relationships from deep mutational scanning data. *Proc Natl Acad Sci U S A* **118** (2021).
45 Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
46 Hsu, C., Nisonoff, H., Fannjiang, C. & Listgarten, J. Learning protein fitness models from evolutionary and assay-labeled data. *Nat Biotechnol* **40**, 1114-1122 (2022).
47 Chennubhotla, C., Rader, A., Yang, L.-W. & Bahar, I. Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies. *Physical biology* **2**, S173 (2005).
48 Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).
49 Ma, E. J. *et al.* Machine-Directed Evolution of an Imine Reductase for Activity and Stereoselectivity. *ACS Catalysis* **11**, 12433-12445 (2021).
50 Rapp, J. T., Bremer, B. J. & Romero, P. A. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nature Chemical Engineering* **1**, 97-107 (2024).
51 Wittmann, B. J., Johnston, K. E., Almhjell, P. J. & Arnold, F. H. EvSeq: cost-effective amplicon sequencing of every variant in a protein library. *ACS Synthetic Biology* **11**, 1313-1324 (2022).
52 Medcalf, E. J., Gantz, M., Kaminski, T. S. & Hollfelder, F. Ultra-high-throughput absorbance-activated droplet sorting for enzyme screening at kilohertz frequencies. *Anal Chem* **95**, 4597-4604 (2023).
53 Zinchenko, A. *et al.* One in a million: flow cytometric sorting of single cell-lysate assays in monodisperse picolitre double emulsion droplets for directed evolution. *Anal Chem* **86**, 2526-2533 (2014).
54 Ladeveze, S., Zurek, P. J., Kaminski, T. S., Emond, S. & Hollfelder, F. Versatile product detection via coupled assays for ultrahigh-throughput screening of carbohydrate-active enzymes in microfluidic droplets. *ACS catalysis* **13**, 10232-10243 (2023).
55 Penner, M. *et al.* Sub-single-turnover quantification of enzyme catalysis at ultrahigh throughput via a versatile NAD (P) H coupled assay in microdroplets. *bioRxiv*, 2023.2011. 2022.568356 (2023).
56 Scheele, R. *et al.* Ultrahigh throughput evolution of tryptophan synthase in droplets via an aptamer-biosensor. *bioRxiv*, 2023.2010. 2011.561886 (2023).
57 Reetz, M. T. *et al.* Expanding the substrate scope of enzymes: combining mutations obtained by CASTing. *Chemistry–A European Journal* **12**, 6031-6038 (2006).
58 Sumbalova, L., Stourac, J., Martinek, T., Bednar, D. & Damborsky, J. HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. *Nucleic Acids Research* **46**, W356-W362 (2018).
59 Romero, P. A., Tran, T. M. & Abate, A. R. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proc Natl Acad Sci U S A* **112**, 7159-7164 (2015).

60 Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397-401 (2016).
61 Starita, L. M. *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc Natl Acad Sci U S A* **110**, E1263-1272 (2013).
62 Tiessen, A., Pérez-Rodríguez, P. & Delaye-Arredondo, L. J. Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC research notes* **5**, 1-23 (2012).
63 Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310 e1220 (2020).
64 Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N protein engineering with data-efficient deep learning. *Nature methods* **18**, 389-396 (2021).
65 Yeh, A. H.-W. *et al.* De novo design of luciferases using deep learning. *Nature* **614**, 774-780 (2023).
66 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
67 Buchler, J. *et al.* Algorithm-aided engineering of aliphatic halogenase WelO5* for the asymmetric late-stage functionalization of soraphens. *Nat Commun* **13**, 371 (2022).
68 Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature methods* **6**, 343-345 (2009).
69 Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nature communications* **9**, 2542 (2018).
70 Faure, A. J., Schmiedel, J. M., Baeza-Centurion, P. & Lehner, B. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol* **21**, 207 (2020).
71 Chiu, J., March, P. E., Lee, R. & Tillett, D. Site-directed, Ligase-Independent Mutagenesis (SLIM): a single-tube methodology approaching 100% efficiency in 4 h. *Nucleic acids research* **32**, e174-e174 (2004).
72 Hie, B. L. *et al.* Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology* **42**, 275-283 (2024).
73 Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences (USA)* **118**, e2016239118 (2021).
74 Chen, T., He, T., Benesty, M. & Khotilovich, V. Package 'xgboost'. *R version* **90**, 40 (2019).
75 Chen, T. & Guestrin, C. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 785-794.

Methods

Ultra-high throughput screening in droplets

All solutions were filtered before they were applied to microfluidic chips. Aqueous solutions without protein components were filtered using Millex-GV filters (Merck), Novec HFE-7500 (3M) with 2% (w/v) 008-Fluorsurfactant (RAN Biotechnologies) was filtered using GD/X 13 filters (Whatman), and aqueous solutions containing protein components were filtered using Minisart syringe filters (Sartorius). 0.5x1mm (I.DxO.D.) PTFE tubing (Bola), glass syringes (Hamilton, SGE) and a syringe pump system (Nemesys, Cetoni) were used to inject solution onto polydimethylsiloxane (PDMS) chips. A high-speed camera (Phantom, Miro eX4, Vision Research) was used to determine droplet size and frequency. Electric currents were applied

via a function generator (TG2000, TTi), a pulse generator (TGP110M 10MHz, TTi) and an amplifier (Trek, Model 601C). Droplets were stored in custom-made droplet storage devices[20]. Droplets were generated at 2000 Hz using a flow-focusing device (**Supplementary Figure 1, Supplementary methods**) with cells solution (OD_{600} 0.01, to reach an average occupancy of 0.5 cells per droplet, in 20% Percoll, Sigma) and substrate solution (0.2 mg/ml polymyxin B (Sigma), 2 µl/ml R-lysozyme (Merck), 10 mM NADPH (Carbolution), 20 mM cyclohexanone (Fluka), 40 mM cyclopropylamine (from pH-adjusted stock of 500 mM cyclopropylamine (Sigma) in 100 mM Tris pH 8.0) and HFE (Fluorochem) with 2% RAN surfactant. Flowrates were adjusted to yield droplets with a volume of approximately 170 pl. Droplets were incubated for either 3h or 22h. Picoinjection was performed at 250-500 Hz (1.0 - 2.0 V constant sine wave) with 5 mM WST-1 (NBS biologicals), 5 µg/mL mPMS (Sigma) and 6 mM Tartrazine (Sigma) in a pico-injection device (**Supplementary Figure 1, Supplementary methods**)). 85 pl were injected and flowrates at injection speed were carefully adjusted to rule out droplet fusion. Sorting was performed at 150-200 Hz using custom Arduino and LabVIEW scripts [26] with 4-7 V voltage. A custom-made sorting device was used (**Supplementary Figure 1, Supplementary methods**)**)** that was fabricated as previously described. Droplets were collected in a DNA LoBind tube (Eppendorf). The emulsion of collected droplets was broken by adding 1H,1H,2H-2H-Perfluoro-octanol in a 1:1 (v/v) ratio with HFE-7500 (3M) and 100 µl 2 ng/µl salmon sperm DNA (Thermo), and subsequent vortexing (1 min, full-speed). After centrifugation (14000xg, 1 min), the aqueous solution was carefully pipetted off to avoid transferring the oil phase and DNA was extracted three times by repeated addition of 2 ng/µl Salmon sperm DNA, vortexing and centrifugation. The resulting aqueous solution was concentrated using a spin column (Monarch, NEB) following the manufacturer's instructions. DNA was either eluted in 6 µl nuclease-free water (Invitrogen) and transformed into electrocompetent *E. coli* (E cloni, Lucigen) for secondary screening, or eluted in 20 µl nuclease-free water for NGS.

Cloning of UMI-tagged *Sr*IRED error-prone PCR library

The *Sr*IRED gene was amplified with error-prone PCR using the GeneMorph II random mutagenesis kit (Agilent) following the instructions of the manufacturer using 2.4 µg template plasmid and 25 PCR cycles. The product was subjected to DpnI digest (NEB) overnight and purified using SPRIselect beads (Beckman coulter) following the manufacturer's instructions. The UMI was ordered as oligonucleotide (IDT) and 3 cycles of PCR with a single primer annealing to its 3' end were performed to produce double-stranded DNA. Two more fragments (vector and promoter region) were produced by PCR. Gibson assembly mix[68] was added on ice, and the mix was incubated for 1h at 50°C. After purification using SPRIselect beads (Beckman Coulter), plasmid DNA was transformed into electrocompetent *E. coli* (Lucigen).

Oxford Nanopore sequencing

After library transformation into *E. coli*, DNA for sequencing was prepared by adding 10 ml LB media to the grown agar plate, scraping the colonies using a sterile L-scraper, recovering the cell suspension, harvesting the cells by centrifugation (4000xg), and minipreping following the kit manufacturer's instructions (Genejet). Restriction digest with AclI (NEB) and Bsu36I (NEB) yielded a DNA fragment with UMI and the *Sr*IRED gene. The fragment was purified via agarose gel electrophoresis and SPRIselect beads. 500 fmol of DNA were sequenced on three Minion10.4 flowcells (Oxford nanopore technologies) using the Oxford Nanopore ligation sequencing kit following the manufacturers protocol.

Preparation of libraires for NGS

For NGS library preparation, primers including Illumina flow-cell binding sequences and indices in overhangs were used to amplify the UMI region of the recovered DNA (**Supplementary Table 1**). The different microfluidics runs were multiplexed using primer pairs with different NGS indices. Upon preparation of the PCR mix (NEBNext Q5 mix, NEB) following the instructions of the manufacturer, PCR conditions were optimized using a 15% fraction of the recovered DNA and qPCR to determine the cycle number at 10% of the maximum fluorescence. 18-22 cycles were employed for subsequent recovery. DNA was purified using SPRIselect beads (Beckman Coulter), with a beads-to-sample volume ratio of 0.7 to remove primer dimers. The DNA concentration was determined using NEBnext library quant kit (NEB) and capillary electrophoresis (Bioanalyzer, Agilent) following the manufacturers protocols. The samples were pooled and prepared into Illumina TruSeq libraries by the University of Cambridge, Department of Biochemistry sequencing facility according to the manufacturer's instructions. Next-generation sequencing was performed using an Illumina next-seq 2x150bp run (with 20% PhiX spike in).

