SUBSTRATE-AWARE ZERO-SHOT PREDICTORS FOR NON-NATIVE ENZYME ACTIVITIES

Francesca-Zhoufan Li*, **Lukas A. Radtke, Kadina E. Johnston, Cheng-Hao Liu**, **Yisong Yue**, **Frances H. Arnold***
California Institute of Technology, Pasadena, CA 91125, USA

ABSTRACT

Enzymes can be engineered to catalyze reactions with non-native substrates or even perform entirely new reactions unknown in nature. However, developing such novel activities through wet-lab engineering is time- and resource-intensive. By estimating enzyme activity without new experimental data, zero-shot (ZS) predictors can accelerate enzyme engineering. While ZS predictors have been demonstrated in various contexts, they have yet to be evaluated on non-native substrates and *new-to-nature* chemistry. Critically, many existing predictors do not explicitly encode substrate or transition-state properties, which we propose are essential for predicting *new-to-nature* chemistry. Here, we systematically studied two types of mechanistically distinct enzymes using 16 ZS predictors—including six general and ten substrate-aware scores derived from generative modeling, molecular docking, and active-site heuristics. We curated new experimental and literature mutation datasets spanning 11 non-native substrates and three *new-to-nature* reactions with 11 additional substrates. The six general ZS predictors could generalize to non-native substrates, but failed for *new-to-nature* chemistries. In contrast, certain substrate-aware approaches could predict *new-to-nature* chemistries, with AlphaFold 3's chain-predicted aligned error being the most predictive of both activity and stereoselectivity. A weighted ensemble of AlphaFold 3 and EVmutation scores generalized to all chemistries that we tested. Our findings highlight the potential of ZS predictors to accelerate enzyme engineering, advancing the expansion of the chemical universe beyond nature's repertoire.

1 INTRODUCTION

Enzymes, nature's catalysts, perform life-sustaining chemistry. Due to their exquisite specificity and selectivity, engineered enzymes have applications in therapeutics, bioremediation, and biocatalysis, where they can offer greener and more sustainable alternatives to conventional chemical methods (Buller et al., 2023; Lutz & Iamurri, 2018). Beyond merely enhancing their native functions, significant efforts have focused on expanding enzyme repertoires to catalyze reactions with non-native substrates—or even perform entirely new chemistries unknown in biological systems, termed new-to-nature (Arnold, 2017; Renata et al., 2015; Chen & Arnold, 2020; Bell et al., 2021). The development of such enzymes often starts by identifying a promiscuous or side activity, which is then optimized for desired functions (termed "fitness", Figure 1a) (O'Brien & Herschlag, 1999; Leveson-Gower et al., 2019). Fitness optimization typically employs directed evolution, a widely used method for accumulating beneficial mutations through iterative rounds of mutagenesis (to generate variants) and functional assessment by selection or screening (Packer & Liu, 2015; Wang et al., 2021). However, this process is labor- and resource-intensive, particularly for challenging chemistries constrained by low-throughput data collection.

Emerging computational tools, especially machine learning (ML)-based methods, have shown promise in accelerating enzyme engineering, from starting point discovery, to *de novo* designs and fitness optimization (Yang et al., 2024; Mak et al., 2015; Siegel et al., 2010; Kalvet et al., 2023; Ding et al., 2024b; Yang et al., 2025; Thomas et al., 2024; Rapp et al., 2024). A particularly appealing avenue is zero-shot (ZS) prediction: estimating variant fitness without relying on experimental

*Correspondence to: fzl@caltech.edu, frances@cheme.caltech.edu

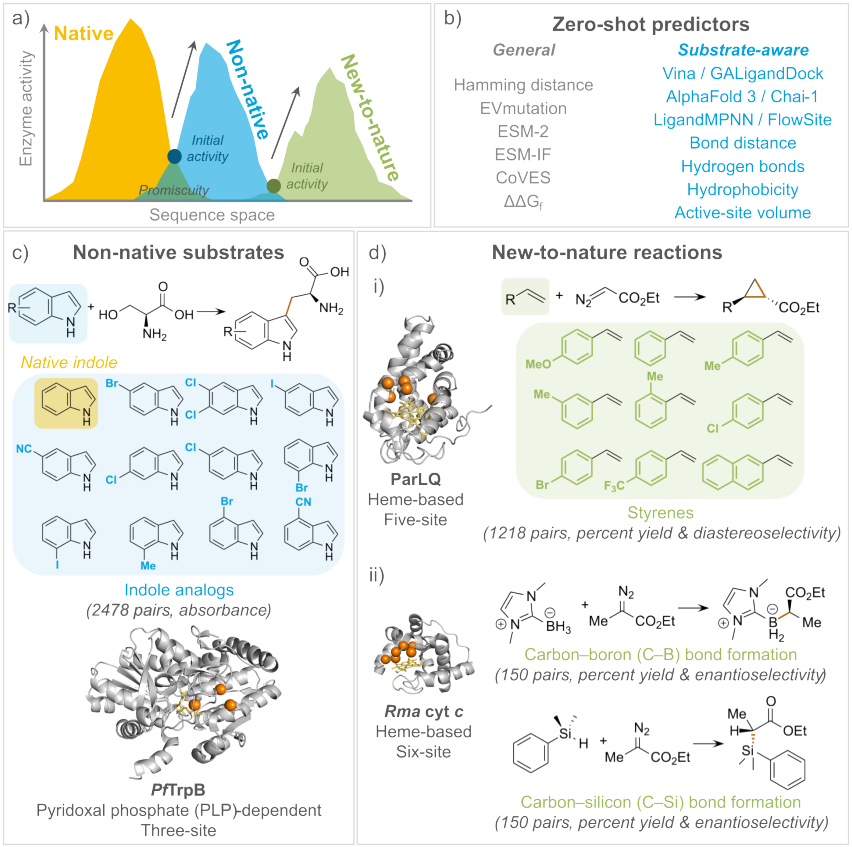

Figure 1: Overview of datasets and zero-shot (ZS) predictors. a) Enzyme engineering for reactions with non-native substrates and new-to-nature chemistries. b) Six general ZS predictors, covering different modalities and auxiliary information, alongside ten substrate-aware predictors derived from generative modeling, molecular docking, and active-site properties. c) The *Pf*TrpB reaction with the native substrate, indole, in a pyridoxal phosphate (PLP)-dependent manner, along with 11 non-native indole analogs. d) Heme-based new-to-nature carbene transfer reactions with activity and stereoselectivity labels: i) ParLQ: protoglobin-catalyzed olefin cyclopropanation with 9 styrenes and ii) *Rma* cyt *c*-mediated formation of C–B (*Rma*-CB) and C–Si (*Rma*-CSi) bonds. Crystal structures illustrate the cofactors and active-site residues targeted for engineering. The number of unique enzyme-substrate pairs is listed under each reaction.

data. ZS predictors leverage auxiliary information such as protein stability, evolutionary patterns, or structural features. These predictors have augmented supervised models to identify higher-fitness variants, guided experimental data collection for ML model training, and scored *in silico* designs for reinforcement learning or experimental validation (Wittmann et al., 2021; Li et al., 2024; Hsu et al., 2022a; Landwehr et al., 2025; Johnson et al., 2024; Stocco et al., 2024). Recent benchmarks highlight the broad applicability of ZS predictors Notin et al. (2023). However, these methods have yet to demonstrate capability across diverse enzyme-substrate pairs, particularly for out-of-distribution non-native substrates and new-to-nature chemistries. Few studies also assess the ZS predictability of the reaction product stereoselectivity, a key factor influencing their structural and functional properties (Reisenbauer et al., 2024). Furthermore, existing enzyme-substrate datasets primarily focus on native or near-native activities and rarely address active-site mutations that are critical to enabling fundamentally new chemistries (Goldman et al., 2022; Paton et al., 2024). Most importantly, many existing ZS methods do not explicitly encode substrate or transition-state properties, which we hypothesize are essential for predicting new-to-nature chemistry.

Here, we curated new experimental and literature datasets for two types of mechanistically distinct enzymes and benchmarked six general and ten newly proposed substrate-aware ZS predictors. Specifically, we evaluated their performance across 11 non-native substrates and three new-to-nature chemistries, covering 11 additional substrates (Figure 1). Each dataset includes active-site mutations designed to enhance specificity and selectivity, and introduce new chemistries (Bell et al., 2021). Our study addresses three key questions: 1) Do general ZS predictors generalize

to non-native substrates and new-to-nature chemistries for both activity and stereoselectivity? 2) Are novel substrate-aware ZS scores, derived from generative modeling, molecular docking, and active-site properties, more generalizable than the general predictors? 3) Which combination of ZS predictors can best generalize across these chemistries?

2 METHODS AND DATASETS

We examined the predictability of ZS predictors on activity, defined by absorbance or percent yield of the major product. Where applicable, we also studied stereoselectivity, defined as enantiomeric excess of the desired chiral isomer or diastereomeric excess of the desired diastereomer. We analyzed the effects of mutations in key active-site residues across two types of mechanistically distinct enzymes: *Pf*TrpB, which catalyzes reactions with 11 non-native substrates, and heme-binding proteins (protoglobin and *Rma* cyt *c*), which catalyze three different new-to-nature chemistries (ParLQ, *Rma*-CB, *Rma*-CSi). In ParLQ, we further examined the effects of 9 different substrates. The *Pf*TrpB dataset is presented for the first time, while the rest were previously reported (Ding et al., 2024b; Yang et al., 2025). The 11 *Pf*TrpB datasets and ParLQ-a contain more low-activity variants than the other datasets (Appendix A.4.1). The similarity of the non-native substrate to the native substrate was calculated using the Tanimoto similarity of atom-pair fingerprints (Carhart et al., 1985). The energy barriers of new-to-nature chemistries were determined using density functional theory (DFT) calculations. We tested predictor ensembles using families of linear models. See Appendix A.1 and A.3 for more details. Spearman's correlation is reported in the main text. For recall (true positives of the top 25% ranked variants) and additional results, see Appendix A.4.2 and A.4.3. Data supporting this study are deposited on Zenodo: `https://zenodo.org/records/15226690`. The code is available on GitHub: `https://github.com/fhalab/substrate_aware_zs`.

3 RESULTS

3.1 GENERAL ZS PREDICTORS ARE PREDICTIVE OF ACTIVITIES ON NON-NATIVE SUBSTRATES BUT DO NOT GENERALIZE TO NEW-TO-NATURE CHEMISTRIES

As baselines, we evaluated six general ZS predictors for variant activity scoring. Each ZS predictor leveraged distinct auxiliary information (Figure 2a). Hamming distance assumes that most mutations are deleterious (Romero & Arnold, 2009; Arnold, 2017). It counts the number of amino acid substitutions from the parent, a variant with initial activity, aiming to perform a local search for viable variants. Given similar reaction mechanisms and conserved catalytic residues, we expected the assumption to hold true for substrates similar to their native counterparts. However, new-to-nature chemistry may require exploring a more diverse sequence space. Indeed, we observed a weak positive correlation for non-native activities ($\rho \sim 0.3$), but very little to weak negative correlation for the new-to-nature reactions ($\rho \sim -0.2 - 0.1$).

EVmutation and ESM-2 estimate mutation effects using evolutionary patterns, either through a Potts model applied to multiple-sequence alignments (MSAs) or a mask-prediction protein language models (Hopf et al., 2017; Meier et al., 2021). Both predictors generalized well to non-native substrates ($\rho \sim 0.5$) but not to new-to-nature chemistries ($\rho \sim -0.2 - 0.2$). Notably, their predictability decreased with shallower MSAs (Table A6) and was statistically correlated with substrate similarity to the native substrate ($p < 0.05$, Figure 2c). ESM-IF and CoVES incorporate structural context for ZS predictions. ESM-IF assigns residue likelihoods based on a backbone structure (inverse folding) and CoVES predicts masked residues based on their local atomic environments (Hsu et al., 2022b; Ding et al., 2024a). However, neither predictor outperformed EVmutation or ESM-2, though CoVES scores exhibited a weaker correlation with substrate similarity to the native substrate (Table A10).

Stability is crucial for protein function, as misfolded proteins will be less likely to retain activity (Bloom et al., 2006; Ding et al., 2024b). We estimated variant stability by calculating the change in its free energy of folding relative to the parent ($\Delta\Delta G_f$) using a Rosetta energy function (Wittmann et al., 2021). However, stability did not predict new-to-nature activities ($\rho \sim -0.1 - 0.1$). We reason that once a minimal stability threshold is met, factors such as substrate recognition and transition-state stabilization become the dominant determinants of activity. Moreover, excessive

stability may limit the structural flexibility needed to accommodate new substrates or reaction mechanisms (Teufl et al., 2022).

