

Workflow for Biocatalytic Reaction Performance Prediction and Analysis

Hanna D. Clements^{1†}, Autumn R. Flynn^{1†}, Bryce T. Nicholls², Daria Grosheva², Todd K. Hyster^{*2} and Matthew S. Sigman^{*,1}

[†]These authors contributed equally.

¹Department of Chemistry, University of Utah, 315 South 1400 East, Salt Lake City, Utah 84112, United States

²Department of Chemistry and Chemical Biology, Cornell University, 122 Baker Laboratory, Ithaca, New York 14853

Contact email: matt.sigman@utah.edu; thyster@cornell.edu

Abstract

The development of predictive tools to assess enzyme mutant performance and physical organic approaches to enzyme mechanistic interrogation are crucial to the field of biocatalysis. While many indispensable tools exist to address qualitative aspects of biocatalytic reaction design, they often require extensive experimental data sets or *a priori* knowledge of reaction mechanism. However, quantitative prediction of enzyme performance is lacking. Herein, we present a workflow that merges both computational and experimental data to produce statistical models that predict the performance of new substrates and enzyme mutants while also providing insight into reaction mechanism. As a validating case study, this platform was applied to investigate a non-native enantioselective photoenzymatic radical cyclization. Statistical models enabled interrogation of the reaction mechanism, and the predictive capabilities of these same models led to the quantitative prediction of the enantioselectivities of new substrates with several enzyme mutants. This platform was constructed for application to any biocatalytic system wherein mechanistic interrogation, prediction of reaction performance with new substrates, or quantitative performance of enzyme mutants would be desirable. Overall, this proof of concept study provides a new tool to complement existing protein engineering and reaction design strategies.

Introduction

Enzymes play a significant role as selective, efficient catalysts for biotechnology, biomedicine, biofuels, and industrial pharmacology.^{1–4} Contemporary investigation of non-natural biocatalytic transformations often relies on robust screens of mutant space through protein engineering (e.g., directed evolution (DE)).^{5,6} Although exceptionally effective, engineering campaigns rarely lead to a deeper understanding of the biocatalytic mechanism that affords the desired product, and detailed computational analyses are often required to uncover the interactions between a biocatalyst and reactant that facilitate reactivity and selectivity.^{7–9} Further, translation of the optimized enzyme to new reactants can be challenging. Specifically, expansion of reaction scope often requires additional rounds of engineering or DE optimization,¹⁰ and mechanistic models may need to be (re)developed to explain the observed selectivity of each enzyme/substrate pair. These tasks limit the accessibility of biocatalysis to labs that lack the infrastructure for high throughput experimentation and/or computational resources.

Thus, we were motivated to develop a strategy to concurrently optimize biocatalytic reactions and gain mechanistic insight, with an emphasis on simultaneous exploration of sequence (enzyme mutants) and chemical (substrate scope) space. Specifically, we envisioned a tool that would quantitatively relate the molecular features of enzymes and substrates to observables like enantioselectivity using statistical models. This would require the design and acquisition of empirical results for a matrix of both substrate variants and enzyme mutants, and computational characterization of the reaction components.¹¹ The empirical data would be regressed against computed molecular features, resulting in statistical models that would

provide predictive power and molecular-level insights into the origin of substrate and enzyme performance. Although ideologically similar to quantitative structure/sequence activity relationships, our attention to molecular conformational ensembles, simultaneous parameterization of both enzyme and substrate, use of multivariate linear regression models, and implementation of higher-level descriptors would result in models with greater interpretability and generalizability.^{12–14}

We designed a workflow to accommodate several common types of empirical data, including enantioselectivity, conversion, regioselectivity, or substrate specificity. Notably, this process would not require large data sets and would be adoptable by laboratories with a range of experimental and computational capabilities. Herein, we demonstrate the viability of this general strategy in the context of non-native enantioselective photoenzymatic radical cyclization reactions catalyzed by ‘ene’-reductase variants from *Gluconobacter oxydans* (GluER, Fig. 1a).¹⁵ This strategy successfully guided GluER substrate scope expansion to reduce experimental screening efforts, and the statistical models rationalized the observed and predicted selectivities of previously untested substrates. Unlike current technologies for *in silico* protein engineering, which require additional datasets for each new substrate,^{16–18} our method accounts for both substrate and enzyme features, allowing for knowledge transfer to predict the performance of new substrate and mutant combinations.