Generation of dictionary and fitness scores

Basecalling of Oxford Nanopore sequencing data was performed using guppy (ONT), quality control was performed using NanoPlot and filtering using NanoFilt (length 1,250 --maxlength 1,400 --quality 8). A custom python script was used to extract the UMI sequences [27] and the generated fasta file was clustered using the mmseqs2 suite[69] (min-seq-id 0.8, seq-id-mode 2, cov-mode 0 -c 0.8). The files were then processed in a custom pipeline using minimap2 to align the sequences in every cluster, racon[53] to generate a draft sequence assembly, minialign to align draft assembly and reference (with NNNN in variable regions of the UMI) and medaka (ONT) for generation of a consensus file. The vcf files generated by medaka were converted into .fasta files using custom python scripts. A variant identifier file for downstream data generation was generated by custom python scripts. Fitness scores were calculated from the

variant identifier file and the raw NGS foward reads using the DiMSum suite[70] with the following parameters: retainIntermediateFiles T, cutadaptMinLength 20, vsearchMinQual 10, cutadaptErrorRate 0.6, numCores 8, paired F, indels all, maxSubstitutions 100, mixedSubstitutions T, barcodeErrorRate 0.1, fitnessMinInputCountAll 0.

Cloning, expression and purification of *Sr*IRED variants

*Sr*IRED variants (N-terminal His6-tag, under control of a T7 promoter in a pRSF vector) were either cloned using site-directed ligation-independent mutagenesis (SLIM, [71]), Gibson assembly[68] or using dial out PCR with primers specific to the UMI region from the library[27]. Cloning success was confirmed by Sanger sequencing. Variants were transformed into BL21(DE3) cells (NEB), a single clone was picked and grown to saturation in LB media supplemented with kanamycin (50 µg/ml) (37°C, 200rpm). A culture was diluted to an OD_{600} of 0.05, grown to an OD_{600} of 0.4-0.6 and induced with 0.2 mM IPTG. Cells were expressed overnight (20°C, 200rpm) and harvested by centrifugation (4000xg). Cells were resuspended in lysis buffer (100 mM Pi pH 7, 100 mM NaCl, 1 mM DTT, C0mplete protease inhibitor tablet, Roche) and lysed by sonication for 8 min (Vibracell, 40%, 1 sec on, 1 sec off,). The lysate was cleared by centrifugation (14000xg, 4°C), and incubated with NiNTA agarose (NeoBiotech) on a roll mixer at 4°C. A plastic gravity column with a frit was used to wash the beads three times with 10 ml wash buffer (100 mM phosphate pH7, 100 mM NaCl, 1 mM DTT, 30 mM imidazole) and to elute with elution buffer (100 mM phosphate pH7, 100 mM NaCl, 1 mM DTT, 300 mM imidazole). The protein was rebuffered into storage buffer (100 mM phosphate pH7, 100 mM NaCl, 1 mM DTT, 10% glycerol) using amicon ultra centrifugation (10,000 NMWL) filters (Merck), shock frozen with liquid nitrogen and stored at -80°C. Enzyme concentration was determined in a Bradford assay (Sigma) and purity was controlled by SDS PAGE.

Lysate activity assays

After transformation into chemically competent BL21(DE3) cells (NEB), colonies were picked into deep-well plates (Greiner) and filled with 1 ml LB media supplemented with kanamycin. Plates were incubated overnight (37°C, 750 rpm). 30 µl of the culture were diluted in 920 µl of fresh LB media supplemented with kanamycin in a fresh deep-well plate and grown for 2h (37°C, 750 rpm). Protein expression was induced with 0.2 mM IPTG, and plates were incubated overnight (20-25 °C, 750 rpm). Cells were harvested by centrifugation (30 min, 4000g), the supernatant was removed, and cells were resuspended in 200 µl lysis buffer (100 mM Tris pH 8, 4 mg/ml egg white lysozyme, 0.5 mM EDTA, 1 mg/ml polymyxin B) per well. After incubation (25°C, 1h, 750 rpm), lysate was cleared by centrifugation (30 min, 4000g, 4°C). Assays were performed in Nunc 96-well plates (Thermo). To determine the initial rate of the reaction, the reaction was induced by adding 10 µl lysate to 90 µl reaction buffer (10 mM carbonyl compound, 20 mM amine, 0.5 mM NADPH in 100 mM Tris pH 8) and the absorption

at λ of 340 nm was followed in a plate reader (Spectramax 190, molecular devices). The initial rate in μM/s was calculated using a calibration curve of known NADPH concentrations at 340 nm. To determine the conversion of the reaction in an assay similar to the droplet assay, cleared lysate was used to carry out a reaction with 5 mM NADPH, 10 mM cyclohexanone and 20 mM cyclopropylamine in 100 mM Tris pH 8.0. After 3h and 22h incubation at 25°C, 5 µl of the reaction mix were added to 95 µl 5 mM Iodonitrotetrazolium chloride which is a cheaper alternative to WST-1 (INT, Sigma) and 5 µg/ml mPMS in DMSO. The reaction was diluted 1:20 and the absorbance was measured in a spectrophotometer (Spectramax 190, Molecular Devices) in Nunc 96-well plates (ThermoFisher). In analogy to the calculation of a fitness score, the natural logarithm of the average fold-change of INT absorbance over wild type (proportional to the remaining NADPH concentration) after 3h and 22h was calculated.

Michaelis Menten kinetics

Kinetic constants of *Sr*IRED for cyclohexanone were determined in 100 mM Tris pH 8.0 with 0.5 mM NADPH and varying concentrations of cyclohexanone (0.06 mM - 35 mM) in transparent Nunc 96-well plates (ThermoFisher) and by monitoring the depletion of NADPH concentration at λ of 340nm (Spectramax 190, Molecular Devices) with 30 mM cyclopropylamine and 0.001 - 0.01 mg/ml enzyme at T = 25 °C. Initial rates were measured in 3 replicates and curves were fitted using Prism.

Soluble expression and thermal shift assays

Soluble expression of variants was determined by SDS PAGE analysis of the pellet and the supernatant by densitometry. Variants were expressed analogous to expression for the lysate assay and the lysate was cleared by centrifugation. The pellet was resuspended in 200 µl Laemmli buffer (NuPAGE, invitrogen) and samples of the supernatant were collected; SDS PAGE analysis was conducted and band intensity was compared using ImageJ. For thermal shift assays, 10 µM enzyme with 2x SyproOrange (invitrogen) in 100 mM Tris pH 8.0 were analysed in a Bio-Rad CFX Connect device.

Heuristics for generation of a mutational profile

To guide rational engineering, we define two metrics – mutability and combinability – reflecting how well individual positions accept mutations and how well mutations can be combined into improved higher-order variants. Our discussion is restricted to substitution mutations only. We use (p, a) to denote the substitution of position p in the wild type sequence to the 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 occur at most once. The mutation order $|v|$ is the number of mutations K in variant 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

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

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

$$\text{Combinability}(p_0) = \sum_{\{v\ |\ (p=p_0,\cdot)\ \in\ v\ \wedge\ |v|>1\}} |v|\ \cdot \mathbb{I}_{[f(v)>0]}\ \cdot \mathbb{I}_{[f(v)\geq\sum_{(p,a)\in v} f((p,a))]},$$

with $\mathbb{I}$ 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 $f(v)$. 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 combined plot of the *mutability* and *combinability* heuristics for rational engineering along the *Sr*IRED sequence can be found in **Supplementary Figure 17**.

Machine learning modelling

The single and double mutant AI predictions come from two simple models which were trained on the obtained lrDMS fitness data. The single mutant model is based on a ridge-regression model using fixed, biochemically informed amino acid embeddings and pre-trained language model likelihood while the double mutant model is a gradient boosted tree model on a few curated features deemed important for predicting the fitness of double mutants and tailored to the nature of lrDMS data. We chose these simple models for their relative robustness and interpretability, allowing us to better understand the value of the lrDMS data and to better mitigate the effects of noise in the data, which we expected to be larger than data from plate-based assays on which many current models are evaluated. We note that in the future the predictive performance could be further enhanced by utilizing more complex models which leverage pre-trained information from sequence, structure, or related assay data more extensively, provided they are paired with approaches to reduce model overfitting.

Single mutant machine learning model

The single mutant model is a ridge regression model, which is augmented with an evolutionary density from a large language model as proposed in Hsu et al.[46], to predict fitness scores for unobserved single mutants. For an amino acid sequence $x = (x_{,1}x_{2,} \cdots, x_L)$ we parameterize the fitness as

$$f\big(x; \{\alpha_{i,i}\}, \beta, \gamma\big) = \sum_{i=1}^{L} \sum_{j=1}^{d} \alpha_{i,j}\, \mathrm{EMB}(x_i)_j + \beta + \gamma\, \mathrm{PLL}(x).$$

The embedding $\mathrm{EMB}: \{\mathrm{Ala}, \mathrm{Arg}, \ldots, \mathrm{Tyr}, \mathrm{Val}\} \to \mathbb{R}^d$ for a given amino acid is computed from a principal component analysis of the AAINDEX database[43], which contains reference data on over 500 physiochemical and biochemical properties, such as e.g. hydrophobicity, for each natural amino acid. As in Gelman et al.[44] we reduce the dimensionality via principal component analysis to a 19-dimensional feature vector ($d = 19$) for each amino acid, still capturing over 98% of the variance. In comparison to the one-hot encoding used by Hsu et al.[46], this chemically informed embedding better captures the similarity between amino acids. This chemical similarity between amino acids is crucial to help the model extrapolate 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.

We further include evolutionary information by adding the pseudo-log-likelihood (PLL) of a pre-trained large language model. Based on its successful use in prior work[46,72] we use the ESM2 masked language model[45]. The PLL for masked language models,

$$\mathrm{PLL}(x) = \sum_{i}^{L} \log p(x_i | x_{-i})\,,$$

where x_{-i} denotes masking at position i, requires L function evaluations, making it expensive to compute for many variant sequences x. To reduce the computational cost, we instead use the PLL difference versus the wildtype sequence, which only differs from $\mathrm{PLL}(x)$ by a constant and can further be approximated as[46,73].

$$\begin{aligned}\mathrm{PLL}(x) - \mathrm{PLL}(x_{\mathrm{wt}}) &= \sum_{i=1}^{L} \log p(x_i | x_{-i}) \; - \log p\,\big(x_{\mathrm{wt},i} \big| x_{\mathrm{wt},-i}\big) \\ &\approx \sum_{\{i | x_i \,\neq\, x_{\mathrm{wt},i}\}} \log p(x_i | x_{-i}) - \log\, p\big(x_{\mathrm{wt},i} \big| x_{\mathrm{wt},-i}\big).\end{aligned}$$

This approximation assumes that the conditional log-likelihood of an amino acid at any position is almost the same regardless of whether the wild type or mutant background are given, an assumption which we observed holds well when the mutation order is low for a random subsample of mutants. We used the ESM2 3B parameter model in our experiments.