For the new-to-nature chemistries, we also evaluated the stereoselectivity of the major products (Figure 2b). For the general ZS predictors, activity and stereoselectivity predictions were generally correlated ($\rho \sim 0.68 - 0.95$, Table A7).The chemical mechanisms for the non-native substrates in this study are conserved, thus we postulate that this makes the predictions easier to generalize (Figure A1). In contrast, new-to-nature chemistry involves mechanisms distinct from an enzyme's native chemistry and may require beneficial mutations that are rare in MSAs or occur at conserved residues, thus demanding deeper structural insights into the substrate and/or the active site.

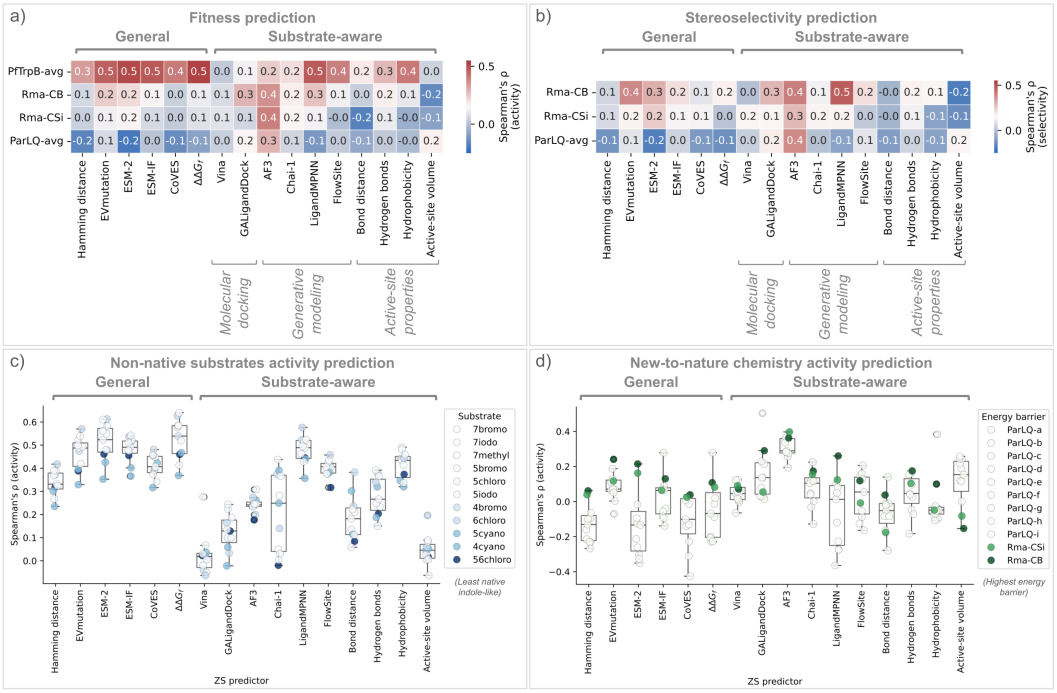

Figure 2: General and substrate-aware ZS predictors. Spearman's ρ for a) activity, b) stereoselectivity, c) *Pf*TrpB non-native substrate activities (Table A10), and d) heme-based new-to-nature activities (Table A11).

3.2 Substrate-aware predictors offer insights for new-to-nature chemistries

Enzyme catalysis involves complex mechanistic steps, including substrate binding and transition-state stabilization. We hypothesize that substrate-aware ZS predictors can better describe substrate recognition and transition-state stabilization for more diverse molecular systems, enhancing the predictability of new-to-nature chemistries. To capture the full enzyme-substrate-cofactor interaction, we considered the cofactors in their catalytically relevant states for each ZS predictor (AppendixA.2).

We first explored enzyme-substrate binding energy as a potential ZS score using physics-based molecular docking with GALigandDock and AutoDock Vina (Park et al., 2021; Eberhardt et al., 2021). Both methods had weak to no correlation on non-native substrates ($\rho \sim 0 - 0.3$), but we noted GALigandDock was slightly better ($\Delta\rho \sim 0.1 - 0.2$) than AutoDock Vina for new-to-nature chemistries (Figure 2a). Correlations of both scores were independent of substrates (Table A10).

Recent advances in generative modeling have significantly advanced molecular docking and structure prediction (Yim et al., 2024; Abramson et al., 2024; Discovery et al., 2024). We hypothesized that the scores pertaining to enzyme-substrate/cofactor binding may predict interactions impacting activity. While not outperforming general ZS predictors for non-native substrates, AlphaFold 3 (AF3) and Chai-1's interface predicted TM-scores (iPTM) for the enzyme-cofactor were predictive of new-to-nature activities (Figure 2a). Interestingly, AF3's chain-predicted aligned error (PAE) for the enzyme-cofactor interaction was the most predictive for both activity and stereoselectivity ($\rho \sim$

0.4), independent of the substrate (Figure 2). In contrast, despite adopting similar algorithmic approaches but without MSAs or templates, Chai-1 exhibited lower predictability, particularly for substrates more dissimilar to the native one (Figure 2c, Table A10). While MSA quality may impact prediction accuracy, further investigation is needed. Generative models can also facilitate binding site design via substrate-aware inverse folding or simultaneous docking and backbone redesign (Krapp et al., 2024; Watson et al., 2023; Dauparas et al., 2023; Stärk et al., 2023). We studied using probability scores from variant generation, conditioned on the parent structure, as a ZS predictor. LigandMPNN and FlowSite were predictive of non-native reactions, performing comparably to existing predictors like ESM-2, but were less effective for new-to-nature chemistries other than *Rma*-CB.

Beyond docking scores, we reasoned that a docked pose can be distilled into key components that reflect enzyme-substrate interactions. We hypothesized that bond-forming atom proximity could indicate higher activity–for instance, the distance between Glu104 and N1-hydrogen in *Pf*TrpB, or between the carbene carbon and boron, silicon, or the styrene double bond (Appendix A.2). However, bond distance was a poor predictor, likely due to noise in docking poses. Stabilization forces, particularly hydrogen bonding, can lower reaction energy barriers in the enzyme's active site (Shan & Herschlag, 1996). In *Pf*TrpB, the number of active-site hydrogen bonds correlated with activity, though it was less evident for heme-based new-to-nature chemistries. Instead, the highly reactive carbene intermediate would be stabilized by the iron in the heme (Chaturvedi et al., 2024). The combined hydrophobicity of the substrate and active site affects their interaction, with optimal binding occurring when their hydrophobicity levels match (Estell et al., 1986; Sriramulu & Lee, 2020). This offered some predictive power for *Pf*TrpB, but not for heme-based chemistries. Lastly, active-site volume has been linked to enzyme promiscuity (Martínez-Martínez et al., 2017), but it showed little predictive power for non-native or new-to-nature chemistries. We reasoned this may only exclude overly large residues, while failing to account for exposed active sites in *Rma* cty *c* or the substrate tunnel in ParLQ (Danelius et al., 2023). Consistent with general ZS predictors, the predictability for activity in substrate-aware predictors generally correlated with its predictability for stereoselectivity ($\rho \sim 0.31 - 0.95$, Table A7). AF3 remained the best predictor for new-to-nature chemistries.

3.3 Predictor ensembles improve generalization across chemistries

We explored ensemble methods to improve generalization and found that many model and predictor combinations outperformed individual predictors (Table A5). The unweighted ensemble of EVmutation and AF3—the top two individual predictors averaged across all chemistries—was consistently predictive ($\rho \sim 0.3 - 0.5$). A learned linear combination of them achieved an average ρ of 0.39 across all chemistries in the test set (Figure 3). This generalization remained robust regardless the training dataset (Figure A17). EVmutation, ESM-IF and AF3 had the highest regression weights when averaged across all chemistries (Figure A16). However, incorporating ESM-IF, the third-best predictors averaged across all chemistries, into any combination did not further improve the generalization (Figure A15). Interestingly, EVmutation was one of the top predictors alongside AF3 in the top 25% recall analysis, whereas ESM-IF was not (Figure A10).

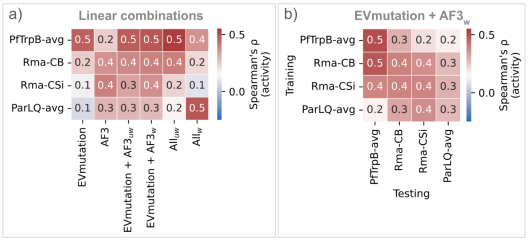

Figure 3: Predictor ensembles. a) Different linear combinations of ZS predictors. w represents weighted linear combination trained on the *Rma*-CB dataset and tested on the rest. uw refers to an unweighted combination of EVmutation and AF3 rankings for each dataset. b) Linear models were trained on EVmutation and AF3 score for each dataset and tested on all the datasets.

4 Discussion

We evaluated six general and ten substrate-aware ZS predictors using two types of enzymes with distinct mechanisms across 22 different substrates and four types of chemistries. General ZS predictors were effective for non-native substrates but failed for new-to-nature chemistries. Among substrate-aware ZS predictors, AF3 was the best for both activity and stereoselectivity prediction in new-to-nature chemistries. A linear combination of AF3 and EVmutation generalized across all studied reactions, which could complement current protein design pipelines that employ a series of

logical filtering steps (Bennett et al., 2023; Johnson et al., 2024). Physics-based docking methods and simple active-site heuristics did not consistently capture enzyme reactivity.

Enzyme engineering for new reactivities remains an out-of-distribution challenge, constrained by limited sequence-activity data. Although we generated new experimental data and curated literature datasets, their scope remains limited, especially since the literature datasets had largely active variants, which would not resemble distributions with mostly inactive variants. While the approaches studied here generalized well in active-site mutation datasets, expanding to more diverse new-to-nature reactions and testing datasets with mutations outside the active site remains a priority. With our growing ability to collect sequence-activity data (Long et al., 2024; Wittmann et al., 2022), more comprehensive datasets will be curated, and systematic benchmarks will be conducted. We are optimistic that increasingly generalizable substrate-aware ZS predictors will accelerate enzyme engineering, unlocking entirely new realms of biocatalysis.

ACKNOWLEDGMENTS

The authors thank Ariane Mora, Yueming Long, Edwin Alfonzo, Ravi Lal, Theophile Lambert, and Jason Yang for their critical reading and discussion of the manuscript, and Patrick Almhjell for assistance with the TrpB methods.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Melissa F Adasme, Katja L Linnemann, Sarah Naomi Bolz, Florian Kaiser, Sebastian Salentin, V Joachim Haupt, and Michael Schroeder. PLIP 2021: expanding the scope of the protein-ligand interaction profiler to DNA and RNA. *Nucleic Acids Res.*, 49(W1):W530–W534, July 2021.

Patrick J Almhjell, Christina E Boville, and Frances H Arnold. Engineering enzymes for noncanonical amino acid synthesis. *Chemical Society Reviews*, 47(24):8980–8997, 2018.

Patrick J Almhjell, Kadina E Johnston, Nicholas J Porter, Jennifer L Kennemur, Vignesh C Bhethanabotla, Julie Ducharme, and Frances H Arnold. The β-subunit of tryptophan synthase is a latent tyrosine synthase. *Nature Chemical Biology*, pp. 1–8, 2024.

Frances H Arnold. Directed evolution: bringing new chemistry to life. *Angewandte Chemie (International Ed. in English)*, 57(16):4143, 2017.

Elizabeth L Bell, William Finnigan, Scott P France, Anthony P Green, Martin A Hayes, Lorna J Hepworth, Sarah L Lovelock, Haruka Niikura, Sílvia Osuna, Elvira Romero, et al. Biocatalysis. *Nature Reviews Methods Primers*, 1(1):1–21, 2021.

Nathaniel R Bennett, Brian Coventry, Inna Goreshnik, Buwei Huang, Aza Allen, Dionne Vafeados, Ying Po Peng, Justas Dauparas, Minkyung Baek, Lance Stewart, et al. Improving de novo protein binder design with deep learning. *Nature Communications*, 14(1):2625, 2023.

A Patrícia Bento, Anne Hersey, Eloy Félix, Greg Landrum, Anna Gaulton, Francis Atkinson, Louisa J Bellis, Marleen De Veij, and Andrew R Leach. An open source chemical structure curation pipeline using rdkit. *Journal of Cheminformatics*, 12:1–16, 2020.