Design of a broadly accessible workflow

We selected GluER-T36A as a biocatalytic framework to study as it has recently been utilized in a range of non-native enantioselective reactions with potential applications in the chemical industry. Specifically, GluER-T36A is a selective catalyst for the photoenzymatic cyclization of many α -chloroamides (Fig. 1a). However, a number of substrates required alternative enzymes to achieve high enantioselectivity in the initial report.¹⁵ Although effective in this instance, the general

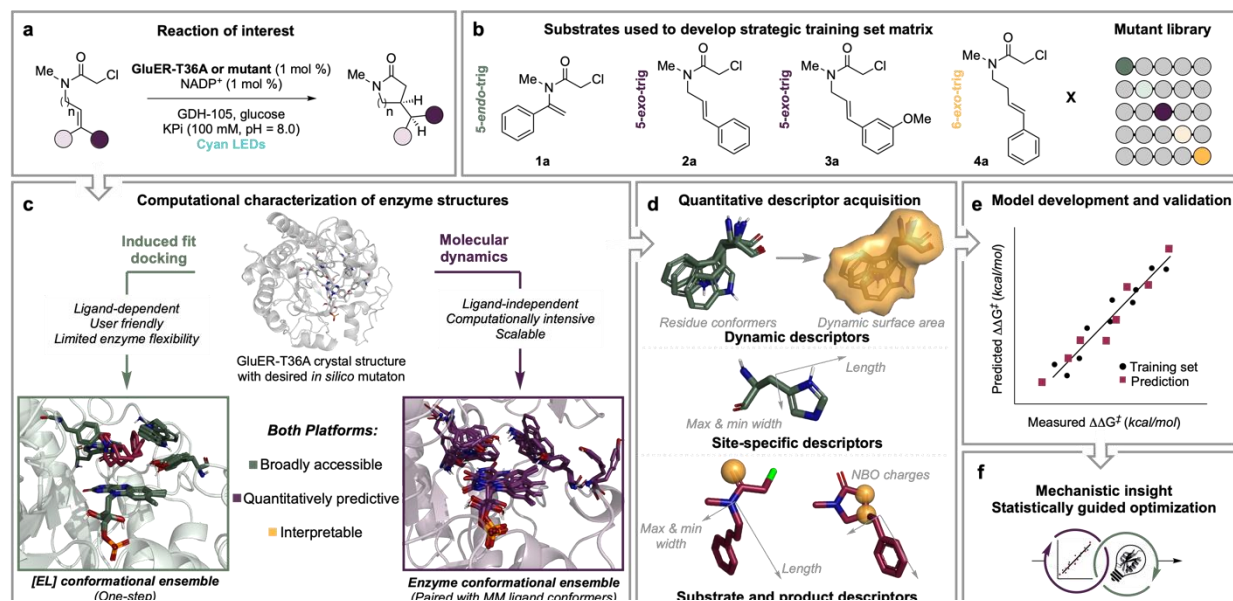


Figure 1. Workflow to develop statistical models of enzyme performance.

a, The workflow was developed using the enantioselective photoenzymatic cyclization of α -chloroamide substrates facilitated by GluER-T36A. **b**, Substrates **1a–4a** were subjected to GluER-T36A or mutant to generate a training set for model development. **c**, Two complementary approaches were developed to generate enzyme conformers from a GluER-T36A crystal structure (PDB ID: 6MYW) after introducing the desired mutation *in silico*. **d**, Enzyme features were quantified using a residue and site-based approach. Dynamic parameters were measured by overlaying a residue conformational ensemble, encapsulating it in a fictitious surface, and measuring the resulting surface area and volume. Site-specific descriptors were generated by measuring the length, width, and backbone angles of each residue conformer, and the fluctuations of these measurements. Ligands were subjected to both a geometric analysis and DFT calculations to acquire electronic descriptors, including the NBO charges of atoms indicated by a yellow sphere. **e**, Descriptors for each enzyme/ligand were regressed against the experimentally determined selectivities, resulting in statistical models that were validated using leave-one-out and k-fold cross-validation. **f**, Select statistical models were interrogated for mechanistic insight and used to predict the selectivity of untested enzyme and substrate combinations.

practice of shifting enzyme frameworks can lead to unexpected results (i.e., enantiodivergent transformations or unexpected products), and introduce biosynthetic challenges including re-optimization of expression and reaction conditions.^{19,20} An alternative tactic to improve underperforming substrates would be to engineer the GluER-T36A active site to each substrate;²¹ however, site-saturation mutagenesis (SSM) of even five active-site residues would require 100 individual variants to be constructed, expressed, and evaluated in the laboratory. Additionally, reaction improvements in a SSM library rarely transfer to new substrates, necessitating evaluation of existing or new mutant libraries to optimize for the best performing variant.²²