The 'augmented' ridge-regression model f is then fitted to the assay-labelled single-mutant data using 5-fold cross validation to tune the L^2 regression hyperparameter λ. Regularization was only applied to the amino acid features, but not the PLL component of the model. To compute the final predictions, the model is re-fitted on the entire single-mutant data using the hyperparameters obtained from cross-validation. A detailed evaluation of the single mutant model performance can be found in **Extended Data Figure 8** and **Supplementary Figure S20.**

Double mutant machine learning model

The double mutant model is a gradient boosted tree model (xgboost[74]) on a selected group of features. The features can be grouped into four categories:

(1) Assay derived features: These include the (1) fitness of the least fit mutation in the double mutant, as well the (2) additive fitness of both single mutations assuming no epistatic effects. 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). For historic reasons, the combinability information used in the double mutant model differs slightly from the *combinability* heuristic used for rational engineering. In the double mutant model, two combinability features, defined as (using the same notation as above)

$$\text{Combinability}^{+}(p_0) = \sum_{\{v \mid (p=p_0,\cdot) \in v \wedge |v|>1\}} \mathbb{I}_{[f(v)\geq 0]} \ ,$$

$$\text{Combinability}^{-}(p_0) = \sum_{\{v \mid (p=p_0,\cdot) \in v \wedge |v|>1\}} \mathbb{I}_{[f(v)<0]}$$

are used. Compared to the rational engineering heuristic, the weighting by mutation order $|v|$ and the indicator function for better-than-additive mutations are removed and we instead define a positive or negative combinability which count the number of times a mutation at the given position p_0 was observed in a better/worse than wildtype variant. The combinability-type information benefits particularly from the epPCR library generation step in the lrDMS workflow, as this step tends to generate a *stacked* library where higher order mutants are based off generated lower order mutants.

(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 proposed catalytic site residues. We also include (4) the pairwise distance between the Cα atoms of the two mutants. A profile of the structural features along the *Sr*IRED sequence is shown in **Supplementary Figure S25**.

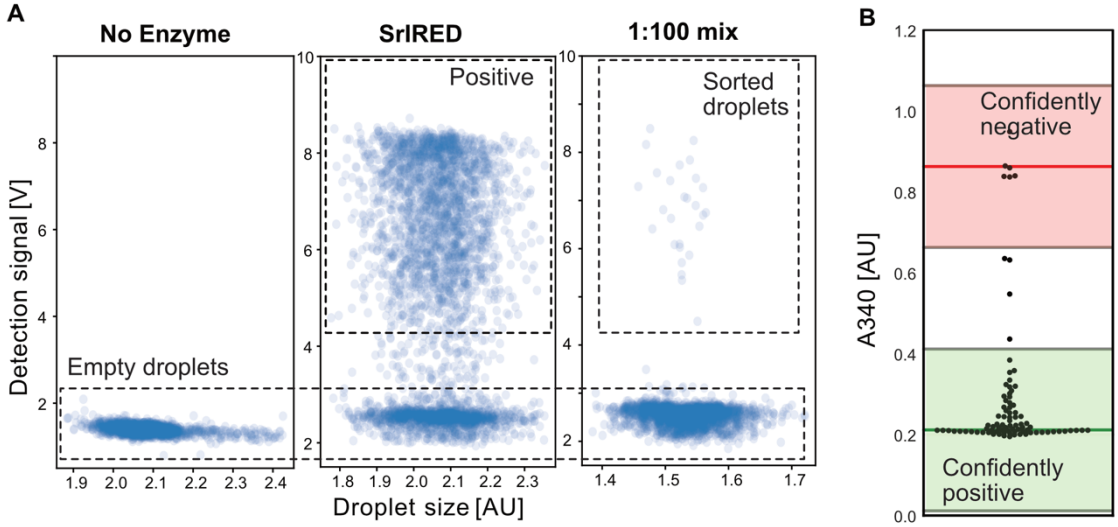
(3) *Coarse grained dynamics information:* We use a coarse-grained normal mode analysis using and invariant elastic network model on the wild type structure (`5OCM`) to derive the mean squared expected fluctuation, a value that mimics b-factors, of each position. We use provide the mean squared fluctuation of both single mutants (min & max) as a feature. A profile of the mean squared residue fluctuation along the *Sr*IRED sequence is shown in **Supplementary Figure S25**.

(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 is defined as the difference between ESM2 PLL for the double mutant versus the sum of the PLL predictions for the single mutants. We use the same PLL approximations described above for the single mutant model and cache intermediate results where possible, to reduce the computational burden of these features when computing them for a library. As demonstrated in the ablations in **Extended Data Figure 9** and in the top double mutant predictions when removing language model features (**Supplementary Figure S12**), the model can perform well even without these more computationally expensive language model features.

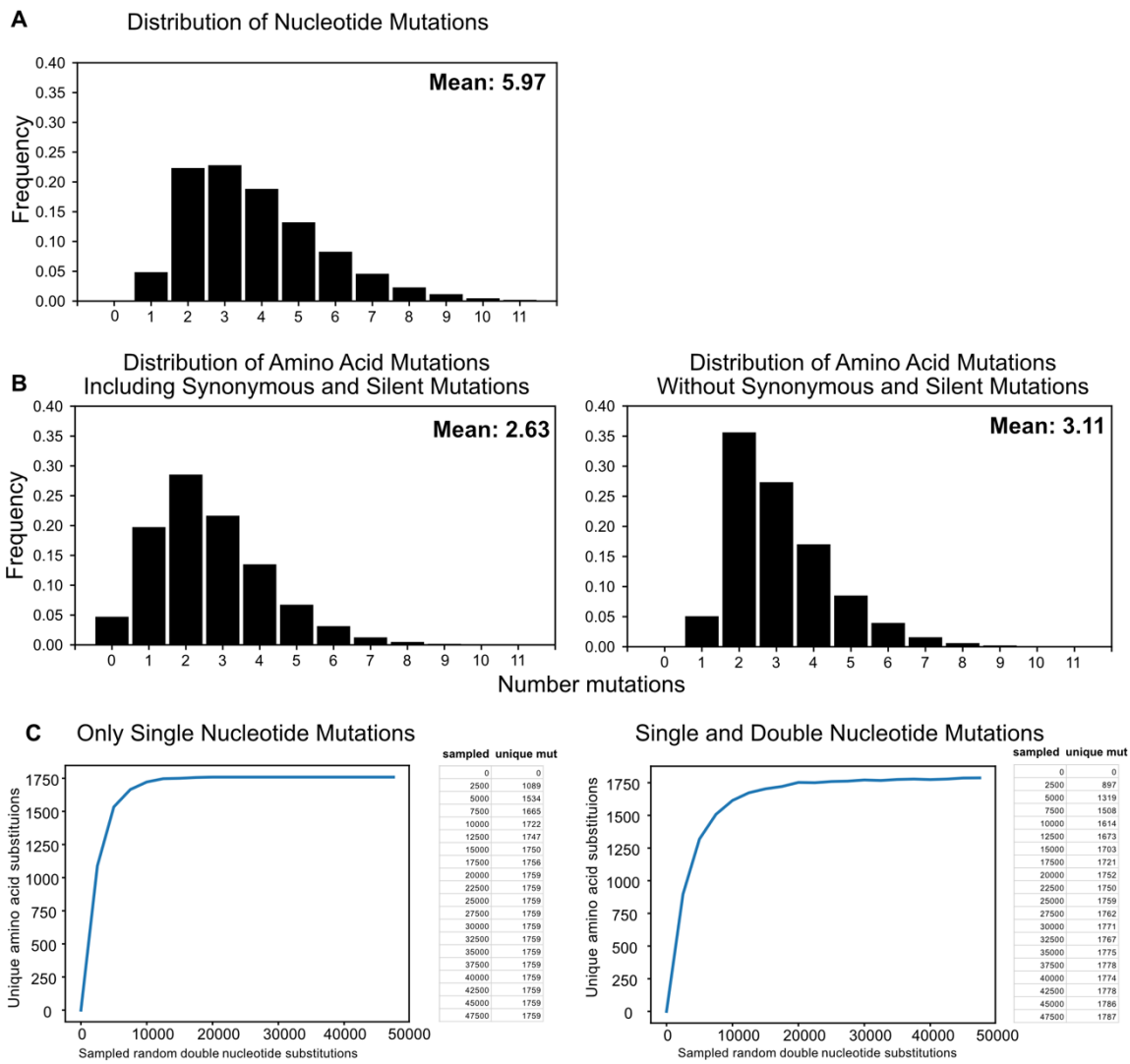
We performed feature selection and hyperparameter tuning of the double-mutant model using cross validation. Our final double-mutant model is composed of 100 estimators, each of maximum depth 3, and was obtained by training with a squared error regression objective, although we also explored training with ranking and multiclass objectives. Detailed learning curves and a cross validation evaluation of the final model can be found in **Extended Data Figure 9** and **Supplementary Figure S21.** To obtain predictions for assaying in the lab we re-trained the models on all higher order lrDMS data with the fixed hyperparameters found during tuning and predicted the fitness of all 721,870 double mutants that could be composed by combining single mutants observed in lrDMS. We chose to restrict ourselves to double mutants obtained from single mutants to increase robustness given our low experimental budget (5 assayed mutants per model). Consequently, T241G, the best *predicted* single mutant for example, was not within the scope of double mutant predictions. However, this choice can

easily be relaxed, and the model can take extrapolated single mutant predictions as input to build higher order mutants. To understand how the various features influenced the final double mutant model's predictions we investigated model internal feature importance indicators (gain) and performed a SHAP analysis, detailed in **Extended Data Figure 10** and **Supplementary figures S18 and S19.**