Jesse D. Bloom, Seth T. Labthavikul, Christopher R. Otey, and Frances H. Arnold. Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences*, 103(15):5869–5874, 2006. doi: 10.1073/pnas.0510098103. URL https://doi.org/10.1073/pnas.0510098103.

Christina E Boville, Remkes A Scheele, Philipp Koch, Sabine Brinkmann-Chen, Andrew R Buller, and Frances H Arnold. Engineered biosynthesis of β-alkyl tryptophan analogues. *Angewandte Chemie*, 130(45):14980–14984, 2018.

Oliver F Brandenberg, Rudi Fasan, and Frances H Arnold. Exploiting and engineering hemoproteins for abiological carbene and nitrene transfer reactions. *Current opinion in biotechnology*, 47:102–111, 2017.

Andrew R Buller, Sabine Brinkmann-Chen, David K Romney, Michael Herger, Javier Murciano-Calles, and Frances H Arnold. Directed evolution of the tryptophan synthase β-subunit for standalone function recapitulates allosteric activation. *Proceedings of the National Academy of Sciences*, 112(47):14599–14604, 2015.

Rebecca Buller, Stefan Lutz, Romas J Kazlauskas, Radka Snajdrova, Jeffrey C Moore, and Uwe T Bornscheuer. From nature to industry: Harnessing enzymes for biocatalysis. *Science*, 382(6673): eadh8615, 2023.

Carla Calvó-Tusell, Zhen Liu, Kai Chen, Frances H Arnold, and Marc Garcia-Borràs. Reversing the enantioselectivity of enzymatic carbene n- h insertion through mechanism-guided protein engineering. *Angewandte Chemie*, 135(35):e202303879, 2023.

Raymond E Carhart, Dennis H Smith, and RENGACHARI Venkataraghavan. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73, 1985.

David A Case, Hasan Metin Aktulga, Kellon Belfon, David S Cerutti, G Andrés Cisneros, Vinícius Wilian D Cruzeiro, Negin Forouzesh, Timothy J Giese, Andreas W Götz, Holger Gohlke, Saeed Izadi, Koushik Kasavajhala, Mehmet C Kaymak, Edward King, Tom Kurtzman, Tai-Sung Lee, Pengfei Li, Jian Liu, Tyler Luchko, Ray Luo, Madushanka Manathunga, Matias R Machado, Hai Minh Nguyen, Kurt A O'Hearn, Alexey V Onufriev, Feng Pan, Sergio Pantano, Ruxi Qi, Ali Rahnamoun, Ali Risheh, Stephan Schott-Verdugo, Akhil Shajan, Jason Swails, Junmei Wang, Haixin Wei, Xiongwu Wu, Yongxian Wu, Shi Zhang, Shiji Zhao, Qiang Zhu, Thomas E III Cheatham, Daniel R Roe, Adrian Roitberg, Carlos Simmerling, Darrin M York, Maria C Nagan, and Kenneth M Jr Merz. Ambertools. *Journal of Chemical Information and Modeling*, 63(20): 6183–6191, 2023. doi: 10.1021/acs.jcim.3c01153. URL https://doi.org/10.1021/acs.jcim.3c01153. PMID: 37805934.

Shobhit S Chaturvedi, Santiago Vargas, Pujan Ajmera, and Anastassia N Alexandrova. Directed evolution of protoglobin optimizes the enzyme electric field. *Journal of the American Chemical Society*, 2024.

Kai Chen and Frances H Arnold. Engineering new catalytic activities in enzymes. *Nature Catalysis*, 3(3):203–213, 2020.

Pedro S Coelho, Eric M Brustad, Arvind Kannan, and Frances H Arnold. Olefin cyclopropanation via carbene transfer catalyzed by engineered cytochrome p450 enzymes. *Science*, 339(6117): 307–310, 2013. doi: 10.1126/science.1231434. URL https://www.science.org/doi/abs/10.1126/science.1231434.

Emma Danelius, Nicholas J Porter, Johan Unge, Frances H Arnold, and Tamir Gonen. Microed structure of a protoglobin reactive carbene intermediate. *Journal of the American Chemical Society*, 145(13):7159–7165, 2023.

Justas Dauparas, Gyu Rie Lee, Robert Pecoraro, Linna An, Ivan Anishchenko, Cameron Glasscock, and D Baker. Atomic context-conditioned protein sequence design using LigandMPNN. December 2023.

David Ding, Andrew Y. Shaw, Samuel Sinai, Nathan Rollins, Nir Prywes, David F. Savage, Michael T. Laub, and Debora S. Marks. Protein design using structure-based residue preferences. *Nature Communications*, 15:1639, 2024a. doi: 10.1038/s41467-024-45621-4. URL https://doi.org/10.1038/s41467-024-45621-4.

Kerr Ding, Michael Chin, Yunlong Zhao, Wei Huang, Binh Khanh Mai, Huanan Wang, Peng Liu, Yang Yang, and Yunan Luo. Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering. *Nature Communications*, 15(1): 6392, 2024b.

Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhonikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955. URL https://www.biorxiv.org/content/early/2024/10/11/2024.10.10.615955.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. Autodock vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61(8):3891–3898, 2021.

David Eisenberg, Elisabeth Schwarz, Michael Komaromy, and Randolph Wall. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *Journal of Molecular Biology*, 179(1):125–142, 1984. ISSN 0022-2836. doi: https://doi.org/10.1016/0022-2836(84)90309-7. URL https://www.sciencedirect.com/science/article/pii/0022283684903097.

David A Estell, Thomas P Graycar, John V Miller, David B Powers, James A Wells, Jean-Pierre Burnier, and Philip G Ng. Probing steric and hydrophobic effects on enzyme-substrate interactions by protein engineering. *Science*, 233(4764):659–663, 1986. doi: 10.1126/science.233.4764.659. URL https://www.science.org/doi/abs/10.1126/science.233.4764.659.

Marc Garcia-Borràs, SB Jennifer Kan, Russell D Lewis, Allison Tang, Gonzalo Jimenez-Osés, Frances H Arnold, and Kendall N Houk. Origin and control of chemoselectivity in cytochrome c catalyzed carbene transfer into si–h and n–h bonds. *Journal of the American Chemical Society*, 143(18):7114–7123, 2021.

Samuel Goldman, Ria Das, Kevin K Yang, and Connor W Coley. Machine learning modeling of family wide enzyme-substrate specificity screens. *PLoS computational biology*, 18(2):e1009853, 2022.

Marcus D Hanwell, Donald E Curtis, David C Lonie, Tim Vandermeersch, Eva Zurek, and Geoffrey R Hutchison. Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics*, 4(1), Aug 2012. doi: 10.1186/1758-2946-4-17.

Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Christoph P I Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, 2017. doi: 10.1038/nbt.3769. URL https://doi.org/10.1038/nbt.3769.

Thomas P Hopp and Kenneth R Woods. A computer program for predicting protein antigenic determinants. *Molecular Immunology*, 20(4):483–489, 1983. ISSN 0161-5890. doi: https://doi.org/10.1016/0161-5890(83)90029-9. URL https://www.sciencedirect.com/science/article/pii/0161589083900299.

Ching-Yu Hsu, Howard Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Learning protein fitness models from evolutionary and assay-labeled data. *Nature Biotechnology*, pp. 1–9, 2022a. doi: 10.1038/s41587-021-01146-5. URL https://doi.org/10.1038/s41587-021-01146-5.

Ching-Yu Hsu, Ruben Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. 2022b. doi: 10.1101/2022.04.10.487779. URL https://doi.org/10.1101/2022.04.10.487779. Preprint.

Xiongyi Huang, Marc Garcia-Borràs, Kun Miao, SB Jennifer Kan, Arjun Zutshi, KN Houk, and Frances H Arnold. A biocatalytic platform for synthesis of chiral α-trifluoromethylated organoborons. *ACS central science*, 5(2):270–276, 2019.

Sean R Johnson, Xiaozhi Fu, Sandra Viknander, Clara Goldin, Sarah Monaco, Aleksej Zelezniak, and Kevin K Yang. Computational scoring and experimental evaluation of enzymes generated by neural networks. *Nature biotechnology*, pp. 1–10, 2024.

Indrek Kalvet, Mary Ortmayer, Jingming Zhao, Rebecca Crawshaw, Nathan M Ennist, Colin Levy, Anindya Roy, Anthony P Green, and David Baker. Design of heme enzymes with a tunable substrate binding pocket adjacent to an open metal coordination site. *Journal of the American Chemical Society*, 145(26):14307–14315, 2023.

SB Jennifer Kan, Russell D Lewis, Kai Chen, and Frances H Arnold. Directed evolution of cytochrome c for carbon–silicon bond formation: Bringing silicon to life. *Science*, 354(6315):1048–1051, 2016. doi: 10.1126/science.aah6219. URL https://www.science.org/doi/abs/10.1126/science.aah6219.

SB Jennifer Kan, Xiongyi Huang, Yosephine Gumulya, Kai Chen, and Frances H Arnold. Genetically programmed chiral organoborane synthesis. *Nature*, 552(7683):132–136, 2017.

Sabrina Kille, Carlos G Acevedo-Rocha, Loreto P Parra, Zhi-Gang Zhang, Diederik J Opperman, Manfred T Reetz, and Juan Pablo Acevedo. Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. *ACS synthetic biology*, 2(2):83–92, 2013.

Lucien F Krapp, Fernando A Meireles, Luciano A Abriata, Jean Devillard, Sarah Vacle, Maria J Marcaida, and Matteo Dal Peraro. Context-aware geometric deep learning for protein sequence design. *Nature Communications*, 15(1):6273, 2024.

Jack Kyte and Russell F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982. ISSN 0022-2836. doi: https://doi.org/10.1016/0022-2836(82)90515-0. URL https://www.sciencedirect.com/science/article/pii/0022283682905150.

Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

Grant M Landwehr, Jonathan W Bogart, Carol Magalhaes, Eric G Hammarlund, Ashty S Karim, and Michael C Jewett. Accelerated enzyme engineering by machine-learning guided cell-free expression. *Nature Communications*, 16(1):865, 2025.

Reuben B Leveson-Gower, Clemens Mayer, and Gerard Roelfes. The importance of catalytic promiscuity for enzyme design and evolution. *Nature Reviews Chemistry*, 3(12):687–705, 2019.

Francesca-Zhoufan Li, Jason Yang, Kadina E. Johnston, Emre Gürsoy, Yisong Yue, and Frances H. Arnold. Evaluation of machine learning-assisted directed evolution across diverse combinatorial landscapes. 2024. doi: 10.1101/2024.10.24.619774.

Yueming Long, Ariane Mora, Francesca-Zhoufan Li, Emre Gursoy, Kadina E Johnston, and Frances H Arnold. Levseq: Rapid generation of sequence-function data for directed evolution and machine learning. *ACS Synthetic Biology*, 14(1):230–238, 2024.

Stefan Lutz and Samantha M Iamurri. Protein engineering: past, present, and future. *Protein Engineering: Methods and Protocols*, pp. 1–12, 2018.

Wai Shun Mak, Stephen Tran, Ryan Marcheschi, Steve Bertolani, James Thompson, David Baker, James C Liao, and Justin B Siegel. Integrative genomic mining for enzyme function to enable engineering of a non-natural biosynthetic pathway. *Nature communications*, 6(1):10005, 2015.

Mónica Martínez-Martínez, Cristina Coscolín, Gerard Santiago, Jennifer Chow, Peter J Stogios, Rafael Bargiela, Christoph Gertler, Jose Navarro-Fernandez, Alexander Bollinger, Stephan Thies, et al. Determinants and prediction of esterase substrate promiscuity patterns. *ACS chemical biology*, 13(1):225–234, 2017.

Joshua Meier, Roshan Rao, Ruben Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021. doi: 10.1101/2021.07.09.450648. URL https://doi.org/10.1101/2021.07.09.450648. Preprint.

Frank Neese. The orca program system. *WIRES Comput. Molec. Sci.*, 2(1):73–78, 2012. doi: 10.1002/wcms.81.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/cac723e5ff29f65e3fcbb0739ae91bee-Paper-Datasets_and_Benchmarks.pdf.

Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3:1–14, 2011.

Patrick J O'Brien and Daniel Herschlag. Catalytic promiscuity and the evolution of new enzymatic activities. *Chemistry Biology*, 6(4):R91–R105, 1999. ISSN 1074-5521. doi: https://doi.org/10.1016/S1074-5521(99)80033-7. URL https://www.sciencedirect.com/science/article/pii/S1074552199800337.