By relying on a small, representative training set that encompasses a range of reaction outputs, we hypothesized that robust statistical models relating function to structural features could be developed to predict the performance of new substrates or substrate/mutant combinations. Importantly, this method would draw from all training set data, including mutants that did not enhance selectivity. In this context, we designed a focused training set with diversity in both substrate characteristics and enzyme mutations. The transformation of substrates **1a–4a** (Fig. 1b) encompass three different cyclization modes (**1a** : 5-*endo*, **2a**, **3a**: 5-*exo*, and **4a**: 6-*exo*), varying electronic properties (as in **2a** and **3a**), and alkene substitution pattern (**1a** vs **2a–4a**). We also identified five residues within the GluER-T36A active site for mutation: W66, Y177, Q232, F269, and Y343 (Fig. S1) and subjected each site to site-directed mutagenesis to introduce residues W, F, D, L, or A, as these mutations would sample a wide range of active site properties. Substrates **1a–4a** were subjected to reaction with each expressible mutant, resulting in a total of 50 datapoints to use in model training and selection. A full table of substrate/enzyme combinations and the resultant e.e.'s are included in Table S1.

We next considered strategies to computationally characterize both enzyme and ligand (substrate and product) structures for descriptor extraction (Fig. 1c). We have previously described methods to characterize small molecule catalysts and extract chemical descriptors by pairing molecular mechanics (MM) based structural analysis with density functional theory (DFT) calculations.^{23,24} However, the shift from small chemical systems to biocatalytic platforms presents a number of unique challenges. The size and elaborate dynamics of enzymes has necessitated bespoke computational strategies to study the dynamics of enzyme ligand complexes [EL], however many of these are best suited for in-depth analysis of one or a few [EL] pairs due to their operational complexity and resource demands.^{25,26} We therefore sought workflows that would account for the dynamic nature of biocatalysts while also introducing operational simplicity, scalability, and consideration of ligand interactions.

We identified two complementary conformational search platforms: Induced fit docking (IFD) and accelerated molecular dynamics (aMD) (Fig. 1c). IFD is a MM-based docking protocol that is widely utilized by medicinal chemists to approximate the docking pose of a ligand and the concomitant repositioning of nearby enzyme residues.^{27–29} We generated ensembles of the [EL] by hijacking the IFD protocol to describe both ligand and enzyme repositioning. In addition to the robust [EL] structural data generated by IFD, we integrated IFD into our workflow because it is easily implemented and does not require sophisticated computing hardware. To complement the IFD workflow, we desired a scalable platform to rapidly scan enzyme mutant space. Therefore, we employed aMD, which allows for flexibility in the entire enzyme, to quickly assemble *apo*-enzyme structures that represent the dynamic ensemble.^{30,31} These structures were paired with free ligand conformational ensembles from MM/DFT. Since enzymes and substrates are assembled separately in the aMD sampling platform, this approach is applicable to large enzyme-ligand matrices and has potential for virtual screening.

Upon acquisition of the enzyme and ligand conformational ensembles with either IFD or aMD, quantitative chemical descriptors were computed, automatically extracted, and curated for the ligands as well as for individual residues in the

active site (Fig. 1d). These descriptors included electronic (e.g., natural bond orbital (NBO) charges),³² steric (e.g., sterimol values),³³ and dynamic descriptors, which describe topographical properties of a collection of conformers (e.g., dynamic surface area, (DSA)).³⁴ Describing the active site by its individual residues in this manner reveals which residues have the most influence on reaction enantioselectivity.

These descriptors were regressed against the experimentally collected dataset (70:30 split of training:test set data points) using a forward-stepwise multivariate linear regression (MLR) algorithm, which resulted in thousands of candidate models for each conformational search platform (Fig. 1e).²³ From these candidate models, we identified a representative high-performing IFD statistical model (Fig. 2a), which had a Training R^2 of 0.83, and a mean absolute error (MAE) of 0.18 kcal/mol, indicating a good correlation between the measured and predicted values of the training set. The Test R^2 is the correlation between measured and predicted values for the test set (the partition of data that was withheld from model training); the IFD model had a Test R^2 of 0.57 and a corresponding Test MAE of 0.29 kcal/mol, which was overall satisfactory. The selected aMD statistical model (Fig. 2b) demonstrated a Training R^2 of 0.82 with a MAE of 0.19 kcal/mol; therefore, the IFD and aMD models performed similarly in their capability to describe the data in the training set. The aMD model exhibited a Test R^2 of 0.73 and Test MAE = 0.19 kcal/mol, indicating it had an improved predictive capability compared to the IFD model. Successful identification of statistical models from both IFD and aMD workflows validated our hypothesis that molecular features of enzymes and ligands can describe the outcome of a biocatalytic reaction. With these statistical models, we were primed to interrogate the mechanistic features responsible for reaction performance and ultimately predict unseen enzyme/substrate combinations.