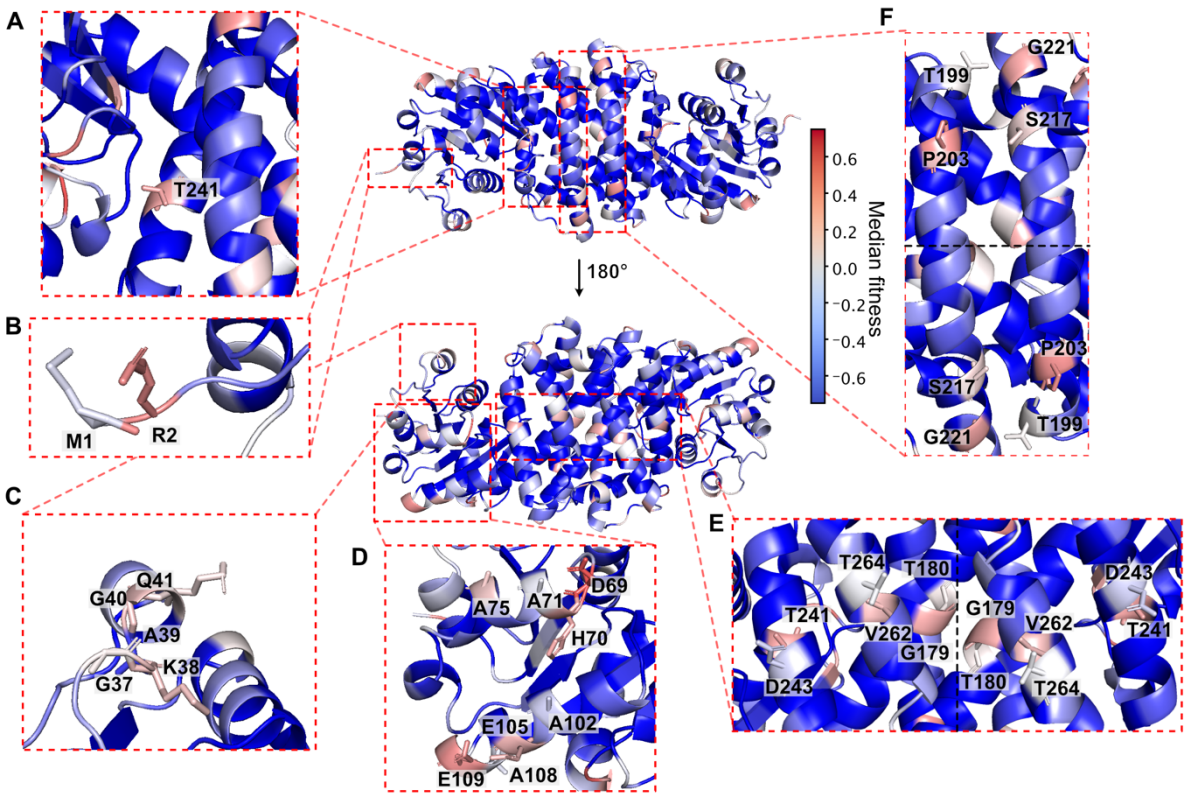
Extended data figures



Extended Data Figure 1: An ultra-high throughput assay for screening imine reductases. A mock-sorting experiment with an IRED from *Streptosporangium roseum* (*Sr*IRED in pRSF vector) diluted with *E. coli* expressing no IRED (empty pRSF vector) shows the capability of the assay to substantially (93-fold) enrich IRED activity over background. (**A**) Droplets with cells expressing no enzyme, cells expressing *Sr*IRED and a 1 to 100 dilution of cells expressing *Sr*IRED with cells expressing no enzyme in the AADS assay. The population with high detection signal represents droplets with IRED-expressing cells. The second population with low detection signal represents empty droplets (due to single-cell encapsulation of a diluted cell suspension, based on the Poisson distribution, 60% of droplets contain no cell) and droplets with non-expressing cells. **(B)** For an enrichment experiment, the positive population was sorted from a 1 to 100 dilution of cells expressing *Sr*IRED and cells expressing no enzyme. The plasmid DNA was recovered by transformation and a secondary assay was performed with 83 clones. The reductive amination activity was determined in lysate with 0.5 mM NADPH, 10 mM cyclohexanone and 20 mM cyclopropylamine. An endpoint was measured after 3 h at 340 nm. Positive and negative clones are assigned using an interval of 0.2 from the positive and negative controls (green and red lines). 6 negative and 77 positive clones were found, giving an enrichment factor of 93.

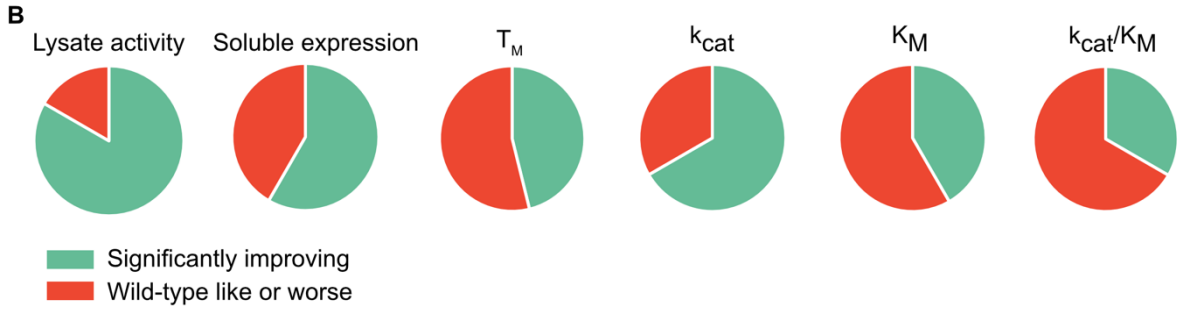


Extended Data Figure 2: Distribution of mutation number in the dictionary. The error-prone PCR library of *Sr*IRED was sequenced using UMIC-seq to generate a_dictionary of 29,841 unique variants present before selection. Analysis of the obtained dictionary shows that the available number of single mutation datapoints number is capped by codon bias. **(A)** Distribution of mutation number of all unique nucleotide mutations in the dictionary. **(B)** Distribution of amino acid mutation number before and after redundant and synonymous variants are removed. The main effect is a loss of apparent diversity in single point mutations and an upwards shift in the average number of mutations from 2.63 to 3.11. (**C**) Simulation of number of expected unique amino acid mutations with different numbers of sampled nucleotide mutations. With only single nucleotide mutations (e.g. C533G; left) and with double nucleotide mutations (e.g. C533G C630T right). If only single nucleotide mutations are allowed, the plateau is reached faster as single mutations can also be achieved without silent mutations. However, after 1759 unique amino acid mutations the possible diversity is exhausted. When double nucleotide mutations are allowed, a higher diversity can be reached, however only after including significantly more diversity (slight increase instead of levelling off). With approximately 60,000 single nucleotide mutations in the library, the 1514 unique amino acid mutations in the dictionary therefore represent a high portion of the achievable theoretical diversity. This shows how the architecture of the genetic code restricts the sequence space available to directed evolution using error-prone PCR. Due to the observed codon bias nearly all higher order mutations are combinations of single point mutants also observed in the dataset promising rich information on the combinability and epistasis in these variants.



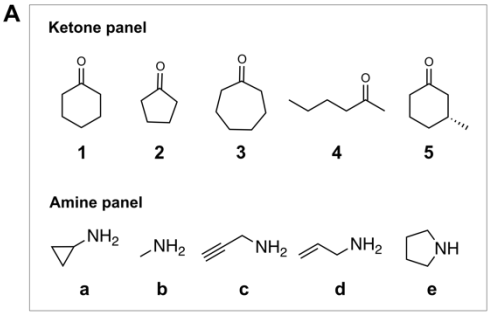
Extended Data Figure 3: Structural grouping of mutability hotspots. (A) T241 is the only residue in the first shell at which mutations show on average improving fitness. (**B**) Mutations at the two N-terminal residues M1 and R2 show high average fitness values indicating a high potential for improvement. M1 can be mutated without a high fitness penalty due to missed translation as a His_6-tag with a second start codon is located at the N-terminus. (**C**) A loop and the beginning of a helix at the edge of the cleft forming the active site shows a pattern of highly mutable positions interspaced by an alanine with low mutability pointing to the inside of the N-terminal domain. The loop is involved in NADPH binding e.g. via K38 interacting with the NADPH phosphate. **(D**) Two helices in the N-terminal domains show high mutability especially at residues pointing to the outside of the protein. (**E**) Patches of residues on a horizontal axis in the central helices show high mutability (the pattern duplicated because of the dimeric structure of the protein as indicated by the black dotted line). Similar to the patches shown in D, a role of the mutations in orientating the central helices at the dimer interface of the protein can be hypothesized. (**F**) In two helices in the central domain two patches of four positions with high mutability were identified (the pattern duplicated because of the dimeric structure of the protein as indicated by the black dotted line). A similar function of improving mutations at these positions for example in orientating the central helices at the dimer interface could be hypothesized.

	Fitness	Lysate activity [fold-change]	Soluble expression [fold-change]	T_M [°C]	k_{cat} [s^{-1}]	K_M [mM]	K_{cat}/K_M [mM-1s^{-1}]
wtype	0	1	1	38.5 ± 0.6	2.1 ± 0.1	3.5 ± 0.3	0.6 ± 0.1
T241A	**1.8 ± 0.5**	**1.3 ± 0.5**	**0.8**	**50.0 ± 0.7**	**5.8 ± 0.2**	**2.3 ± 0.3**	**2.5 ± 0.4**
T241S	**0.7 ± 0.6**	**2.8 ± 0.7**	**1.2**	**37.6 ± 0.6**	**6.0 ± 0.4**	**2.3 ± 0.5**	**2.6 ± 0.6**
M120L	0.3 ± 0.4	3.7 ± 0.9	1.0	33.3 ± 0.1	2.1 ± 0.1	5.3 ± 0.7	0.4 ± 0.1
L174V	1.1 ± 1.4	3.8 ± 1.2	0.7	27.6 ± 0.2	10.7 ± 0.5	13.8 ± 1.5	0.8 ± 0.1
D69H	0.5 ± 2.4	3.3 ± 1.0	1.2	39.5 ± 0.6	4.1 ± 0.2	5.3 ± 0.6	0.8 ± 0.1
D69G	0.5 ± 0.5	5.7 ± 1.8	1.4	38.8 ± 0.6	3.3 ± 0.1	6.0 ± 0.4	0.5 ± 0.04
D236G	0.4 ± 0.6	2.5 ± 0.7	1.4	27.9 ± 0.4	2.0 ± 0.1	3.1 ± 0.4	0.7 ± 0.1
P203A	2.9 ± 1.3	0.2 ± 0.1	n.d.	40.8 ± 0.2	5.0 ± 0.2	6.7 ± 0.8	0.7 ± 0.1
P203S	0.7 ± 0.5	1.6 ± 0.4	0.9	43.7 ± 1.7	2.5 ± 0.1	3.2 ± 0.2	0.8 ± 0.1
E146V	0.9 ± 0.6	4.2 ± 0.9	1.7	31.1 ± 0.5	1.7 ± 0.1	2.4 ± 0.3	0.7 ± 0.1
D267G	0.2 ± 0.5	5.4 ± 1.2	1.9	43.3 ± 0.6	2.7 ± 0.1	4.3 ± 0.3	0.6 ± 0.1
D5H	1.5 ± 2.3	3.0 ± 1.1	1.8	39.6 ± 0.05	1.9 ± 0.1	5.1 ± 0.9	0.4 ± 0.1
V6L V67I	4.8 ± 2.3	n.d.	n.d.	29.1 ± 0.6	1.7 ± 0.0	2.3 ± 0.2	0.7 ± 0.1