Michael S Packer and David R Liu. Methods for the directed evolution of proteins. *Nature Reviews Genetics*, 16(7):379–394, 2015.

Hahnbeom Park, Guangfeng Zhou, Minkyung Baek, David Baker, and Frank DiMaio. Force field optimization guided by small molecule crystal lattice data enables consistent sub-angstrom protein–ligand docking. *Journal of Chemical Theory and Computation*, 17(3):2000–2010, 2021.

Alexandra Paton, Daniil Boiko, Jonathan Perkins, Nicholas Cemalovic, Thiago Reschützegger, Gabe Gomes, and Alison Narayan. Generation of connections between protein sequence space and chemical space to enable a predictive model for biocatalysis. 2024.

Jacob T Rapp, Bennett J Bremer, and Philip A Romero. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nature chemical engineering*, 1(1):97–107, 2024.

Julia C Reisenbauer, Kathleen M Sicinski, and Frances H Arnold. Catalyzing the future: recent advances in chemical synthesis using enzymes. *Current Opinion in Chemical Biology*, 83:102536, 2024. ISSN 1367-5931. doi: https://doi.org/10.1016/j.cbpa.2024.102536. URL https://www.sciencedirect.com/science/article/pii/S1367593124001121.

Hans Renata, Z Jane Wang, and Frances H Arnold. Expanding the enzyme universe: accessing non-natural reactions by mechanism-guided directed evolution. *Angewandte Chemie International Edition*, 54(11):3351–3367, 2015.

Philip A Romero and Frances H Arnold. Exploring protein fitness landscapes by directed evolution. *Nature reviews Molecular cell biology*, 10(12):866–876, 2009.

David K Romney, Javier Murciano-Calles, Jori E Wehrmuller, and Frances H Arnold. Unlocking reactivity of trpb: a general biocatalytic platform for synthesis of tryptophan analogues. *Journal of the American Chemical Society*, 139(31):10769–10776, 2017.

Shu-ou Shan and Daniel Herschlag. The change in hydrogen bond strength accompanying charge rearrangement: Implications for enzymatic catalysis. *Proceedings of the National Academy of Sciences*, 93(25):14474–14479, 1996. doi: 10.1073/pnas.93.25.14474. URL https://www.pnas.org/doi/abs/10.1073/pnas.93.25.14474.

Justin B Siegel, Alexandre Zanghellini, Helena M Lovick, Gert Kiss, Abigail R Lambert, Jennifer L St.Clair, Jasmine L Gallaher, Donald Hilvert, Michael H Gelb, Barry L Stoddard, Kendall N Houk, Forrest E Michael, and David Baker. Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science*, 329(5989):309–313, 2010. doi: 10.1126/science.1190239. URL https://www.science.org/doi/abs/10.1126/science.1190239.

Dinesh Kumar Sriramulu and Sun-Gu Lee. Combinatorial effect of ligand and ligand-binding site hydrophobicities on binding affinity. *Journal of Chemical Information and Modeling*, 60(3): 1678–1684, 2020. doi: 10.1021/acs.jcim.9b01143. URL https://doi.org/10.1021/acs.jcim.9b01143. PMID: 32108477.

Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Harmonic self-conditioned flow matching for multi-ligand docking and binding site design. 2023.

Filippo Stocco, Maria Artigues-Lleixa, Andrea Hunklinger, Talal Widatalla, Marc Guell, and Noelia Ferruz. Guiding generative protein language models with reinforcement learning. *arXiv preprint arXiv:2412.12979*, 2024.

Magdalena Teufl, Charlotte U Zajc, and Michael W Traxlmayr. Engineering strategies to overcome the stability–function trade-off in proteins. *ACS Synthetic Biology*, 11(3):1030–1039, 2022.

Neil Thomas, David Belanger, Chenling Xu, Hanson Lee, Kathleen Hirano, Kosuke Iwai, Vanja Polic, Kendra D Nyberg, Kevin G Hoff, Lucas Frenz, Charlie A Emrich, Jun W Kim, Mariya Chavarha, Abi Ramanan, Jeremy J Agresti, and Lucy J Colwell. Engineering of highly active and diverse nuclease enzymes by combining machine learning and ultra-high-throughput screening. 2024. doi: 10.1101/2024.03.21.585615.

Wei Tian, Chang Chen, Xue Lei, Jieling Zhao, and Jie Liang. Castp 3.0: computed atlas of surface topography of proteins. *Nucleic acids research*, 46(W1):W363–W367, 2018.

Matthew Z Tien, Austin G Meyer, Dariya K Sydykova, Stephanie J Spielman, and Claus O Wilke. Maximum allowed solvent accessibilites of residues in proteins. *PloS one*, 8(11):e80635, 2013.

Antonio Tinoco, Yang Wei, John-Paul Bacik, Daniela M Carminati, Eric J Moore, Nozomi Ando, Yong Zhang, and Rudi Fasan. Origin of high stereocontrol in olefin cyclopropanation catalyzed by an engineered carbene transferase. *ACS catalysis*, 9(2):1514–1524, 2018.

Yajie Wang, Pu Xue, Mingfeng Cao, Tianhao Yu, Stephan T Lane, and Huimin Zhao. Directed evolution: methodologies and applications. *Chemical reviews*, 121(20):12384–12444, 2021.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Yang Wei, Antonio Tinoco, Viktoria Steck, Rudi Fasan, and Yong Zhang. Cyclopropanations via heme carbenes: Basic mechanism and effects of carbene substituent, protein axial ligand, and porphyrin substitution. *Journal of the American Chemical Society*, 140(5):1649–1662, 2017.

Bradley J. Wittmann, Yanfeng Yue, and Frances H. Arnold. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Systems*, 2021. doi: 10.1016/j.cels.2021.07.008. URL https://doi.org/10.1016/j.cels.2021.07.008.

Bruce J Wittmann, Kadina E Johnston, Patrick J Almhjell, and Frances H Arnold. Evseq: cost-effective amplicon sequencing of every variant in a protein library. *ACS Synthetic Biology*, 11(3): 1313–1324, 2022.

Jason Yang, Francesca-Zhoufan Li, and Frances H Arnold. Opportunities and challenges for machine learning-assisted enzyme engineering. *ACS Central Science*, 10(2):226–241, 2024.

Jason Yang, Ravi G Lal, James C Bowden, Raul Astudillo, Mikhail A Hameedi, Sukhvinder Kaur, Matthew Hill, Yisong Yue, and Frances H Arnold. Active learning-assisted directed evolution. *Nature Communications*, 16(1):714, 2025.

Jason Yim, Hannes Stärk, Gabriele Corso, Bowen Jing, Regina Barzilay, and Tommi S Jaakkola. Diffusion models in protein structure and docking. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 14(2):e1711, 2024.

A Appendix

A.1 Datasets

A.1.1 Summary

All datasets are available on Zenodo: `https://zenodo.org/records/15226690`.

Table A1: Dataset summary

Enzyme	Substrates	Cofactor	# Pairs	# Sites	Sites	Activity	Selectivity
*Pf*TrpB	4bromo indole + L-serine	PLP	241	3	I165, I183, Y301	Absorbance	N.A.
*Pf*TrpB	4cyano indole + L-serine	PLP	241	3	I165, I183, Y301	Absorbance	N.A.
*Pf*TrpB	5bromo indole + L-serine	PLP	241	3	I165, I183, Y301	Absorbance	N.A.
*Pf*TrpB	5chloro indole + L-serine	PLP	241	3	I165, I183, Y301	Absorbance	N.A.
*Pf*TrpB	5cyano indole + L-serine	PLP	241	3	I165, I183, Y301	Absorbance	N.A.
*Pf*TrpB	5iodo indole + L-serine	PLP	241	3	I165, I183, Y301	Absorbance	N.A.
*Pf*TrpB	6chloro indole + L-serine	PLP	241	3	I165, I183, Y301	Absorbance	N.A.
*Pf*TrpB	7bromo indole + L-serine	PLP	68	3	I165, I183, Y301	Absorbance	N.A.
*Pf*TrpB	7iodo indole + L-serine	PLP	241	3	I165, I183, Y301	Absorbance	N.A.
*Pf*TrpB	7methyl indole + L-serine	PLP	241	3	I165, I183, Y301	Absorbance	N.A.
*Pf*TrpB	5,6chloro indole + L-serine	PLP	241	3	I165, I183, Y301	Absorbance	N.A.
Rma cyt *c*	NHC-borane + Me-EDA	heme	150	6	V75, M99, M100, T101T, D102, M103	% yield	Enantio-
Rma cyt *c*	phenyldimethyl-silane + Me-EDA	heme	150	6	V75, M99, M100, T101T, D102, M103	% yield	Enantio-
ParLQ	a: 4-vinylanisole + EDA	heme	490	5	W56, Y57, L59, Q60, F89	% yield	Diastereo-
ParLQ	b: styrene + EDA	heme	91	5	W56, Y57, L59, Q60, F89	% yield	Diastereo-
ParLQ	c: 1-methyl-4-vinylbenzene + EDA	heme	91	5	W56, Y57, L59, Q60, F89	% yield	Diastereo-
ParLQ	d: 1-methyl-3-vinylbenzene + EDA	heme	91	5	W56, Y57, L59, Q60, F89	% yield	Diastereo-
ParLQ	e: 1-methyl-2-vinylbenzene + EDA	heme	91	5	W56, Y57, L59, Q60, F89	% yield	Diastereo-
ParLQ	f: 1-chloro-4-vinylbenzene + EDA	heme	91	5	W56, Y57, L59, Q60, F89	% yield	Diastereo-
ParLQ	g: 1-bromo-4-vinylbenzene + EDA	heme	91	5	W56, Y57, L59, Q60, F89	% yield	Diastereo-
ParLQ	h: 1-(trifluoromethyl)-4-vinylbenzene + EDA	heme	91	5	W56, Y57, L59, Q60, F89	% yield	Diastereo-
ParLQ	i: 2-vinylnaphthylene +EDA	heme	91	5	W56, Y57, L59, Q60, F89	% yield	Diastereo-

A.1.2 Dataset backgrounds

The tryptophan synthase β -subunit (TrpB) catalyzes a native reaction between L-serine and indole to form tryptophan. Engineered TrpBs extend this function to non-native substrates such as serine analogues and substituted indoles, enabling the synthesis of tryptophan analogs and other noncanonical amino acids that are important precursors to pharmaceuticals and natural products (Figure 1c) (Buller et al., 2015; Romney et al., 2017; Almhjell et al., 2018; Boville et al., 2018).

Heme-containing enzymes have been engineered to carry out a plethora of valuable reactions that have not been found in biological systems (Brandenberg et al., 2017). These new-to-nature reactivities include carbene transfers for stereoselective olefin cyclopropanation, traditionally requiring unsustainable transition metals (Renata et al., 2015; Coelho et al., 2013), and the formation of carbon–silicon (C–Si) (Kan et al., 2016) and carbon–boron (C–B) (Kan et al., 2017) bonds (Figure 1d).

A.1.3 Multi-substrate *Pf*TrpB dataset

Library Generation Beginning with a TrpB variant discovered in a directed evolution campaign for 4-nitroTrp formation, *Pf*5G8 (Romney et al., 2017), a triple-site saturation mutagenesis library was generated. Primers from Table A2 were used to amplify out the vector in three pieces and install variation at positions 165, 183, and 301 via the 22-codon trick (Kille et al., 2013). Amplification was performed with Phusion® High-Fidelity DNA Polymerase according to manufacturer recommendations (New England Biolabs, Catalog M0530L). A Gibson assembly was used to generate full-length vectors which were transformed via electroporation into electrocompetent BL21-DE3 *E. coli* cells. These cells were plated onto LB agar containing 100 µg/mL ampicillin.

Single colonies were picked into 96-well deep well plates containing 300 µL TB containing 100 µg/mL ampicillin (TB_{Amp}) and grown overnight at 37 °C, 250 rpm, and 80% humidity. The following day, expression 96-well deep well plates were filled with 630 µL TB_{Amp} and 20 µL culture and grown for 3 h at 37 °C, 250 rpm, and 80% humidity. These plates were then cooled on ice for 20 min prior to adding 50 µL 14 mM IPTG (1 mM final) and incubating overnight at 25 °C and 250 rpm overnight. Cells were pelleted at 3500–4000 rpm for 10 min, the supernatant was decanted, and the plates were then frozen at -20 °C overnight.