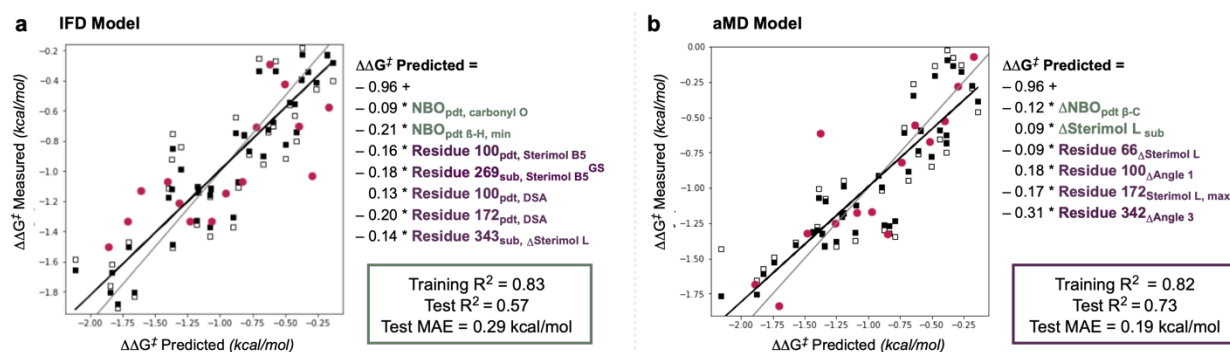


Figure 2. Statistical models of GluER-T36A and variant selectivity.

a. The IFD model had a training and test R^2 of 0.83 and 0.57, respectively, a leave-one-out (LOO) R^2 of 0.73, and a 4-fold R^2 of 0.70. The model had ligand descriptors (green) and enzyme descriptors (purple). The $NBO_{pdt, carbonyl O}$ is the NBO charge of the carbonyl oxygen from product structures. The $NBO_{pdt, 8-H, min}$ describes the minimum NBO charge of the hydrogen bound to the β -carbon in product structures. Residue 100_{pdt, Sterimol B5} is the maximum width of residue 100 from product-docked enzyme structures. Residue 269_{sub, Sterimol B5^{GS}} is the G-Score (docking-score) weighted maximum width of residue 269 from the substrate-docked enzyme structures. Residue 100_{pdt, DSA} and Residue 172_{pdt, DSA} are the dynamic surface areas of residues 100 and 172 from product-docked enzyme structures, respectively. Residue 343_{sub, ΔSterimol L} is the difference in the maximum and minimum length values (flexibility) of residue 343 from substrate-docked enzyme structures. **b.** The aMD model had a training and test R^2 of 0.82 and 0.73, respectively, LOO R^2 of 0.70, and 4-fold R^2 of 0.67. The aMD model also included ligand descriptors (green), and enzyme descriptors (purple). The $\Delta NBO_{pdt, \beta-C}$ is the difference in the maximum and minimum values of the NBO charge on the β -carbon of product structures. The $\Delta Sterimol L_{sub}$ is the difference in the maximum and minimum substituent length values (flexibility) of substrate structures. Residue 66_{ΔSterimol L} is the difference in the maximum and minimum residue length values (flexibility) of residue 66. Residue 100_{ΔAngle 1} and Residue 342_{ΔAngle 3} are the difference in the maximum and minimum Angle 1 and Angle 3 (see SI) values of residues 100 and 343, respectively. Residue 172_{Sterimol L, max} is the maximum length of residue 172.

Mechanistic Interpretability

Unlike other machine learning (ML) technologies, the statistical models resulting from this strategy are interpretable at the molecular level. The parameters in the IFD model included the NBO charge of the hydrogen on the β -carbon from

product structures, which classified whether the radical cyclization mechanism is 5-*exo* or not, and the NBO charge of the carbonyl oxygen, which further differentiated between ring sizes (Fig. 3a, left). The first free ligand parameter in the aMD model (the NBO charge of the β -carbon) similarly classified 5-*exo* cyclization. The second parameter in the aMD model (Δ Sterimol L_{sub} , describes the flexibility of the substrate's alkene substituent) differentiated the resulting ring sizes in product structures. The positive coefficient accompanying the Δ Sterimol L_{sub} term was associated with lower selectivity when Δ Sterimol L_{sub} is large, as in **4a** (Fig. 3a).