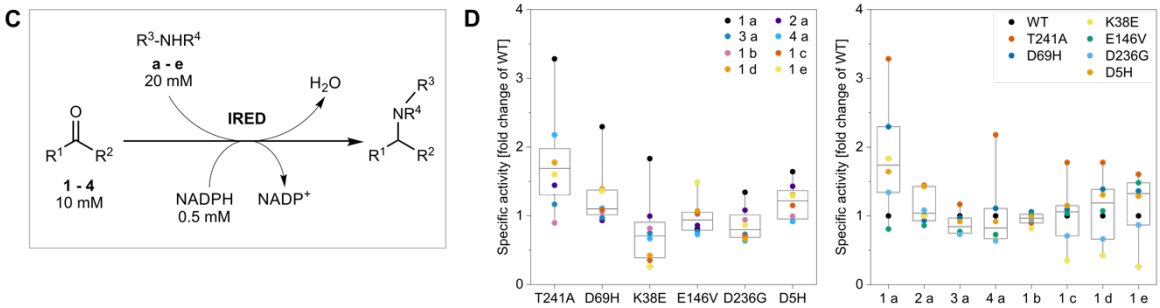
Extended Data Figure 4: Characterization of 12 selected single point mutants with positive fitness and the variant with the highest fitness in the library V6L V67I. (**A**) Lysate activity is given as initial rates measured in three replicates (error bars represent standard deviations) with **1** (10 mM), **2** (20 mM), NADPH (0.5 mM) in 100 mM Tris pH 8.0 following the decline of NADPH absorbance at 340 nm and normalized to lysate without IRED. Soluble expression was determined by densitometry of SDS-PAGE bands of the soluble and insoluble fraction after lysis. Values are normalized to wild type. Melting temperatures were measured in thermal shift assays in three replicates, errors represent standard deviations. Michaelis Menten kinetics were conducted with variable concentrations of ketone substrate (0.05 - 35 mM) with a constant concentration of amine (30 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 at T = 25 °C. Each curve was measured in three replicates. Errors represent standard errors of the fit and three technical replicates with a 95% confidence interval. Amine and NADPH kinetics as well as Michaelis-Menten curves for wild type are provided in **Supplementary Figure 6**. (**B**) Pie

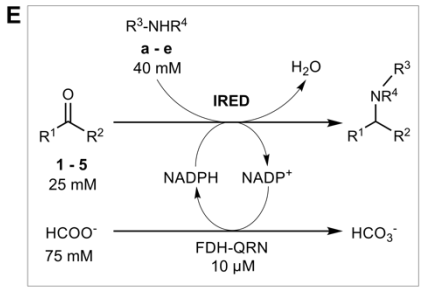
charts representing the fraction of variants significantly improved (1 standard deviation over wild type).



B

Variant	Significance
WT	wild type
T241A	Best variant from DMS and global combinability hotspot
D69H	Positive DMS fitness scores and positively epistatic
K38E	Neutral DMS fitness scores, improving k_{cat} and lower expression
E146V	Positive DMS fitness scores
D236G	Positive DMS fitness scores
D5H	Positive DMS fitness scores
P203A T241A	Best rationally engineered double mutation
T241G	Best AI predicted single mutation
E36A T241A	Best AI predicted double variant
Q41P T241A	Second best AI predicted double variant
E36A T241G	Best AI predicted hybrid variant from single and double mutations
P203A T241G	Best hybrid double mutation from AI-informed and rational engineering





F

Conversion after 4 h ([IRED] = 20 µM, [NADPH] = 0.5 mM)

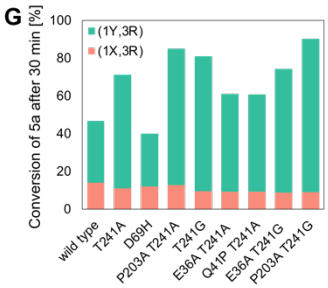
	1a	2a	3a	4a	1b	1c	1d	1e
WT	83%	43%	34%	9%	79%	44%	35%	9%
T241A	97%	53%	39%	19%	72%	54%	46%	18%

Conversion after 24 h ([IRED] = 20 µM, [NADPH] = 0.5 mM)

	1a	2a	3a	4a	1b	1c	1d	1e
WT	98%	70%	64%	41%	100%	78%	56%	33%
T241A	100%	83%	73%	67%	100%	89%	80%	34%

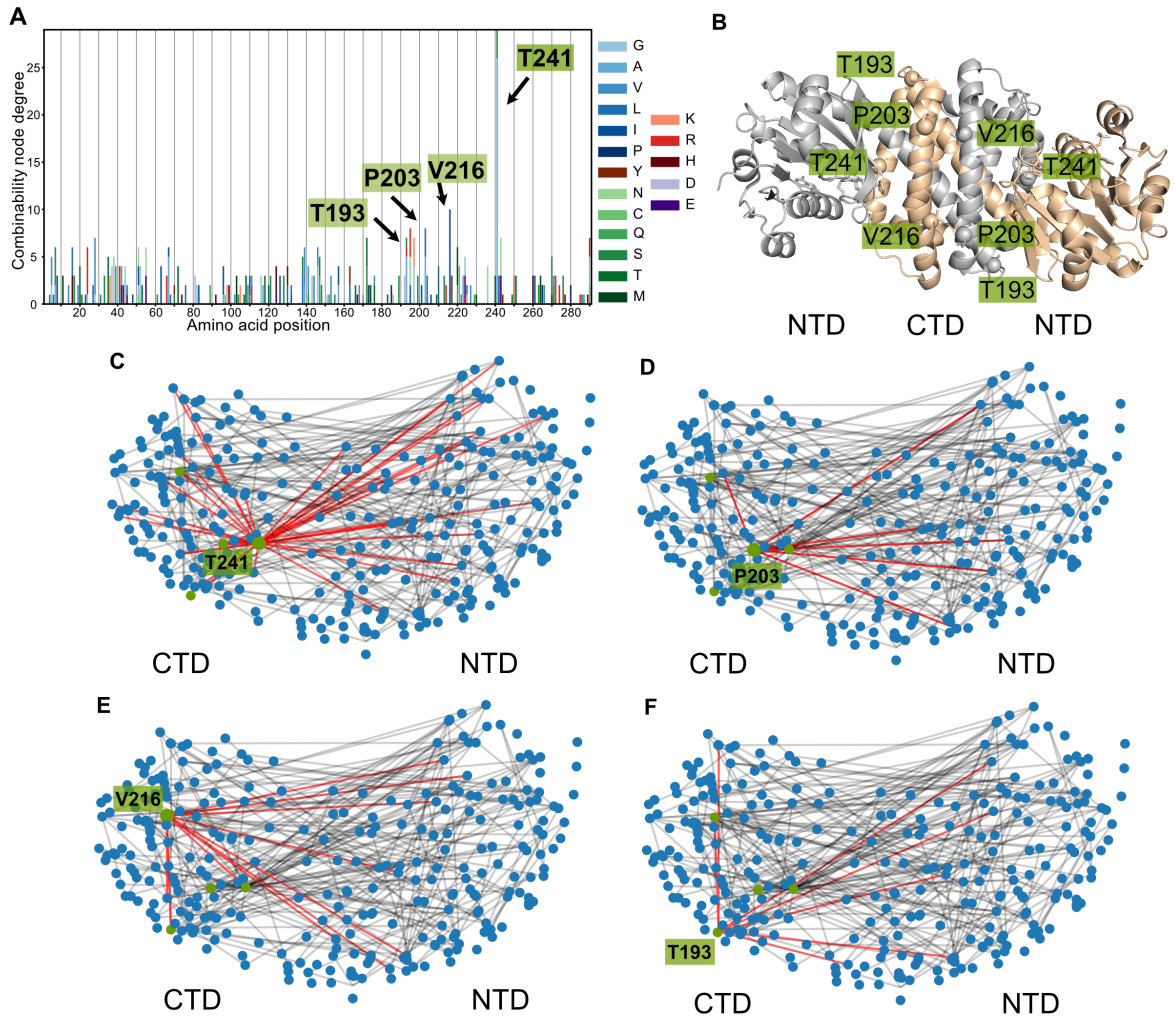
Conversion of 1a ([IRED] = 5 µM, [NADPH] = 1 mM)

	WT	T241A	D69H	P203A T241A	T241G	E36A T241A	Q41P T241A	E36A T241G	P203A T241G
30 min	22%	33%	28%	39%	43%	35%	35%	42%	42%
4 h	63%	83%	62%	82%	81%	81%	78%	79%	83%



Extended Data Figure 5: Substrate Scope. (A) *Sr*IRED wild type and selected mutants were tested with a panel of ketones and amines with varying degree of similarity to the model substrate combination (ketone parameters: ring size, structure; amine parameters: size, secondary/primary, pi-system). (**B**) Criteria for choosing variants for characterization in substrate scopes. **(C)** The single variants were characterized by using 0.5 mM NADPH and following the progress of the reaction by monitoring the depletion of NADPH absorbance at 340 nm (**Supplementary methods**). **(D)** Fold-change in initial rate compared to wild type with the six chosen mutations and a set of substrate combinations (color-coded) with ketone (10 mM), amine (20 mM), NADPH (0.5 mM) in Tris buffer (100 mM pH 8.0). Absolute values of specific activity in mU/ mg are listed in **Supplementary Figure 7**. T241A increases activity for a broad range of different ketones and amines. It shows increased activity with **1** and **a** (3-fold) but lower activity increases with smaller and larger rings. D69H is improving with a narrower spectrum of different substrates: **1e** and **1d**. K38E is strictly specialized to **1** combined with **a** and shows decreasing or unchanged rate with all other substrates. E146V mainly improves lysate activity by increasing the expression level but also shows a slight decrease in K_M. With purified enzyme it therefore shows next to no improvement with the model substrate combination **1a**. D5H (improved expression level and slightly lower K_M) show slight improvements with the model substrate combination but only very low improvements with all other substrates. D236G which only improves expression level shows next to no change in substrate specificity. Fold-change in initial rate compared to wild type for all substrate combinations with color-coded variants shows that T241A is the best variant for all substrate combinations apart from **1b** and **3a** and the only variant significantly improving activity with 2-hexanone **4a**. With all other substrates than **1a** there is a slight decrease in median activity

showing that the discovered mutations overall specialize for **1a**. With methylamine (**1b**) and cycloheptanone (**3a**) most mutations are neutral and no improving mutations were identified suggesting low evolvability for these substrate combinations. Interestingly, with pyrrolidine (**e**) E146V, D5H and D69H are rivalling the improvement with T241A. (**E**) Reaction conditions used for monitoring conversion using GC. As the recycling-enzyme for the NADPH coenzyme, a variant of the formate dehydrogenase from *Candida boidinii* (FDH-QRN) was used. Sodium formate served as the ultimate hydride donor (**Supplementary methods**). (**F**) *Sr*IRED wild type and T241A were tested in biocatalytic reactions for the amination of ketone substrates **1** to **4** with amine donors **a** to **e**. T241A improves conversions with all substrates but **1b**. Raw values with standard deviations are listed in **Supplementary Figure 8** and **Supplementary Figure 15**. **(G)** The top variants' conversion with the chiral substrate (*R*)-3-methyl cyclohexanone with nonchiral amine **a** and the stereoselectivity at the newly formed stereogenic centre (green and red) showing that T241A and G significantly improve conversion on their own and combined with other mutations (up to 90% conversion with a 90:10 diastereomeric ratio compared to a 70:30 diastereomeric ratio for wild type) but only accelerate the reaction towards one diastereomer with conversion to the other diastereomer staying nearly constant or even decreasing. Raw values with standard deviations are listed in **Supplementary Figure 16.**