Table A2: Primer Sequences, where XXX = 22 codon trick (Kille et al., 2013)

Name	Sequence
I165_f	GTTCTCGCACCCTGAAAGACGCAXXXGACGAGGCTCTGCGTGATTGG
I165_r	TGCGTCTTTCAGGGTGCGAGAAC
I183_f	GTGGCTACTTTTGAATACACCCACTACCTAXXXGGTTCCGTGGTCGGTCCAC
I183_r	TAGGTAGTGGGTGTATTCAAAAGTAGCCAC
Y301_f	CTCCATCGCACCAGGTCTGGATXXXCCAGGTGTTGGTCCAGAACACG
Y301_r	ATCCAGACCTGGTGCGATGGAG

Screening To prepare cell lysate, pellets were first resuspended in lysis buffer composed of 1mg/mL HEWL, 2 mM $MgCl_2$, 10X bug buster, 200 µM PLP, and a small amount of DNAse in 50 mM potassium phosphate buffer, pH 8.0 (KPi). They were then incubated at 37 °C for 30 min and heat treated at 75 °C for at least 30 min. Plates were then spun down at 5000 rpm for 10 min and the supernatant was used as cell lysate.

Nucleophile stocks were made for all indole analogs at 200 mM in either EtOH (4bromo, 5bromo) or DMSO (7bromo, 5chloro, 6chloro, 5,6chloro, 5iodo, 7iodo, 7methyl, 4cyano, 5cyano). Reactions were set up in 96-well deep well plates. Reactions were prepared with 10 µL nucleophile stock (10 mM final), 20 µL lysate, 10 µL L-serine (25 mM final), and 160 µL KPi and incubated in a tightly sealed plate at 75 °C overnight. The next day the reactions were acidified with 200 µL 1M HCl and the unreacted indole extracted with 500 µL EtOAc. The plates were sealed tightly and shaken vigorously, then spun down at 1000 rpm for 3 min to separate the layers before drawing the bottom (aqueous) layer of the mixture into a 96-well UV-transparent flat-bottom plate. Absorbance was collected every 2 nm from 290–310 nm for every substrate using a Tecan InfiniTe. The absorbance wavelength used for quantification for each substrate was selected according to its absorbance properties (Table A3).

Table A3: Wavelengths

Compound	Wavelength (nm)
4bromo	304
5bromo	306
7bromo	300
5chloro	306
6chloro	304
5,6chloro	310
5iodo	306
7iodo	306
4cyano	294
5cyano	310
7methyl	296

A.1.4 MULTI-SUBSTRATE PARLQ DATASET

The dataset was sourced from the study by Yang and Lal *et al.* (Yang et al., 2025). The model substrate was presented in their main text, while the substrate scope data was provided in the supplementary information. Additional experimental details were confirmed through direct communication with the authors.

Activity was calculated based on GC-FID measurements, where the product area was normalized to the internal standard area and converted using the calibration curve from Figures S11–S28. The yield calculation followed the author's notebook (GitHub repository), normalizing the area to the maximum possible product concentration and accounting for the 1.5X dilution from the reaction. The *cis* isomer was the major product. Selectivity was determined by calculating the ratio of the *cis* to *trans* isomer.

A.1.5 *Rma* CYTOCHROME *c* C–B AND C–SI DATASET

The dataset was sourced from the study by Ding *et al.* (Supplementary Tables 3 and 4) (Ding et al., 2024b). Upon communication with the corresponding author, we confirmed that no sequence information was collected from the random mutagenesis libraries presented in Supplementary Tables 5 and 6.

A.2 MECHANISM

A.2.1 PLP-DEPENDENT TRPB REACTIONS

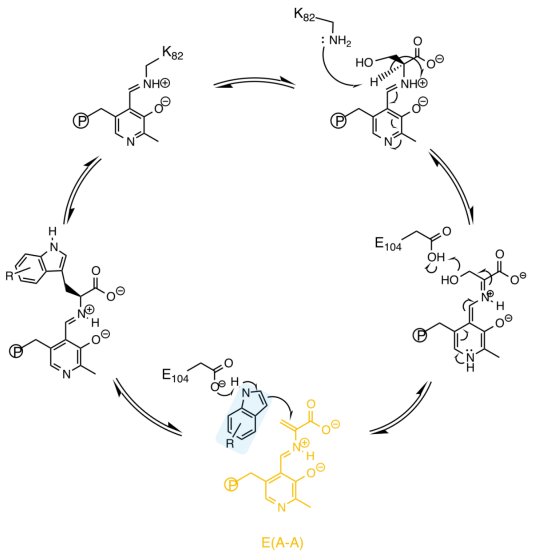

Figure A1: TrpB mechanism based on published studies (Buller et al., 2015; Romney et al., 2017; Almhjell et al., 2024). The E(A-A) intermediate together with the substrates were used for ZS predictors.

A.2.2 HEME-BASED CARBENE TRANSFER REACTIONS

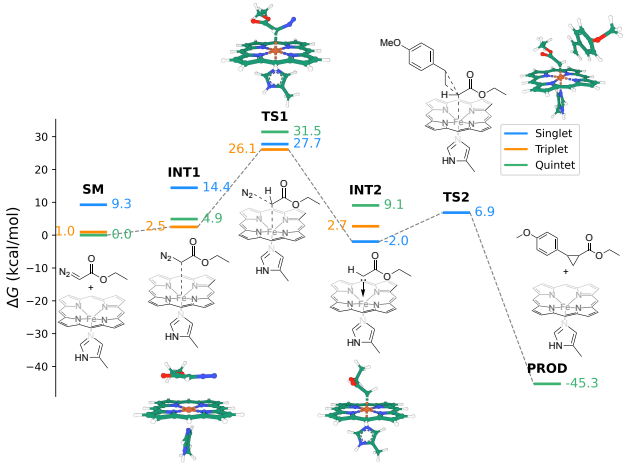

Figure A2: Mechanism for heme-based carbene transfer reaction as computed by DFT (Appendix A.3.15) (Tinoco et al., 2018; Calvó-Tusell et al., 2023). TS2 was used for ZS predictors.

A.3 METHODS

The code is available on GitHub: `https://github.com/fhalab/substrate_aware_zs`.

A.3.1 GENERAL ZS PREDICTORS

Hamming distance, EVmutation, ESM, ESM-IF, CoVES, and ($\Delta\Delta G_f$) ZS scores were calculated based on the study by Li *et al.* (Li et al., 2024).

A.3.2 VINA

AutoDock Vina v1.2.5 was used. PDBQT files for substrates were prepared from corresponding SMILES strings using RDKit at pH 7.4 and Open Babel (Bento et al., 2020; Landrum, 2013; O'Boyle et al., 2011). The cofactor was extracted from the parent PDB and converted to PDBQT using Open Babel, while metal ions were prepared separately. Receptor structures were derived from parent PDB structures (PDB ID: 5DW0 for *Pf* TrpB and 3CP5 for *Rma* cyt *c*), while the structure for ParLQ was modeled by the authors using Alphafold 3 with a bound-heme. Variant structures were generated using MDAnalysis.

Docking coordinates were defined by the centroid of the substrate-cofactor complex with a box size of 20Å. Each docking experiment was performed in five replicates, with nine docking modes and an exhaustiveness setting of 32. The lowest energy from each replicate was recorded, and the final energy was averaged across replicates. The negative values of the energies were used as the ZS predictor.

A.3.3 ROSETTA GALIGANDDOCK

The Pyrosetta GALigandDock-based ZS scores were obtained from a local copy of the pyrosetta distribution *pyrosetta-2025.3+release.1f5080a079-py3.12-linux-x86_64.egg/pyrosetta/distributed*. Two conda environments (*anaconda.org*) were created. One for the input preprocessing and one for inference of Pyrosetta GALigandDock. To set them up, download the corresponding .yaml files (*ambertools.yml* and *pyrosetta_env.yml*) and execute the following console commands:

```
$ conda env create -f <path/to/ambertools.yml>
```

and

```
$ conda env create -f <path/to/pyrosetta_env.yml>
```

respectively.

To preproccess the inputs, the first conda environment was activated and the script *pyrosetta_pipeline.py* was executed with the following parameters:

```
python -m substrate_aware.zs.pyrosetta_pipeline
--meta_list <path/to/campaign_1_meta.csv>
            <path/to/campaign_2_meta.csv>
            <...>
            <path/to/campaign_n_meta.csv>
--struc_dir <path/to/structures_dir>
--tmp_dir <path/to/dir/for/tempfiles>
--out_dir <path/to/dir/for/output_files>
--rosettascript_path <rosetta/source/.../mol2genparams.py>
--net_charge_unit_1 <net charge>
--net_charge_unit_2 <net charge>
```

The script takes docked structures for a given campaign as input and returns them adequately reformatted for Pyrosetta, alongside a Pyrosetta-specific parameter file for each of the campaign's substrates. The script runs into a tracepoint and prompts the user to manually correct a newly created mol2 file of the substrates and then save it under a printed location. For this purpose, the file was then downloaded, observed in a 3D molecular viewer, such as Avogadro (Hanwell et al., 2012) and edited

to meet antechambers (Case et al., 2023) requirements for am1bcc charge generation. This includes adding hydrogen, correcting unnatural bond orders, and ensuring that the molecule only contains atoms of the element set {H, C, N, O, F, P, S, Cl, Br, I} on which antechamber is parametrized. In the case of iron coordination centers, the metal atom was replaced by a phosphorous. Boron and silicon were substituted with carbon. Lastly, each connected unit must consist of 4 or more atoms. Units with less than that (e.g. ions) were omitted. To finish the preprocessing, the console prompts were followed.

The second script, *pyrosetta_inference.py*, runs the actual GALigandDock docking by executing it with the following parameters:

```
python -m substrate_aware.zs.pyrosetta_inference
--meta_list  <path/to/campaign_1_meta.csv>
             <path/to/campaign_2_meta.csv>
             <...>
             <path/to/campaign_n_meta.csv>
--preprocessed_dir <path/to/directory_containing_pdbs_and_params>
--results_dir <path/to/results_dir>
```

The docking mover is parametrized within this script. This will create two output files for each variant of all campaigns. The *variantname_aligned_enzyme_final.pkl* contains the best scoring docked poses and *variantname_aligned_enzyme_final.csv* contains a table with Rosetta metrics of these poses. Finally, for each campaign *campaignname.csv* summarizes the Rosetta-metrics of the best docked pose of each variant together with variant ground-truth data.

The negative values of all energy terms were extracted. The dH value, representing enthalpy, was used as the score to indicate the thermodynamic stability of the binding event, where more exothermic values correspond to stronger binding.

A.3.4 ALPHAFOLD 3 (AF3)

For *Pf*TrpB, the substrate SMILES was joint with the E(A-A) intermediate (Figure A1) and the crystallographic sodium ion to prevent the substrate from docking onto the enzyme surface. For heme-based reactions, the substrate SMILES was assigned to chain B, while the carbene-heme intermediate complex (Figure A2) TS2 was assigned to chain C. All scores were extracted from five replicates, and the final structure for each variant was aggregated. The scores from the replicates were averaged. For chain-predicted aligned errors (PAE), the negative values were used as the predictor. The confidence scores of each residue at the targeted site were also extracted and averaged as a predictor.

A.3.5 CHAI-1

Chai-1 version 0.1.0 was used, following the same process as AF3, except without MSAs and using PAE as scores.

A.3.6 LIGANDMPNN

Code from LigandMPNN GitHub was adopted to extract the ZS scores (Dauparas et al., 2023). The model with 20Å Gaussian noise was chosen. Only the mutated residues of the campaign were redesigned with autoregressive scoring. To mitigate biases introduced through decoding order, the number of batches was set to 100. Variant likelihoods were thus obtained through:

$$P_{\text{Variant}} = \frac{1}{100} \sum_{i=1}^{100} \prod_{n=1}^{N_{\text{mut}}} P\big(AA_n \mid \text{Backbone}, \{AA_j \mid j < n\}_i\big). \tag{1}$$

A.3.7 FLOWSITE

Code from FlowSite GitHub was adopted to extract ZS scores (Stärk et al., 2023). The parameters were chosen according to the author's suggestion. To evaluate the docking and sequence co-generation as appropriate to directed evolution campaigns, both the residues to design and the

pocket were defined via the mutated sites. For each variant, 100 inference trajectories were generated. Predicted likelihoods were averaged among inferences and position to yield the final variant ZS score.