Both aMD and IFD statistical models include features that describe how particular H172 conformations enhance selectivity, indicated by the negative sign associated with the coefficients. The aMD parameter corresponding to this residue measures the maximally extended conformation of H172 for each GluER-T36A variant. When H172 was extended, nearby residues N175 and Y177 were concomitantly displaced, resulting in the formation of a distinct binding pocket, shown as yellow spheres in both the cartoon and the aMD structure of GluER-T36A-F269L (Fig. 3b). Enzyme variants where H172 was retracted (Fig. 3b, GluER-T36A-Y177W), resulted in occlusion of the putative binding pocket by nearby residues, leading to diminished enantioselectivity. We hypothesized an open binding site, as in GluER-T36A-F269L, resulted in facile substrate binding without significant steric rejection, which could induce substrate detrimental dissociation or rotation. These observations are also corroborated by the high conservation of H172 across 'ene'-reductase families, and studies on oxyanion holes in 'ene'-reductases.³⁵

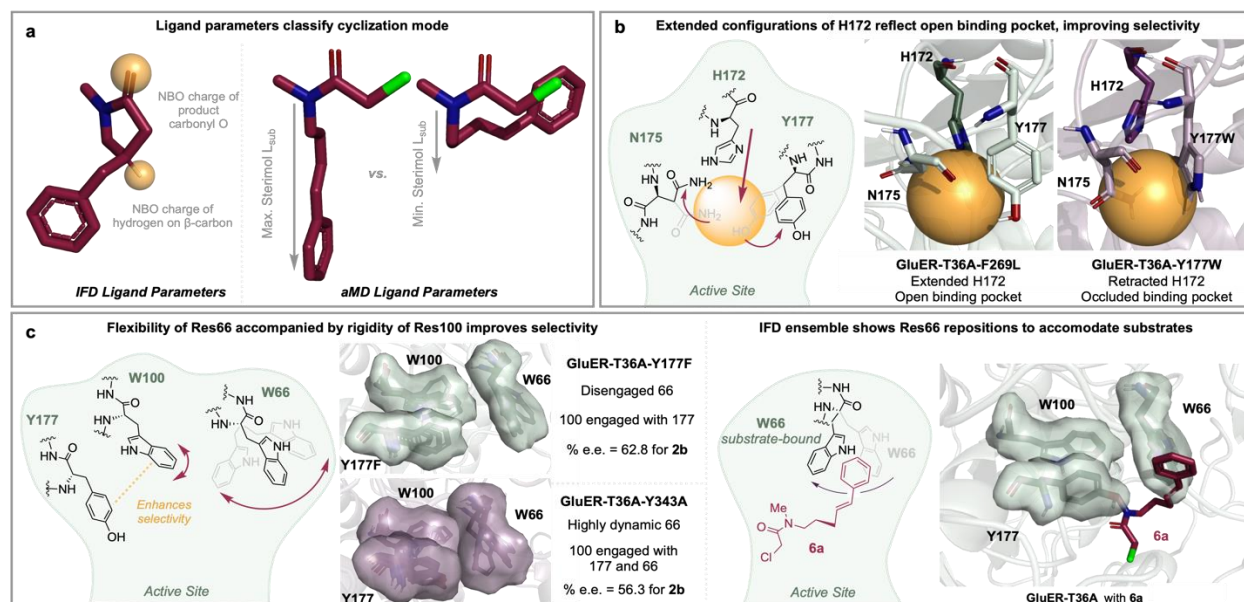


Figure 3. Mechanistic interpretation of descriptors in statistical models.

a, Illustration of IFD and aMD ligand parameters from selected statistical models. Five-membered rings with aryl substituents typically had a more positive carbonyl oxygen NBO charge (greater than -0.73). The NBO charge of the hydrogen attached to the product β -carbon was generally less than 0.2 for the products that resulted from 5-*exo* cyclizations (**2a** and **3a**). The aMD steric descriptor (Δ Sterimol L_{sub}) describes the flexibility of the substrate; it distinguished the most flexible substrate **4a** (unscaled Δ Sterimol L_{sub} = 6.64 Å) from the less flexible substrates (**1a** = 0.14 Å, **2b** = 1.19 Å). **b**, The Residue 172_{Sterimol L_{max}} term from the aMD model indicated that extended configurations of H172 facilitated selectivity. Examination of enzyme conformers where this term was large (GluER-T36A-F269L = 6.7 Å) showed H172 to be extended (green) and revealed an open binding pocket (yellow sphere); this binding pocket was occluded in structures where values of this parameter were small (GluER-T36A-Y177W = 5.2 Å, purple). **c**, The aMD conformers demonstrate that when aromatic residues 100 and 177 were closely associated (green), interactions between residues 66 and 100 were precluded, which induced residue 66 flexibility and higher selectivity. The IFD conformational ensembles (right) corroborated that flexibility of residue 66 is necessary for substrate binding.