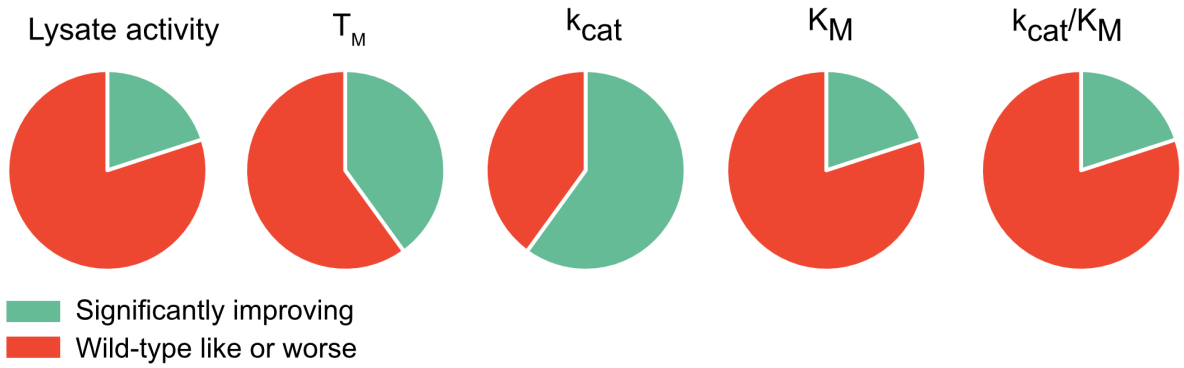


G	Fitness	Combinability score	Lysate activity [fold-change]	T_M [°C]	k_{cat} [s^{-1}]	K_M [mM]	K_{cat}/K_M [mM-1s^{-1}]
wt	0			38.5 ± 0.6	2.1 ± 0.1	3.5 ± 0.3	0.6 ± 0.1
T241A	1.8 ± 0.5	29	1.3 ± 0.5	50.0 ± 0.7	5.8 ± 0.2	2.3 ± 0.3	2.5 ± 0.4
T241S	0.7 ± 0.6	29	2.8 ± 0.7	37.6 ± 0.6	6.0 ± 0.4	2.3 ± 0.5	2.6 ± 0.6
V216I	-0.6 ± 0.6	10	1.1 ± 0.3	38.8 ± 0.9	3.5 ± 0.1	4.3 ± 0.4	0.8 ± 0.1
P203A	2.9 ± 1.3	8	0.2 ± 0.1	40.8 ± 0.2	5.0 ± 0.2	6.7 ± 0.8	0.7 ± 0.1
P203S	0.7 ± 0.5	8	1.6 ± 0.4	43.7 ± 1.7	2.5 ± 0.1	3.2 ± 0.2	0.8 ± 0.1
T193A	0.1 ± 0.4	8	0.8 ± 0.3	33.7 ± 0.5	2.6 ± 0.1	2.0 ± 0.2	1.3 ± 0.2
K38E	0.0 ± 0.7	5	0.6 ± 0.1	37.7 ± 1.1	3.2 ± 0.1	6.3 ± 0.6	0.5 ± 0.05

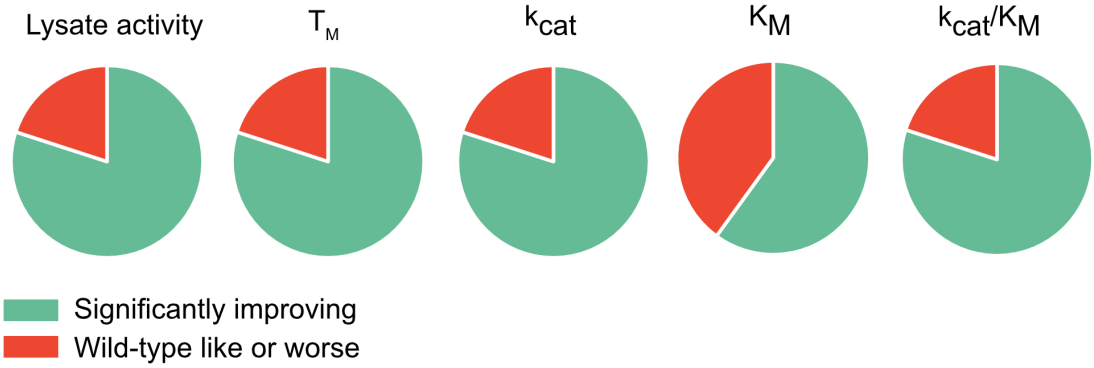
D69H	0.5 ± 2.4	0	3.3 ± 1.0	39.5 ± 0.6	4.1 ± 0.2	5.3 ± 0.6	0.8 ± 0.1
L174V	1.1 ± 1.4	2	3.8 ± 1.2	27.6 ± 0.2	10.7 ± 0.5	13.8 ± 1.5	0.8 ± 0.1
T241A P203A	-		1.5 ± 0.3	50.6 ± 0.4	15.0 ± 0.4	2.3 ± 0.2	6.6 ± 0.6
T241A D69H	-		8.7 ± 1.7	47.0 ± 0.6	12.1 ± 0.3	2.4 ± 0.2	5.1 ± 0.5
T241A K38E	-		3.1 ± 0.8	50.8 ± 0.4	11.3 ± 0.3	1.9 ± 0.2	6.1 ± 0.6
T241A L174V	-		2.8 ± 0.2	37.2 ± 0.2	12.7 ± 0.3	5.3 ± 0.4	2.4 ± 0.2
V216I D69H	-		3.3 ± 0.7	37.8 ± 0.1	5.4 ± 0.2	4.3 ± 0.5	1.3 ± 0.2
V216I K38E	-		1.0 ± 0.3	37.5 ± 0.8	5.2 ± 0.2	5.3 ± 0.7	1.0 ± 0.1
K38E D69H	-		0.3 ± 0.01	36.7 ± 0.1	4.3 ± 0.3	12.7 ± 2.0	0.3 ± 0.1
T241A K38E D69H	-		1.4 ± 0.1	50.2 ± 0.9	24.5 ± 1.0	7.1 ± 0.8	3.4 ± 0.4
T241A V216I	-		2.9 ± 0.7	48.8 ± 0.6	7.7 ± 0.2	4.9 ± 0.4	1.6 ± 0.1

Extended Data Figure 6: Combinability hotspots. (A) Positional frequency in highly active and combinable variants. Amino acid identity of mutations is colour-coded (blues: non-polar aliphatic, brown: aromatic, greens: polar, uncharged, reds: positively charged, violets: negatively charged). T241 is the global hotspots of combinability, followed by V216. **(B)** *Sr*IRED structure (monomers indicated in grey and orange; PDB ID: 5ocm) with hotspots represented as spheres identifies them in the C-terminal domain and first shell. **(C-F)** Combinability graphs showing combinable interactions of mutations at distinct positions (CTD = C-terminal domain; NTD = N-terminal domain). Nodes represent positions in the IRED sequence that are grouped by distance in the IRED structure. Connections represent cooccurrence in a higher-order variant whose corresponding fitness is larger or equal to the combined effect in single mutations (i.e. combinability). The high node degree of T241 (edges in red) indicates high potential for combinability and positive epistasis with all parts of the protein while interactions of P203, V216 and T193 are more focused on the N-terminal domain. **(G)** Lysate activity was measured as initial rates in three replicates (errors represent standard deviations) with **1** (10 mM), **a** (20 mM), NADPH (0.5 mM) in Tris buffer (100 mM pH 8.0) following the decline of NADPH absorbance at 340 nm and normalized to lysate without IRED. Soluble expression was determined by densitometry of SDS-PAGE bands of the soluble and insoluble fraction after lysis. Values are normalized to wild type. Melting temperatures were measured in thermal shift assays in three replicates, errors represent standard deviations. Michaelis Menten kinetics were conducted with variable concentrations of ketone substrate (0.05 - 35 mM) and a constant concentration of amine (30 mM) in Tris pH 8 with NADPH (0.5 mM) and variable enzyme concentrations (from 0.03 μM to 0.3 μM) depending on the rate at T = 25 °C. Each curve was measured in three replicates. Errors represent standard errors of the fit and three technical replicates with a 95% confidence interval.

A	Rank AI model	Lysate activity [fold change]	T_M [°C]	k_{cat} [s^{-1}]	K_M [mM]	k_{cat}/K_M [$mM^{-1}s^{-1}$]
wt	n.a.	1	38.5 ± 0.6	2.1 ± 0.1	3.5 ± 0.3	0.6 ± 0.1
P203G	1	0.5 ± 0.2	39.7 ± 0.3	2.8 ± 0.1	5.1 ± 0.6	0.5 ± 0.1
T241G	2	3.2 ± 0.4	45.8 ± 0.6	11.7 ± 0.4	3.0 ± 0.3	3.9 ± 0.4
M1P	3	0.7 ± 0.03	36.7 ± 0.4	3.6 ± 0.1	5.1 ± 0.5	0.7 ± 0.1
A251I	4	0	n.d.	n.d.	n.d.	n.d.
E146L	5	0.4 ± 0.1	38.5 ± 0.9	2.1 ± 0.1	2.9 ± 0.3	0.7 ± 0.1