A.3.8 BOND DISTANCES

Bond distances were derived from AF3 docked structures based on the mechanisms for bond-forming atoms (Appendix A.2). For *Pf*TrpB, distances were measured between the catalytic Glu104 and N1-hydrogen. For heme-based carbene transfer reactions, the distances were measured between the carbene carbon and either boron, silicon, or the styrene double bond.

Distances were calculated for each replicate and averaged. The negative value of the bond distance was used as the predictor, based on the hypothesis that closer reactive atoms lead to stronger reactivity and, consequently, higher activity.

A.3.9 PROTEIN-LIGAND-INTERACTION-PROFILER (PLIP)

A local copy of the PLIP software (release 2.4.0) was obtained from the GitHub (Adasme et al., 2021). The AF3 docked strucutres were used as inputs. An output XML report file was generated to characterize each variant's ligand-active-site-interactions.

A.3.10 ACTIVE-SITE IDENTIFICATION

Two different active site extraction heuristics were explored. The first heuristic defines all residues to belong to the active site, that bear the centroid of their side-chain atoms within a 10Å distance threshold of the ligand's centroid.

The second extraction heuristic used PLIP to define the active site. The residues tagged with "bindingsite" were considered (Appendix A.3.9).

A.3.11 HYDROGEN BONDS

The AF3 docked structures were used to run PLIP (Appendix A.3.9). The number of hydrogen bonds identified in the active site was extracted from the output files and used as a ZS predictor.

A.3.12 HYDROPHOBICITY

For the enzyme, active-site hydrophobicity was calculated based on different active-site identification methods (Appendix A.3.10) using various scales, including the Kyte-Doolittle scale (Kyte & Doolittle, 1982), the Hopp-Woods scale (Hopp & Woods, 1983), the Eisenberg scale (Eisenberg et al., 1984), and theoretical and empirical residue solvent accessibility (Tien et al., 2013).

For the ligand, logP was calculated. While previous literature used the Kyte-Doolittle scale to identify hydrophobic regions likely to be in transmembrane segments (Sriramulu & Lee, 2020), we instead chose the Hopp-Woods scale, which highlights antigenic (hydrophilic) regions on protein surfaces.

A.3.13 ACTIVE-SITE VOLUME

The substrate volume was estimated based on the ConvexHull of the substrate. The active-site volume of the parent was estimated with CASTp based on PDB ID 5DW0 for *Pf*TrpB, 3CP5 for *Rma* cyt *c* and 3ZJI for ParLQ (Tian et al., 2018). The variant active-site volume was estimated by the different in different amino acid side chain at the targeted sites.

A.3.14 SIMILARITY CALCULATIONS

To quantify the similarity between the indole analogs to the native indole, Tanimoto similarity of atom-pair fingerprints (Carhart et al., 1985) was calculated with RDKit (Bento et al., 2020; Landrum, 2013).

A.3.15 DFT CALCULATIONS

DFT calculations were conducted using Orca 6.0 (Neese, 2012). We constructed a model containing the porphyrin core, Fe center and an imidazole to mimic the histidine ligand. Geometry optimizations and frequency calculations were performed using the unrestricted B3LYP hybrid functional with def2-TZVP basis set and with D3(BJ) dispersion correction. All geometries were verified as minima or first-order saddle points by frequency analysis. Enthalpies and entropies were calculated for 1 atm and 298.15 K. The SMD continuum solvation model was used in all optimizations and single point calculations with water as the implicit solvent to approximate the energy otherwise required when the reaction is performed without the enzyme. In Figure A2, we show the complete energetics of heme-catalyzed cyclopropanation with different spin-states. In other carbenoid reactions, we report energetic barriers derived from open-shell singlet calculations of the C-Si insertion/borylation transition state, in comparison against the carbene-porphryin intermediate. Readers should note that while DFT can derive reasonable geometries for transition states, the absolute energy values can have significant margin of error and should only serve as qualitative estimates. Surprisingly, DFT calculations yielded similar activation energies of $\sim$9-13 kcal/mol for the three new-to-nature reactions, as shown by other studies (Garcia-Borràs et al., 2021; Huang et al., 2019; Wei et al., 2017). We also obtained the ΔG of the reaction considering all substrates and products (Table A12).

A.3.16 ENSEMBLE MODELS

To ensemble ZS predictors into a unified score, unweighted ensemble and different types of learned linear models were explored. Results from the shallow neural network were excluded due to overfitting.

Unweighted ensemble Each ZS predictor was ranked, and the ranks of different chosen ZS predictors were summed up for the final score.

Learned ensemble Each model was fitted on one specific enzyme optimization campaign and successively tested on all other campaigns. The models included linear regression, piecewise linear regression with a threshold. By doing so, we tested whether a model's learned relationship between feature scores and measured activities. During prediction, these models thereby weighted individual ZS scores and introduced nonlinearities. Given the data:

$$\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{m}, \quad \mathbf{x}^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}\right), \quad y^{(i)} \in \mathbb{R}.$$

The goal is defined as:

$$\min \sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)}\right)^2,$$

where $\hat{y}^{(i)}$ depends on the chosen transformation. And that fitting on 1 set of $y^{(i)}$ generalizes to other sets of ys.

Inputs where normalized according to:

$$x_i = \frac{x_i - \mu_x}{\sigma_x}$$

Linear regression In the case of linear regression (w), the prediction is obtained by the transformation:

$$\hat{y}^{(i)} = w_0 + \sum_{j=1}^{n} w_j \, x_j^{(i)}.$$

where w_0, w_j are obtained through the optimization problem:

$$\min_{w_0,\ldots,w_n} \sum_{i=1}^{m} \left(y^{(i)} - [w_0 + \sum_{j=1}^{n} w_j \, x_j^{(i)}]\right)^2$$

Piecewise linear regression Although linear regression is straightforward to fit and interpret, it may fail to capture threshold-dependent behaviors (e.g., scores only become useful after a certain threshold and optimization is capped after a certain cutoff). To address this, we additionally considered a piecewise linear regression model, which introduces simple nonlinearities via a learned threshold for each feature. The prediction is obtained by the transformation:

$$\hat{y}^{(i)} = w_0 + \sum_{j=1}^{n} w_j \, \phi_j\big(x_j^{(i)}; \alpha_{j1}, \alpha_{j2}\big).$$

Where the mapping function ϕ_j introduces the nonlinearity:

$$\phi_j(x_j; \alpha_{j1}, \alpha_{j2}) = \begin{cases} 0, & x_j < \alpha_{j1}, \\ \dfrac{x_j - \alpha_{j1}}{\alpha_{j2} - \alpha_{j1}}, & \alpha_{j1} \leqslant x_j < \alpha_{j2}, \\ 1, & x_j \geqslant \alpha_{j2}. \end{cases}$$

The piecewise model parameters $\{w_0, w_j\}$ and thresholds $\{\alpha_{j1}, \alpha_{j2}\}$ are fit by minimizing the sum of squared errors:

$$\min_{\substack{w_0, w_1, \ldots, w_n \\ \alpha_{j1}, \alpha_{j2}}} \sum_{i=1}^{m} \Big(y^{(i)} - \big[w_0 + \sum_{j=1}^{n} w_j \, \phi_j(x_j^{(i)}; \alpha_{j1}, \alpha_{j2}) \big] \Big)^2,$$

subject to $\alpha_{j1} < \alpha_{j2}$ for each feature j.

A.4 ADDITIONAL RESULTS

A.4.1 DATASET VISUALIZATION

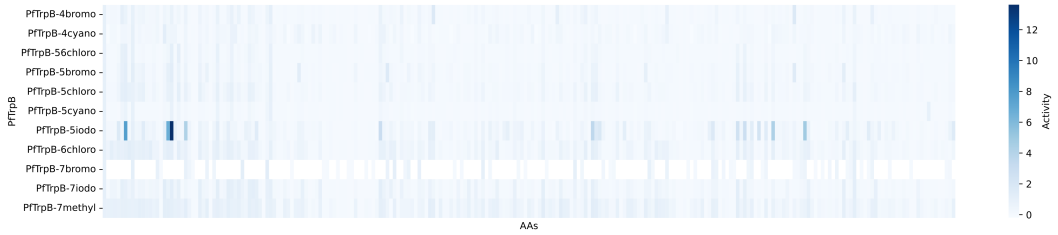

Figure A3: *Pf* TrpB activity

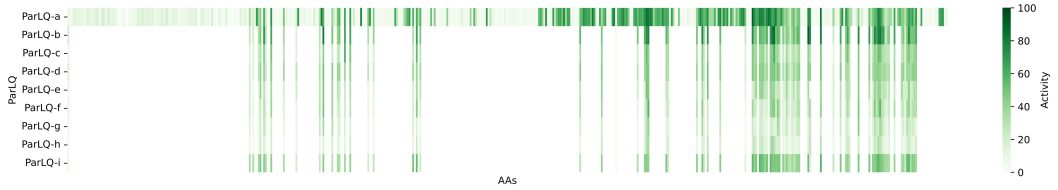

Figure A4: ParLQ activity

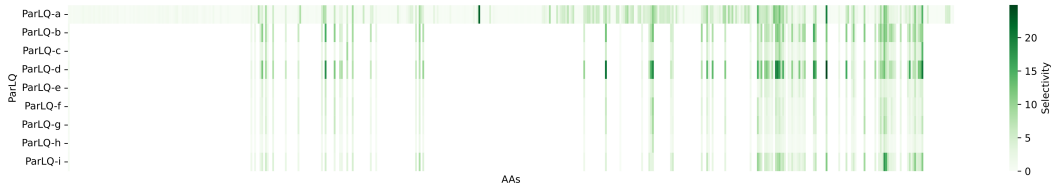

Figure A5: ParLQ diastereoselectivity

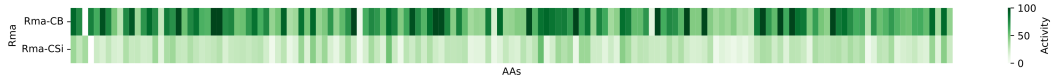

Figure A6: *Rma* cyt *c* activity

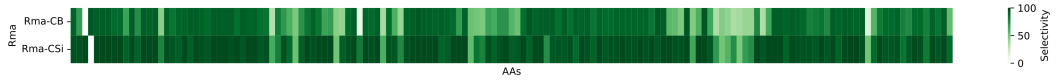

Figure A7: *Rma* cyt *c* enantioselectivity

A.4.2 INDIVIDUAL ZS PREDICTOR PERFORMANCE

Table A4: ZS predictors are averaged across all, non-native, and new-to-nature datasets. **Bold** indicates the best predictor, ***bold italics*** indicates the second-best predictor, and *italics* highlight the third-best predictor within each category.

ZS predictor	All	Non-native substrate	New-to-nature chemistry
Hamming distance	0.1056	0.3378	-0.1266
EVmutation	**0.2768**	0.4652	0.0885
ESM-2	0.1904	***0.5125***	-0.1316
ESM-IF	*0.2534*	*0.4810*	0.0257
CoVES	0.1473	0.4075	-0.1129
$\Delta\Delta G_f$	0.2379	**0.5253**	-0.0495
Vina	0.0361	0.0257	0.0465
GALigandDock	0.1393	0.1228	***0.1559***
AF3	***0.2751***	0.2416	**0.3086**
Chai-1	0.1420	0.2094	0.0746
LigandMPNN	0.2105	0.4780	-0.0570
FlowSite	0.2176	0.4007	0.0345
Bond distance	0.0607	0.1853	-0.0639
Hydrogen bonds	0.1633	0.2802	0.0464
Hydrophobicity	0.2028	0.4128	-0.0072
Active-site volume	0.0828	0.0469	*0.1187*