The statistical models also revealed that positioning of the (natively) aromatic residues 66 and 100 affected selectivity in the radical cyclization.^{36,37} The positive coefficients associated with features describing residue 100 flexibility communicated that rigidity of this residue lead to greater enantioselectivity, while the negative coefficient associated with residue 66 in the aMD model (Residue 66 Δ sterimol L) indicated that dynamic behavior of this residue facilitated selective transformations. To gain a deeper understanding of these effects, we interrogated the aMD conformational ensembles, which implied that W100 was involved in a network of competitive non-covalent interactions (NCIs) with flanking residues 66 and 177. Our analysis led us to postulate that when W100 preferentially interacts with residue 177, engagement with residue 66 is precluded. This results in increased mobility of residue 66, which allows residue repositioning and substrate binding. In concordance, IFD revealed significant repositioning of residue 66 upon substrate **6a** binding (relative to *apo*, Fig. 3c, right).

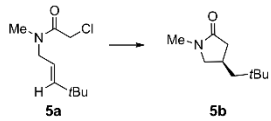
The importance of these aromatic NCIs is further supported by the lack-of-function observed when Y177 is mutated to a non-aromatic residue. Y177A/D/L mutants did not afford products with any substrate, with the exception of Y177A with **1a**. Interestingly, the differential alkene geometry of **1a** (compared to **2a-4a**) positioned the substrate alkene substituent on the opposite side of the active site such that the ligand did not necessitate residue 66 repositioning, as revealed by IFD (Fig. S4).

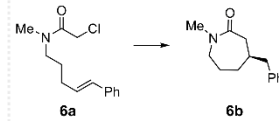
Demonstration of predictive capability

As a final validation of the workflow, the prediction of out-of-sample combinations of substrates and enzyme variants were explored. Compared to established procedures to predict biocatalyst selectivity, the statistical models presented herein possessed the unique ability to evaluate substrates that were not included in model training. We therefore used both IFD and aMD statistical models to predict the performance of various GluER-T36A mutants with two new substrates: **5a** and **6a** (Table 1). These were selected to incorporate substrate characteristics that were not represented in the training set, including an alkyl substituted example (**5a**), and a different cyclization mode (**6a**, 7-*exo*). For the aMD model, **5a** and **6a** conformers were collected and combined with existing enzyme trajectories, and for IFD the relevant [EL] poses were generated. Descriptors for these ensembles were automatically extracted. Using these descriptors, the enantioselectivities of **5a** and **6a** with each GluER-T36A variant were predicted using the statistical models from Figure 2 (Table 1).

Gratifyingly, experimental evaluation of these combinations revealed the IFD model successfully predicted the enantioselectivity of 50% of the reactions within one MAE (0.29 kcal/mol). Within two MAE (0.58 kcal/mol), the IFD

Table 1.^a Predicted selectivity of new substrates with GluER-T36A mutants.





	Entry	% e.e. pred.	% e.e. meas.		Entry	% e.e. pred.	% e.e. meas.
IFD Predictions	5-W66A	17.5 - 56.7	21.1	6-W66A	19.8 - 58.3	55.1	
	5-W66L	61.4 - 82.8	8.8	6-W66L	15.9 - 55.6	46.2	
	5-Y177F	51.0 - 77.4	36.0	6-Y177F	56.7 - 80.4	54.6	
	5-Q232F	70.1 - 87.0	2.9	6-Q232F	47.3 - 75.3	71.2	
	5-Y343A	70.1 - 87.0	29.1	6-Y343A	27.4 - 63.4	58.3	
	5-Y343F	53.3 - 78.6	71.6	6-Y343F	41.5 - 72.0	59.7	
	5-Y343W	62.9 - 83.5	37.8	6-Y343W	73.2 - 88.5	66.5	
	Entry	% e.e. pred. ^b	% e.e. meas.		Entry	% e.e. pred.	% e.e. meas.
aMD Predictions	5-W66A	10.0 - 39.9	21.2	6-W66A	42.2 - 65.3	55.1	
	5-W66L	6.1 - 36.5	8.8	6-W66L	41.5 - 63.4	46.2	
	5-Y177F	20.1 - 48.1	36.0	6-Y177F	52.8 - 71.3	54.6	
	5-Q232F	26.6 - 55.3	2.9	6-Q232F	68.8 - 73.9	71.2	
	5-Y343A	43.2 - 65.5	29.1	6-Y343A	83.5 - 90.7	58.3	
	5-Y343F	66.7 - 81.0	71.6	6-Y343F	83.3 - 90.9	59.7	
	5-Y343W	70.8 - 83.5	37.8	6-Y343W	85.2 - 91.7	66.5	