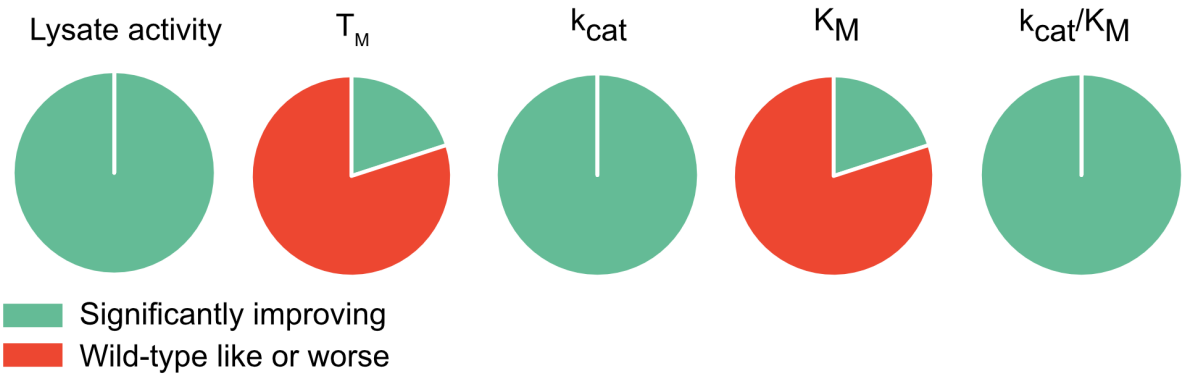


B	Rank AI model	Lysate activity [fold change]	T_M [°C]	k_{cat} [s^{-1}]	K_M [mM]	k_{cat}/K_M [$mM^{-1}s^{-1}$]
wt	n.a.	1	38.5 ± 0.6	2.1 ± 0.1	3.5 ± 0.3	0.6 ± 0.1
S35T S51G	1	0.7 ± 0.1	40.1 ± 1.0	1.9 ± 0.05	3.6 ± 0.3	0.6 ± 0.04
Q41P T241A	2	1.7 ± 0.1	50.0 ± 0.7	6.4 ± 0.2	1.6 ± 0.1	4.0 ± 0.4
E36A T241A	3	4.2 ± 1.1	49.2 ± 0.1	9.4 ± 0.4	2.1 ± 0.3	4.4 ± 0.7
S35T T241A	4	2.1 ± 0.3	51.3 ± 0.1	3.3 ± 0.1	1.3 ± 0.1	2.6 ± 0.2
Q41P V262L	5	2.1 ± 0.4	29.4 ± 0.4	3.6 ± 0.1	4.4 ± 0.5	0.8 ± 0.1

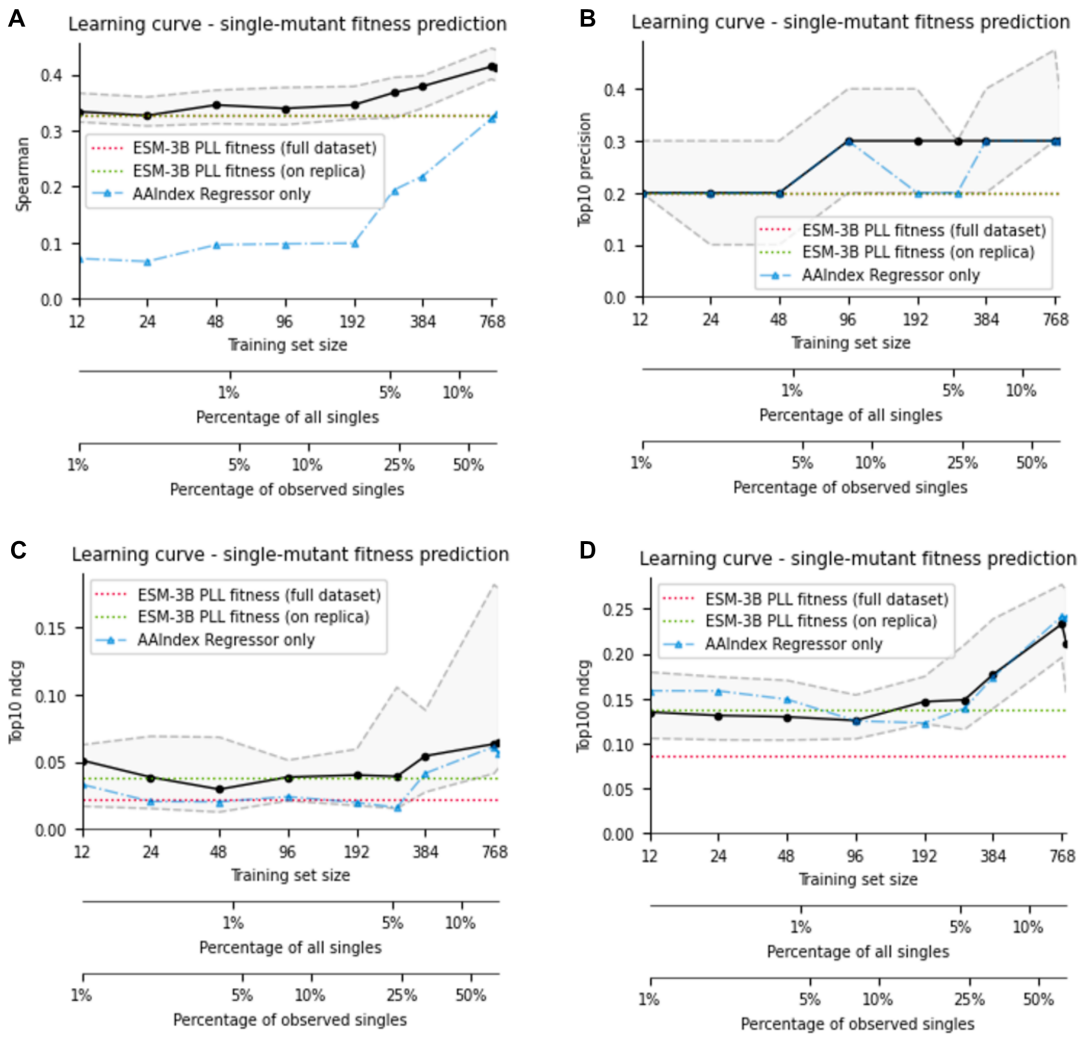


C	Lysate activity [fold change]	T_M [°C]	k_{cat} [s^{-1}]	K_M [mM]	k_{cat}/K_M [$mM^{-1}s^{-1}$]
wt	1	38.5 ± 0.6	2.1 ± 0.1	3.5 ± 0.3	0.6 ± 0.1
T241G K38E	6.4 ± 1.4	37.7 ± 1.0	10.2 ± 0.6	3.9 ± 0.7	2.6 ± 0.5
T241G D69H	9.7 ± 1.3	36.2 ± 1.9	28.6 ± 0.9	4.1 ± 0.4	6.9 ± 0.7
T241G L174V	6.7 ± 0.6	29.8 ± 0.6	22.7 ± 0.8	6.2 ± 0.6	3.6 ± 0.4
T241G P203A	7.5 ± 1.2	39.3 ± 1.2	49.3 ± 1.5	5.1 ± 0.4	9.7 ± 0.9
E36A T241G	3.0 ± 0.2	45.6 ± 0.5	14.0 ± 0.3	2.5 ± 0.2	5.7 ± 0.5

n.d …. not determined

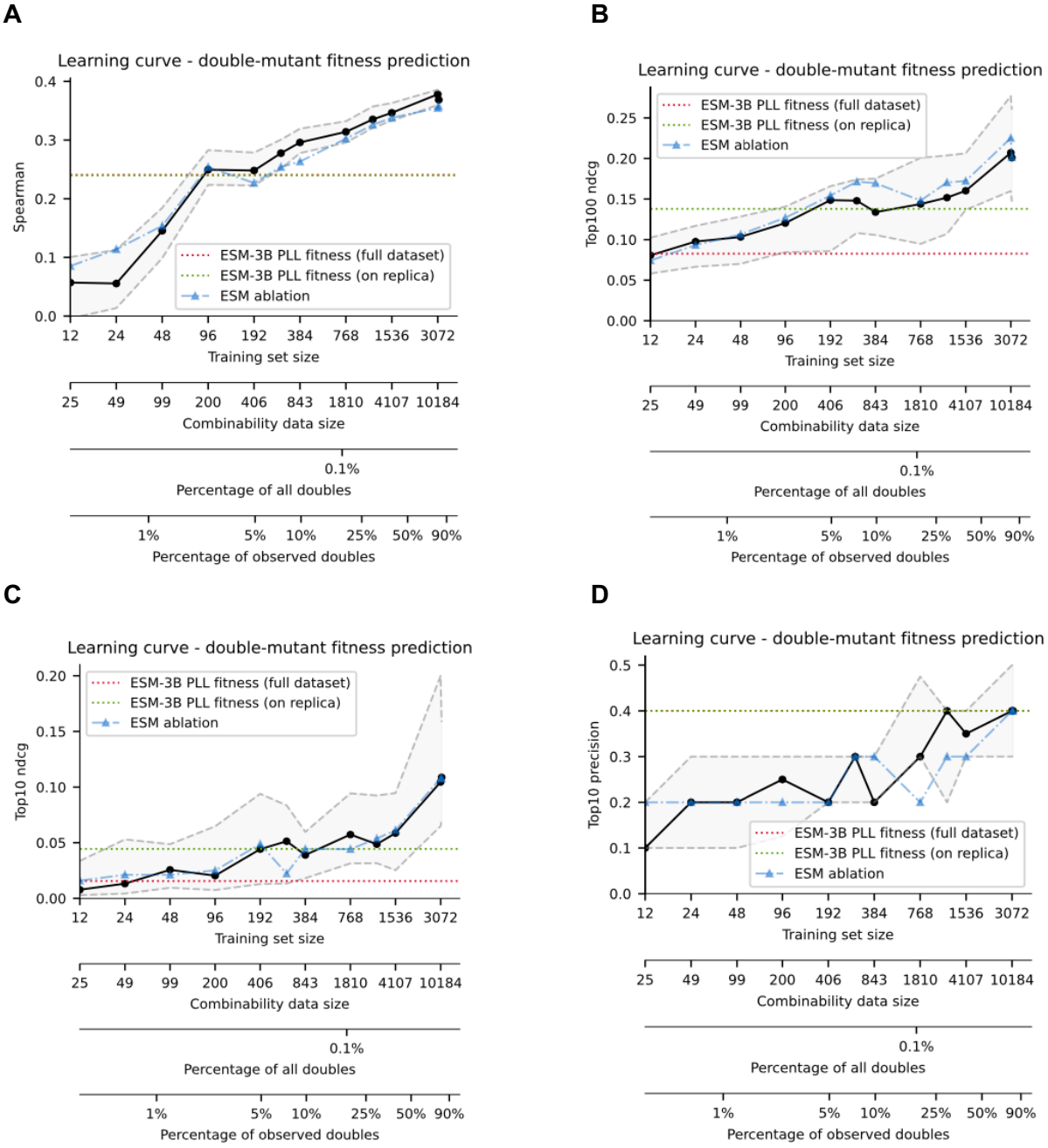


Extended Data Figure 7: Characterization of AI-predicted and hybrid variants (**A**) Lysate activity was measured as initial rates in three replicates (errors represent standard deviations) with **1** (10 mM), **2** (20 mM), NADPH (0.5 mM) in 100 mM Tris pH 8.0 following the decline of NADPH absorbance at 340 nm and normalized to lysate without IRED. Soluble expression was determined by densitometry of SDS-PAGE bands of the soluble and insoluble fraction after lysis. Values are normalized to wild type. Melting temperatures were measured in thermal shift assays in three replicates, errors represent standard deviations. Michaelis Menten kinetics were conducted with variable concentrations of ketone substrate (0.05 - 35 mM) and a constant concentration of amine (30 mM) in Tris pH 8 (T = 25°C) with NADPH (0.5 mM) and 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.