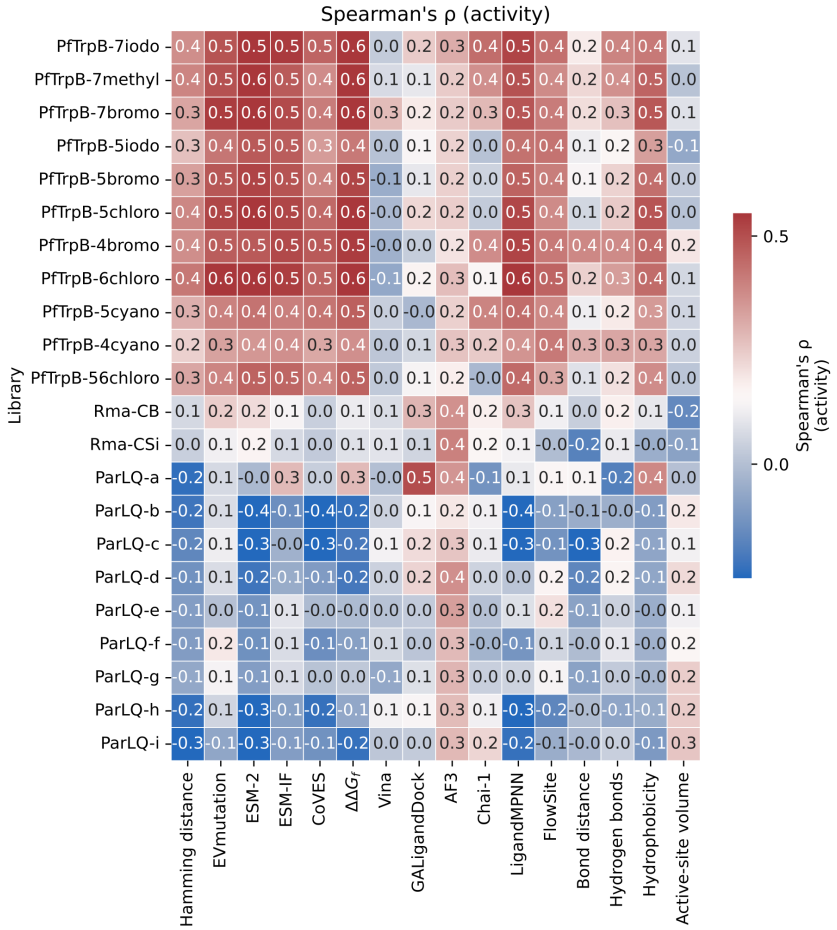

Figure A8: Spearmen's correlation for activity

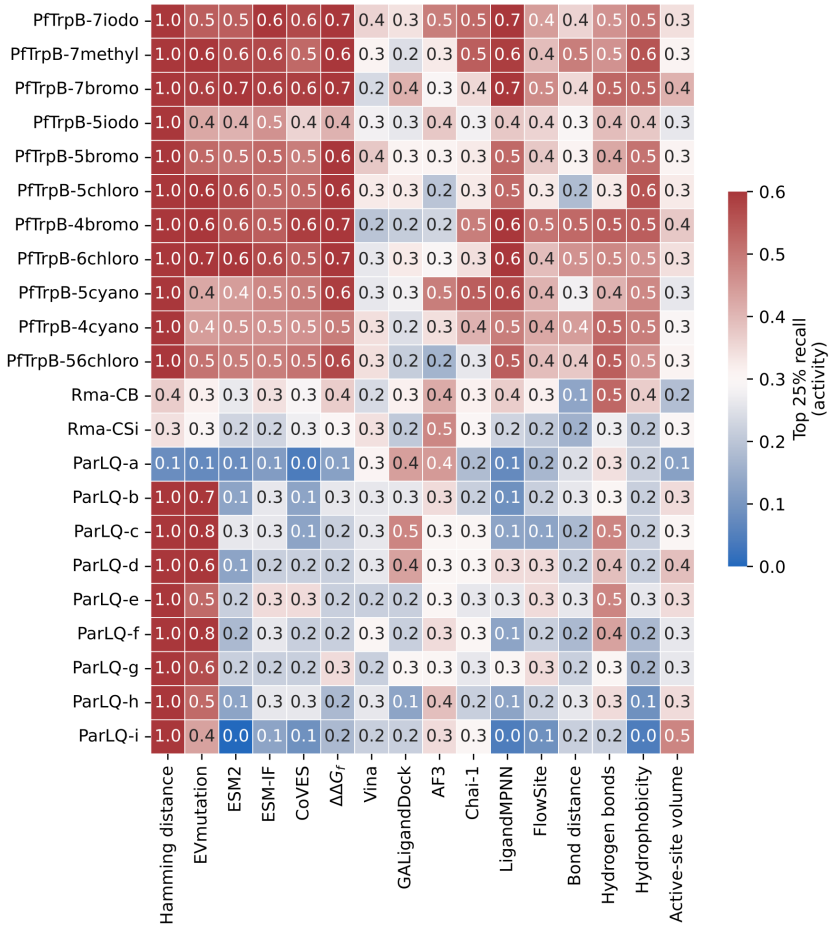

Figure A9: Top 25% recall for activity

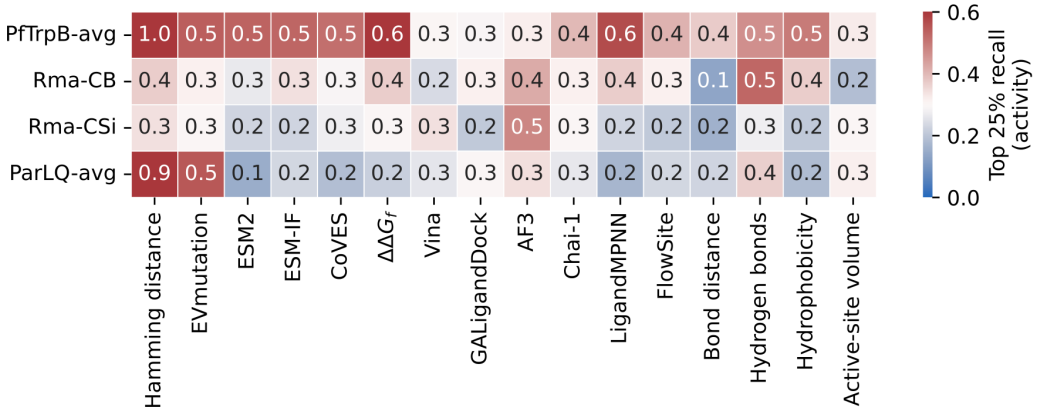

Figure A10: Top 25% recall for activity, averaged by chemistry

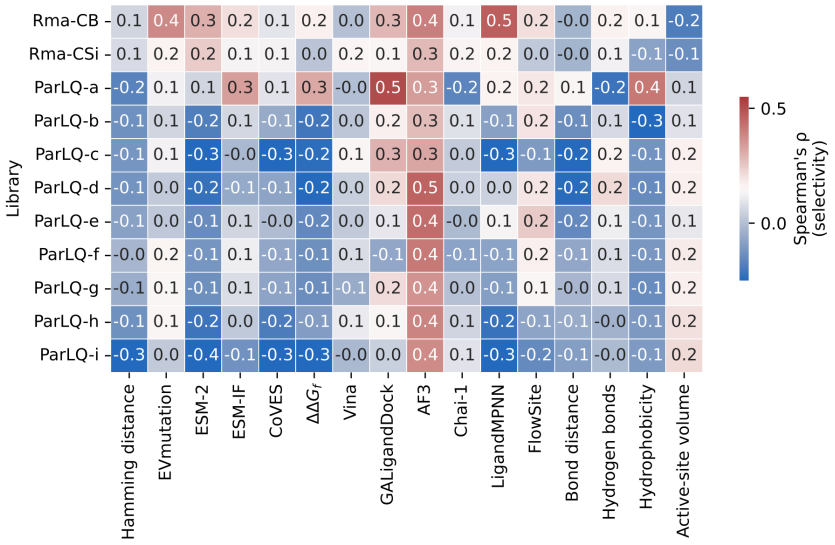

Figure A11: Spearmen's correlation for selectivity

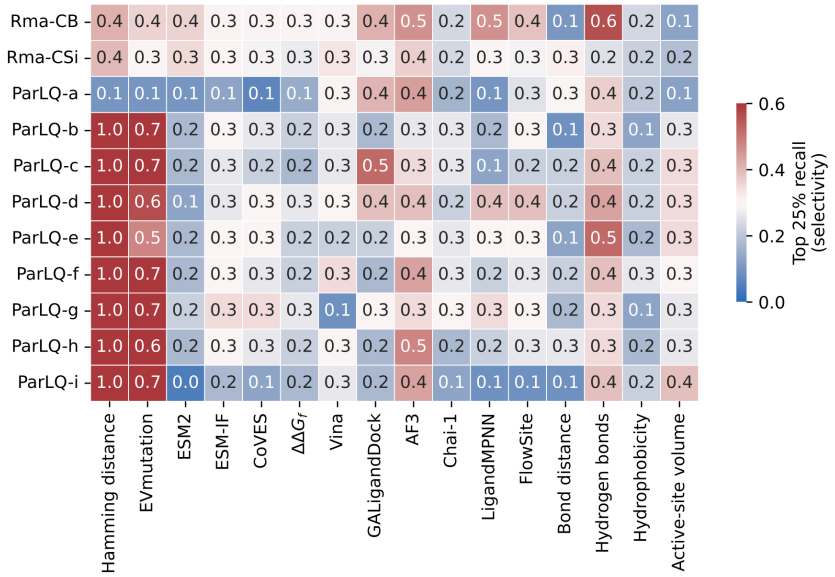

Figure A12: Top 25% recall for selectivity

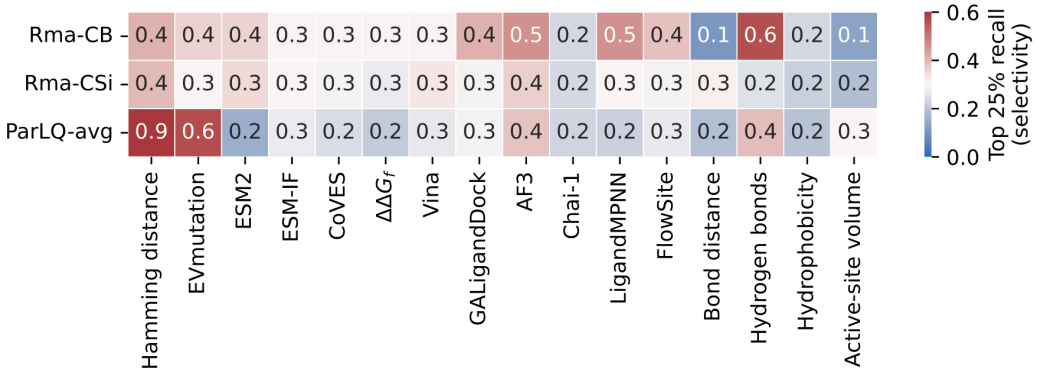

Figure A13: Top 25% recall for selectivity, averaged by chemistry

A.4.3 COMBINATION OF ZS PREDICTORS

Table A5: Spearman's ρ of 34 unweighted ensembles of ZS predictors generalized better than the top individual ZS across all chemistries. See Table A4 for topN predictors.

Predictor Combination	Average Spearman's ρ across all chemistries
All top2	0.3859
New-to-nature top8	0.3802
New-to-nature top9	0.3769
New-to-nature top5	0.3757
All top3	0.3738
New-to-nature top7	0.3708
New-to-nature top11	0.3598
New-to-nature top10	0.3588
New-to-nature top6	0.3579
All top4	0.3531
New-to-nature top12	0.3460
All top5	0.3457
New-to-nature top13	0.3395
New-to-nature top4	0.3390
All top6	0.3316
New-to-nature top14	0.3179
All top7	0.3162
New-to-nature top15	0.3091
All top12	0.3079
All top9	0.3046
All top14	0.3008
Non-native top16	0.3007
All top16	0.3007
New-to-nature top16	0.3007
All top13	0.3000
All top8	0.2983
Non-native top15	0.2978
All top15	0.2978
All top11	0.2943
Non-native top14	0.2938
All top10	0.2901
Non-native top12	0.2887
Non-native top11	0.2843
Non-native top13	0.2802