■ correct within one MAE

■ incorrect

■ correct within two MAE

■ correct within one MAE ■ incorrect ■ correct within two MAE

^a Enantioselectivities of reactions with **5a** and **6a** to form **5b** and **6b**, predicted from the IFD and aMD models. The range of predicted selectivities from IFD was derived from the mean absolute error (MAE) of 0.29 kcal/mol. The range for the aMD predictions was derived from an MAE of 0.19 kcal/mol. ^b Predicted from aMD Model 2, Fig. S3.

model successfully predicted 71% of enantioselectivities. The aMD model successfully predicted 57% of reaction enantioselectivities within one MAE (0.19 kcal/mol) and 64% within two MAE (0.38 kcal/mol). These results demonstrated the ability of these statistical models to predict *quantitative* enantioselectivity data of unseen substrates and enzyme combinations, which has been a historic limitation in other ML, rational design, or computational enzyme investigation^{38,39}

Conclusion

In summary, a general approach was disclosed to quantitatively relate enzyme and ligand features to experimental observables; specifically, enzyme and ligand structures from either IFD or aMD/MM were related to the observed enantioselectivity in a photoenzymatic radical cyclization using statistical models. The resultant statistical models were mechanistically insightful and provided new perspectives on the origin of enantioselectivity in GluER systems. Furthermore, the utility of the statistical modeling strategy was demonstrated by predicting the enantioselectivity for out-of-sample combinations of mutants and substrates, with a particular emphasis on substrate scope to complement contemporary predictive approaches. Future applications of this workflow will include enhancement of enzymatic descriptors and virtual screening of enzyme mutants for reaction engineering.

Acknowledgments

The computational portion of this work was supported by the Center for High Performance Computing (CHPC) at the University of Utah. This research was supported by the National Institutes of Health National Institute of General Medical Sciences (R01 GM127703 to T.K.H and R35 GM136271 to M.S.S.). This research made use of the Cornell University NMR Facility, which is supported, in part, by the NSF through MRI Award CHE-1531632. DG was supported by a Swiss National Science Foundation Early Postdoc Mobility Fellowship (no number)

References

1. Huffman, M. A. *et al.* Design of an in vitro biocatalytic cascade for the manufacture of islatravir. *Science*. **366**, 1255–1259 (2019).
2. Meghwanshi, G. K. *et al.* Enzymes for pharmaceutical and therapeutic applications. *Biotechnol. Appl. Biochem.* **67**, 586–601 (2020).
3. Duza, M. B. & Mastan, S. A. Microbial Enzymes and Their Applications – a Review. *Indo Am. J. Pharm. Res.* **3**, 651–657 (2013).
4. Akyilmaz, E., Yorganci, E. & Asav, E. Do copper ions activate tyrosinase enzyme? A biosensor model for the solution. *Bioelectrochemistry* **78**, 155–160 (2010).
5. Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem. Int. Ed.* **57**, 4143–4148 (2018).
6. Bornscheuer, U. T. *et al.* Engineering the third wave of biocatalysis. *Nature* **485**, 185–194 (2012).
7. Narayan, A. R. H. *et al.* Enzymatic hydroxylation of an unactivated methylene C-H bond guided by molecular dynamics simulations. *Nat. Chem.* **7**, 653–660 (2015).
8. Sheng, X., Kazemi, M., Ządło-Dobrowolska, A., Kroutil, W. & Himo, F. Mechanism of Biocatalytic Friedel-Crafts Acylation by Acyltransferase from *Pseudomonas protegens*. *ACS Catal.* **10**, 570–577 (2019).