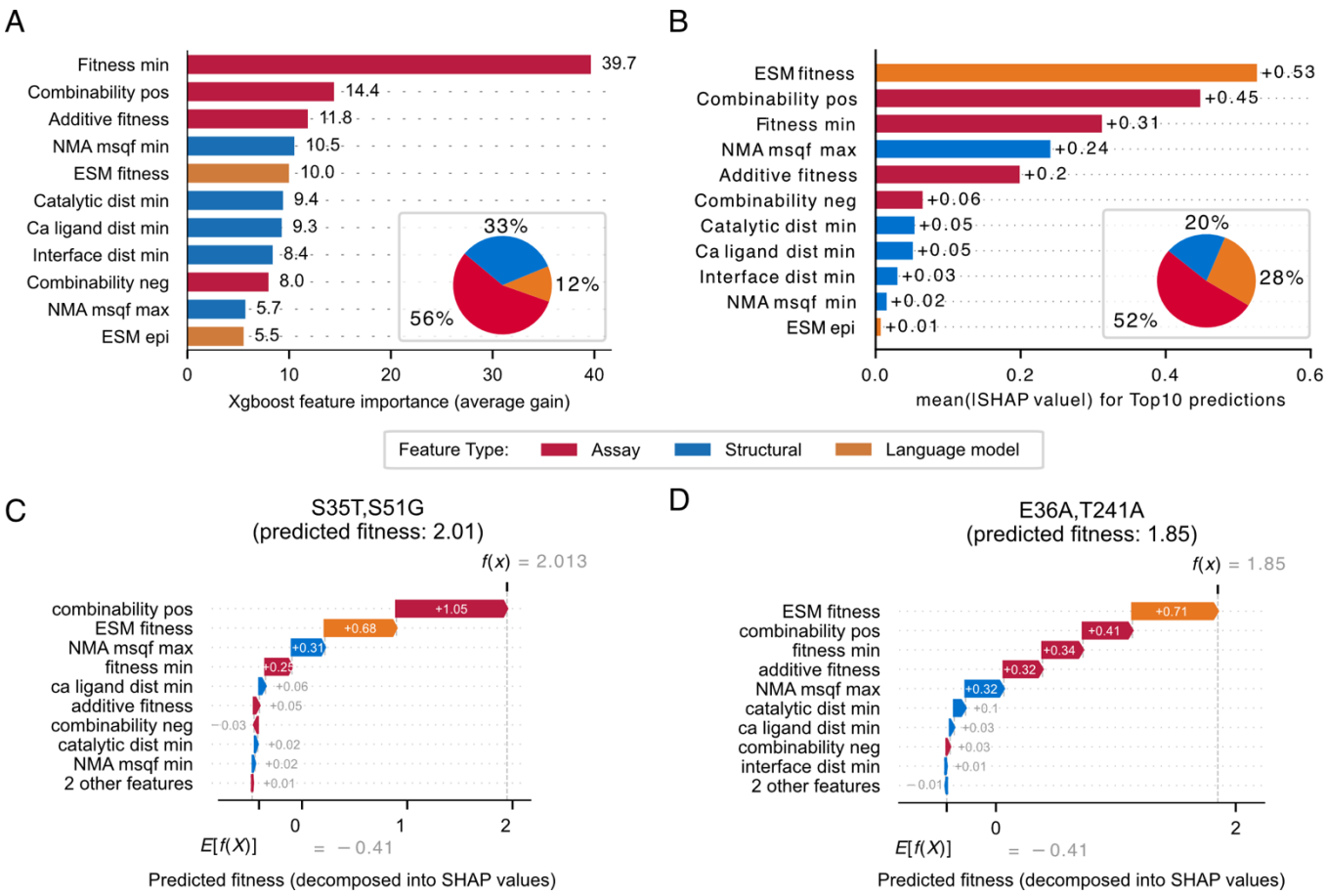
Extended Data Figure 8: Single mutant model learning curves for 4 key metrics: Spearman, Top10 precision and NDCG and Top 100 NDCG. Learning curves for the single mutant model and several ablations are computed on 10 replicas, using 3-fold cross validation with a different random seed per replica, resulting in 30 experiments. The black line indicates the single mutant model's median across all experiments, while the top/bottom dashed gray line indicates the 75th and 25th percentile – corresponding to a 'good' and 'bad' case respectively. The blue line ablates the ESM features and shows the median performance of the assay-trained linear regression model based on AAIndex embeddings alone. The red and green lines respectively indicate ESM PLL zero-shot performance on all single mutants and the median across the 30 folds used to test the other models, for fair comparison in the presence of batch effects. **(A)** Spearman correlation. For low dataset sizes (<200) the model's ranking performance is dominated by the performance of ESM, as expected: By construction the single mutant model treats each position independently and can therefore not generalize to unseen positions. From dataset sizes above ~200 (*Sr*IRED is 290 AA long) the model has on average observed 2 amino acids (wild type and a mutant) per position and the droplet data starts to clearly improve global ranking performance. The overall Spearman (~0.41) is modest, but not uncommon for noisy biological datasets[24,46]. Note that while Spearman is a useful measure of overall model performance, it poorly reflects the performance on the few top mutants, which are most relevant for engineering. **(B)** Top 10 precision. The model's hit-rate for discerning variants better than wild type (WT) (fitness > 0) among the top 10 variants does not significantly improve with increasing lrDMS training dataset size apart from the uptick of the 25th percentile. On average, 2-3 of the top 10 predicted variants are consistently better than wildtype, which is comparable to the hit-rate of ESM-PLL ranking (2). This is likely due to the limited resolution of lrDMS data for variants of similar activity, and because most single mutants are of WT-like fitness (c.f. **Figure 2B**, single-mutants are strongly peaked around WT-

fitness compared to higher order mutants). Instead, lrDMS data contains a more coarse-grained signal for when a variant is significantly better or worse than WT. As a result, if the test data for a given replica only contains mediocre fitness variants, the model is unlikely to be able to say with confidence if this variant is slightly better or worse than WT, limiting precision improvements. (**C-D)** Top 10 / Top 100 Normalised discounted cumulative gain (NDCG). The NDCG computes the quality of recommendations by encompassing both their relevance (fitness) and their rank in the results (ranking). The restriction to the top 10/top 100 predictions further focuses the NDCG on the top few predictions which are most relevant for protein engineering at low test budgets. Similarly to the Spearman correlation, for small training dataset sizes below the sequence length (<290) the Top 10/100 NDCG are mostly constant. For training dataset sizes higher than the sequence length, NDCG starts to improve. Note that not only the median, but also the 25th percentile ("bad luck scenario") exceeds the both zero-shot baselines. Overall, the NDCG values clearly show improvements for larger dataset sizes. Similar magnitudes for NDCG are not uncommon for noisy biological datasets, and the low NDCG value achieved by ESM points to the single mutant data in this dataset being a more challenging than most others in ProteinGym[24].
Further detailed cross-validation analysis of the single mutant model is available in **Supplementary Figure S20.**



Extended Data Figure 9: Double mutant model learning curves for 4 key metrics: Spearman, Top10 precision and NDCG and Top 100 NDCG. Learning curves for the double mutant model and several ablations were computed on 6 replicas using 5-fold cross validation with a different random seed per replica, resulting in 30 overall experiments. The black line indicates the double mutant model's median performance across all experiments, while the top/bottom dashed gray line indicates the 75th and 25th percentile – corresponding to a 'good' and 'bad' case respectively. The blue line ablates the ESM features, which are the most expensive to compute, and demonstrate that our model is performant even when omitting ESM features, e.g. when no GPUs but only a table-top computer is available. It also demonstrates that most useful information is carried by the lrDMS data itself, rather than the language model. The red and green lines respectively indicate ESM PLL zero-shot performance on all single mutants and the median across the 30 folds used to test the other models, for fair comparison in the presence of batch effects. The two x-axes ("Training set size" & "Combinability data size") correspond to the number of double mutants available for training the model and the number of higher order mutants available to inform combinability features respectively. **(A)** Spearman correlation. The model's ranking performance consistently increases with

increasing dataset size. Notably, even the 25th percentile outperforms the zero-shot baseslines as soon as about 300 datapoints are available. **(B-C)** Top 100 and Top 10 Normalized discounted cumulative gain (NDCG). The NDCG computes the quality of recommendations by encompassing both their relevance (fitness) and their rank in the results (ranking). The restriction to the top 10/top 100 predictions further focuses the NDCG on the top few predictions which are most relevant for protein engineering at low test budgets. **(D)** Top 10 precision. The model's hit-rate for discerning variants better than wild type (WT) (fitness > 0) among the top 10 variants also shows a consistent uptick as dataset size is increased. It is the only metric for which the "bad case" (25th percentile) is not consistently better than the zero-shot ESM performance. Similarly to the single mutant case, we attribute this behaviour to the limited resolution of lrDMS assay data for variants of similar fitness but note that the double mutant case is less severe than in the single mutant case.
For all metrics we observe significant improvements in the quality of the model with increasing amount of training data and observe that ablating ESM features appears not to affect performance. Further detailed cross-validation analysis of the double mutant model is available in **Supplementary Figure S21**.



Extended Data Figure 10: Evaluation of the influence of different features in the double mutant model. Irrespective of which feature attribution method is used, the assay derived features consistently have the highest influence on the model's predictions. We investigate the extent to which different features influence the double mutant model that was used for hit combination with AI. (**A**) The features are ranked by their average gain[75] across all splits in which that feature is used within the decision trees constructed by xgboost. Loosely, this gain corresponds to the extent of improvement in training error when the training data is split at a threshold value of the given feature. The so-derived 'gain'-feature importance values indicate that assay derived features (fitness min, combinability and additive fitness) have the strongest impact for enhancing model performance during training. (**B**) To assess each feature's the impact on our top predictions, we plot the absolute SHAP values[48] for each feature, averaged across the top 10 predicted mutants. SHAP values are a well-studied, alternative way of assigning feature importances based on cooperative game theory and provide insight into the contribution of each feature to the prediction for an individual instance. As in (A), assay derived features are responsible the largest contribution to the predicted scores of the top10 mutants, followed by language model features and structural features. (**C**) and (**D**) show a detailed breakdown for double mutant predicted to have the highest score (C) and the model-predicted double mutant with the highest improvement in post-hoc characterization (D). Breakdowns for the other predicted variants and a more detailed SHAP analysis are available in **Supplementary Figures S18 and S19**. Note that Combinability, NMA, Catalytic site distance and Ligand distance features only influence the position of the chosen mutant. The amino acid identity is purely informed via the fitness min, additive fitness, and ESM fitness features, therefore at least one of these three features must give significant input to select the amino acid identity at the selected position, which may overweight the importance of these features over the position-only features.