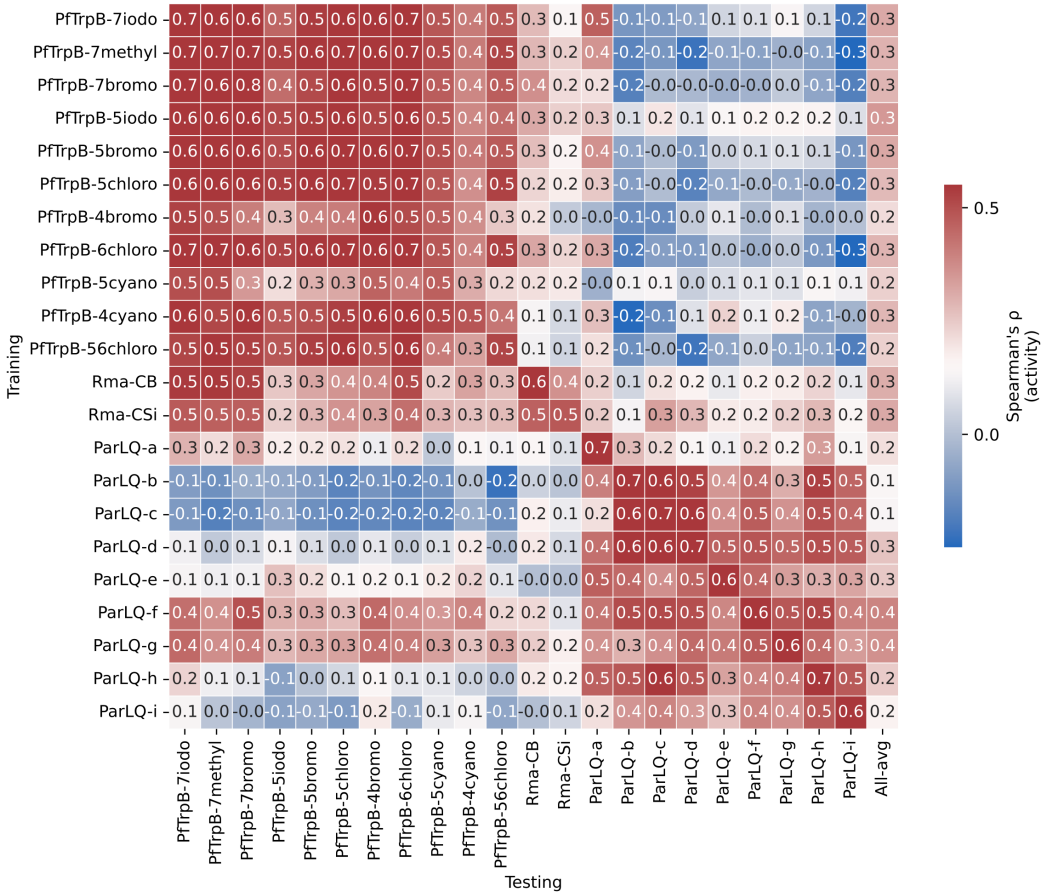

Figure A14: Linear regression model trained on one library with 16 ZS and tested on all

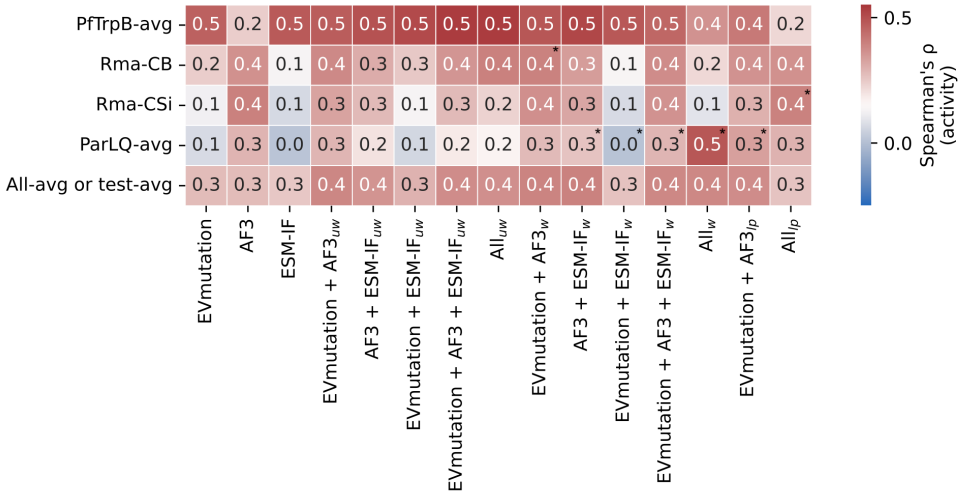

Figure A15: Different model and ZS predictor combinations for ensembling. uw refers to an unweighted combination. w refers weighted linear combination trained on the best dataset and tested on the rest. lp refers to piecewise linear models trained on the best dataset and tested on the rest. The $*$ symbol indicates the training set, which is excluded from the test-avg calculation.

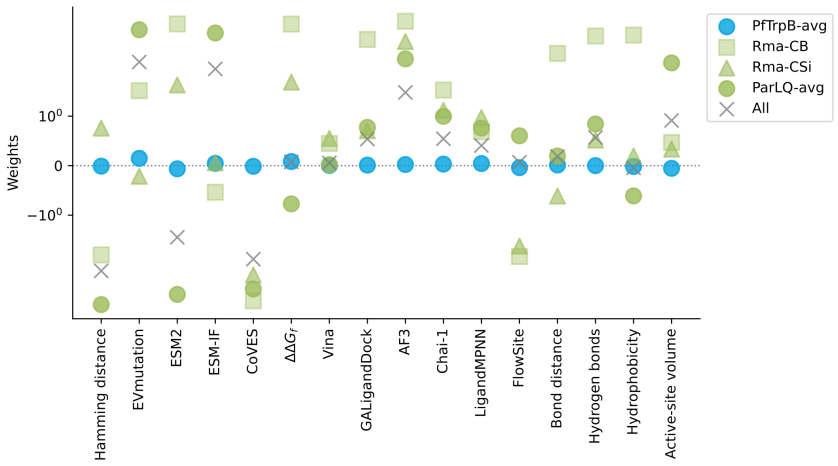

Figure A16: Averaged weights for linear regression model trained on one library with 16 ZS and tested on all

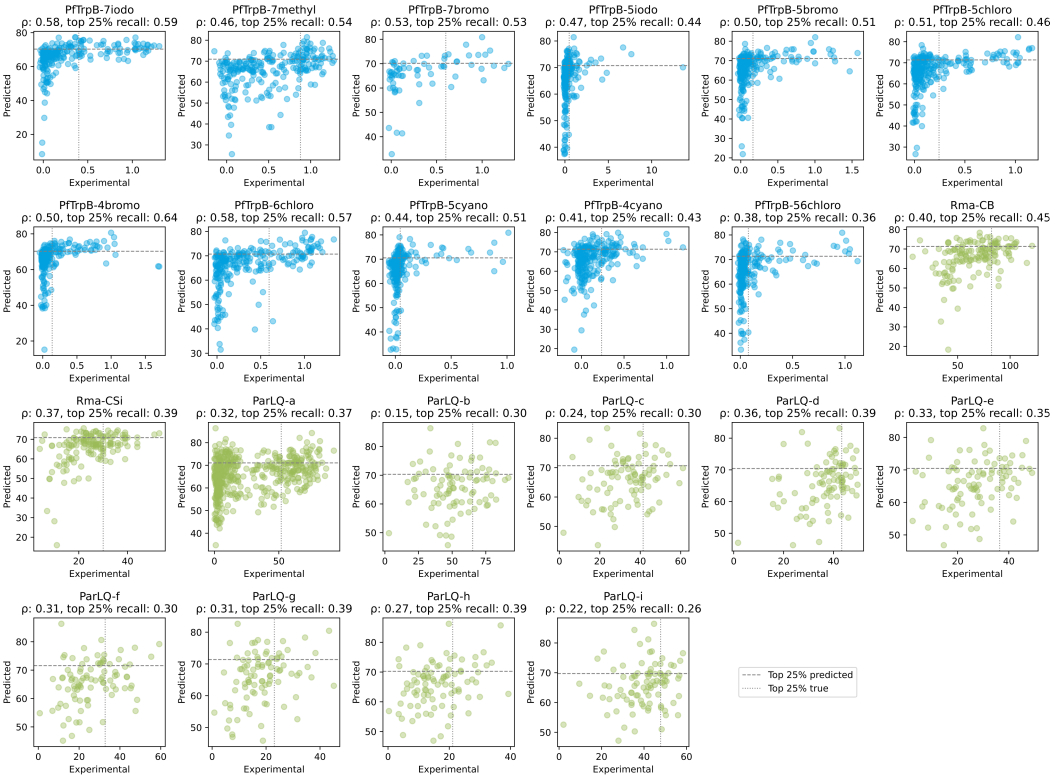

Figure A17: Scatter plot for EVmutation + AF3$_w$

A.4.4 ADDITIONAL TABLES

Table A6: Bitscore and sequence counts for *Pf*TrpB, *Rma* cyt *c*, and ParLQ. The **bold** row indicates the chosen MSA covering all the targeted sites.

Enzyme	Bitscore	Sequences
*Pf*TrpB	0.1	74795
	0.3	**5996**
	0.5	5935
	0.7	4647
Rma cyt *c*	0.1	Job exceeded resources
	0.3	79025
	0.5	**3042**
	0.7	1940
ParLQ	0.1	15086
	0.3	875
	0.5	343
	0.7	**343**

Table A7: Activity and selectivity Spearman's correlation

Library	Spearman's ρ
ParLQ-a	0.9610
ParLQ-b	0.7527
ParLQ-c	0.9335
ParLQ-d	0.9326
ParLQ-e	0.8971
ParLQ-f	0.8197
ParLQ-g	0.7097
ParLQ-h	0.9257
ParLQ-i	0.7618
Rma-CB	0.6438
Rma-CSi	0.4554

Table A8: Correlation between ZS predictions for activity and for selectivity, both measured by Spearman's correlation

ZS predictor	Spearman's ρ	p-value
Hamming distance	0.9455	1.12e-05
EVmutation	0.7818	0.0045
ESM-2	0.8545	0.0008
ESM-IF	0.7273	0.0112
CoVES	0.6818	0.0208
$\Delta\Delta G_f$	0.8727	0.0005
Vina	0.8091	0.0026
GALigandDock	0.9182	6.66e-05
AF3	0.3091	0.3550
Chai-1	0.9545	4.99e-06
LigandMPNN	0.8273	0.0017
FlowSite	0.6545	0.0289
Bond distance	0.5636	0.0710
Hydrophobicity	0.4909	0.1252
Hydrogen bonds	0.9364	2.21e-05
Active-site volume	0.9000	0.0002

Table A9: Tanimoto similarity of atom-pair fingerprints for *Pf*TrpB non-native substrates

Indole Analogs	Similarity to Indole
7iodo	0.6053
7methyl	0.6053
7bromo	0.6053
5iodo	0.6000
5bromo	0.6000
5chloro	0.6000
4bromo	0.5500
6chloro	0.5366
5cyano	0.4898
4cyano	0.4894
56chloro	0.3333

Table A10: Correlation between predictors and substrate similarity to the native substrate.

ZS predictors	Spearman's ρ	p-value
Hamming distance	0.3890	0.2371
EVmutation	0.6390	0.0343
ESM-2	0.6390	0.0343
ESM-IF	0.5371	0.0884
CoVES	0.1574	0.6439
$\Delta\Delta G_f$	0.6390	0.0343
Vina	0.3334	0.3164
GALigandDock	0.5371	0.0884
AF3	0.3982	0.2251
Chai-1	0.3241	0.3308
LigandMPNN	0.3982	0.2251
FlowSite	0.1574	0.6439
Bond distance	0.0185	0.9569
Hydrophobicity	0.5464	0.0820
Hydrogen bonds	0.3612	0.2751
Active-site volume	0.0370	0.9139

Table A11: Calculated reaction energy barrier (kcal/mol)

Chemistry	Energy barrier
ParLQ	~ 9
Rma-CB	~ 11
Rma-CSi	~ 12

Table A12: Calculated reaction energy ΔG (kcal/mol) considering all substrates and products for new-to-nature chemistries

Chemistry	ΔG
ParLQ-a	-44.6075
ParLQ-b	-44.7320
ParLQ-c	-44.5280
ParLQ-d	-46.3590
ParLQ-e	-45.8877
ParLQ-f	-45.5328
ParLQ-g	-46.0807
ParLQ-h	-75.2704
ParLQ-i	-45.5365
Rma-CB	-54.8434
Rma-CSi	-62.6220

Table A13: Correlation between reaction energy and ZS predictor performance

ZS predictor	Spearman's ρ	p-value
Hamming distance	0.4455	0.1697
EVmutation	0.1273	0.7092
ESM-2	0.3000	0.3701
ESM-IF	-0.0545	0.8734
CoVES	0.3455	0.2981
$\Delta\Delta G_f$	0.2818	0.4011
Vina	0.2000	0.5554
GALigandDock	-0.1000	0.7699
AF3	0.4909	0.1252
Chai-1	0.3818	0.2466
LigandMPNN	0.3091	0.3550
FlowSite	0.0273	0.9366
Bond distance	-0.0364	0.9155
Hydrogen bonds	0.1364	0.6893
Hydrophobicity	0.1273	0.7092
Active-site volume	0.0455	0.8944