9. Wittmann, B. J., Yue, Y. & Arnold, F. H. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Syst.* **12**, 1026–1045 (2021).
10. Reetz, M. T. & Wu, S. Laboratory evolution of robust and enantioselective Baeyer-Villiger monooxygenases for asymmetric catalysis. *J. Am. Chem. Soc.* **131**, 15424–15432 (2009).
11. Crawford, J. M., Kingston, C., Toste, F. D. & Sigman, M. S. Data Science Meets Physical Organic Chemistry. *Acc. Chem. Res.* **54**, 3136–3148 (2021).
12. Fox, R. J. *et al.* Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* **25**, 338–344 (2007).
13. Niu, J. & Yu, G. Molecular structural characteristics governing biocatalytic chlorination of PAHs by chloroperoxidase from *Caldariomyces fumago*. *SAR QSAR Environ. Res.* **15**, 159–167 (2004).
14. Wittmann, B. J., Johnston, K. E., Wu, Z. & Arnold, F. H. Advances in machine learning for directed evolution. *Curr. Opin. Struct. Biol.* **69**, 11–18 (2021).
15. Biegasiewicz, K. F. *et al.* Photoexcitation of flavoenzymes enables a stereoselective radical cyclization. *Science (80-.).* **364**, 1166–1169 (2019).
16. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
17. Wu, Z., Jennifer Kan, S. B., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Erratum: Machine learning-assisted directed protein evolution with combinatorial. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 788–789 (2020).
18. Cadet, F. *et al.* A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes. *Sci. Rep.* **8**, 1–15 (2018).
19. Page, C. G. *et al.* Quaternary Charge-Transfer Complex Enables Photoenzymatic Intermolecular Hydroalkylation of Olefins. *J. Am. Chem. Soc.* **143**, 97–102 (2021).
20. Huang, X. *et al.* Photoenzymatic enantioselective intermolecular radical hydroalkylation. *Nature* **584**, 69–74 (2020).
21. Wittmann, B. J. *et al.* Diversity-Oriented Enzymatic Synthesis of Cyclopropane Building Blocks. *ACS Catal.* **10**, 7112–7116 (2020).
22. Sullivan, B., Walton, A. Z. & Stewart, J. D. Library construction and evaluation for site saturation mutagenesis. *Enzyme Microb. Technol.* **53**, 70–77 (2013).
23. Santiago, C. B., Guo, J. Y. & Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.* **9**, 2398–2412 (2018).
24. Sigman, M. S., Harper, K. C., Bess, E. N. & Milo, A. The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and beyond. *Acc. Chem. Res.* **49**, 1292–1301 (2016).
25. Ahmadi, S. *et al.* Multiscale modeling of enzymes: QM-cluster, QM/MM, and QM/MM/MD: A tutorial review. *Int. J. Quantum Chem.* **118**, 1–34 (2018).

26. Van Der Kamp, M. W. & Mulholland, A. J. Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. *Biochemistry* **52**, 2708–2728 (2013).
27. Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A. & Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **49**, 534–553 (2006).
28. Sherman, W., Beard, H. S. & Farid, R. Use of an induced fit receptor structure in virtual screening. *Chem. Biol. Drug Des.* **67**, 83–84 (2006).
29. Farid, R., Day, T., Friesner, R. A. & Pearlstein, R. A. New insights about HERG blockade obtained from protein modeling, potential energy mapping, and docking studies. *Bioorganic Med. Chem.* **14**, 3160–3173 (2006).
30. Case D. A. *et al.* AMBER 2018. University of California, San Francisco. (2018)
31. Hamelberg, D., Mongan, J. & McCammon, J. A. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **120**, 11919–11929 (2004).
32. Glendening, E. D., Landis, C. R. & Weinhold, F. NBO 6.0: Natural bond orbital analysis program. *J. Comput. Chem.* **34**, 1429–1437 (2013).
33. Verloop, A. *Drug Design, Vol. III*. (Academic Press: New York, 1976).
34. Guo, J. Y., Minko, Y., Santiago, C. B. & Sigman, M. S. Developing Comprehensive Computational Parameter Sets to Describe the Performance of Pyridine-Oxazoline and Related Ligands. *ACS Catal.* **7**, 4144–4151 (2017).
35. Richter, N., Gröger, H. & Hummel, W. Asymmetric reduction of activated alkenes using an enoate reductase from *Gluconobacter oxydans*. *Appl. Microbiol. Biotechnol.* **89**, 79–89 (2011).
36. Kress, N., Rapp, J. & Hauer, B. Enantioselective Reduction of Citral Isomers in NCR Ene Reductase: Analysis of an Active-Site Mutant Library. *ChemBioChem* **18**, 717–720 (2017).
37. Ying, X. *et al.* Engineering the enantioselectivity of yeast old yellow enzyme Oye2Y in asymmetric reduction of (E/Z)-citral to (R)-citronellal. *Molecules* **24**, 1–15 (2019).
38. Meng, Q. *et al.* Computational Redesign of an ω -Transaminase from *Pseudomonas jessenii* for Asymmetric Synthesis of Enantiopure Bulky Amines. *ACS Catal.* **11**, 10733–10747 (2021).
39. Garcia-Borràs, M. *et al.* Origin and Control of Chemoselectivity in Cytochrome c Catalyzed Carbene Transfer into Si-H and N-H bonds. *J. Am. Chem. Soc.* **143**, 7114–7123 (2021).