IntEnzyDB: an Integrated Structure-Kinetics Enzymology Database

Bailu Yan,[1,5] Xinchun Ran,[1] Anvita Gollu,[1] Zihao Cheng,[1] Xiang Zhou,[1] Yiwen Chen,[4] and Zhongyue J. Yang[1-4,*]

[1]*Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235, United States* [2]*Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37235, United States* [3]*Vanderbilt Institute of Chemical Biology, Vanderbilt University, Nashville, Tennessee 37235, United States* [4]*Data Science Institute, Vanderbilt University, Nashville, Tennessee, 37235, United States* [5]*Department of Biostatistics, Vanderbilt University, Nashville, Tennessee, 37205, United States*

ABSTRACT: Data-driven modeling has emerged as a new paradigm for biocatalyst design and discovery. Biocatalytic databases that integrate enzyme structure and function data are in urgent need. Here, we described IntEnzyDB as an integrated structure-kinetics database for facile statistical modeling and machine learning. IntEnzyDB employs a relational architecture with flattened data structure, which allows rapid data operation. This architecture also makes it easy for IntEnzyDB to incorporate more types of enzyme function data. IntEnzyDB contains enzyme kinetics and structure data from six enzyme commission classes. Using 1019 enzyme structure-kinetics pairs, we investigated the efficiency-perturbing propensity for mutations that are close or distal to the active site. The statistical results show that efficiency-enhancing mutations are globally encoded; deleterious mutations are much more likely to occur in close mutations than in distal mutations. Finally, we described a web interface that allows public users to access enzymology data stored in IntEnzyDB. IntEnzyDB will provide a computational facility for data-driven modeling in biocatalysis and molecular evolution.

Keywords: Biocatalysis; Enzymology database; Mutations; Statistical analysis

1. Introduction

As a holy-grail challenge in modern chemical sciences, developing new enzyme catalysts provides solutions to transform chemically challenging reactions,[1] expand substrate scope,[2] control complex reaction selectivity,[3] treat metabolic disorders,[4] and degrade inert environmental wastes and pollutants[5]. Data-driven modeling methods have been extensively leveraged to innovate the approaches for enzyme catalyst discovery. They help elucidate the mechanisms of enzyme catalysis,[6] predict the impact of mutations on enzyme functions,[7, 8] and even design artificial enzymes[9].

Central to data-driven modeling, databases have been established for storing enzyme sequence, structure, and kinetics data (Table 1 and Supporting Information, Table S1). For example, Universal Protein Resource Knowledgebase (UniProtKB) contains ~36.7 million unique enzyme sequences.[10] RCSB Protein Databank (PDB) contains 108,000 experimentally-determined enzyme structures.[11] BRENDA[12] and SABIO-RK[13] store enzyme kinetic parameters, including: 80,000 k_{cat} values, 169,000 K_M values, 33,000 k_{cat}/K_M values from BRENDA; and over 56,000 K_Ms or pseudo-dissociation constants, and more than 52,000 velocity constants (V_{max} and k_{cat}) from SABIO-RK. These data cover thousands of enzyme commission (EC) classes that span over seven enzyme types (i.e., oxidoreductases, transferases, hydrolases, lyases, isomerases, ligases, and translocases). In addition, databases have been established to annotate enzyme functions based on its structural, chemical, and metabolic relevance (e.g., EzCatDB,[14] MACiE,[15] KEGG,[16] FunCat,[17] Reactcome,[18] and MetaCyc[19]), to map enzyme sequence, structure, and function relationship (e.g., PDBSWS,[20] SFLD,[21] FunTree,[22] IntEnz,[23] ExploreEnz,[24] and ExPASy[25]), to classify enzymes based structural and functional superfamilies (e.g., CATH[26] and SCOP[27, 28]), and to store designed enzymes (e.g., ProtaBank[29] and Design2Data[4]).

Table 1. A brief summary of enzymology databases.

Database Type	Databases	Data	UniProt	EC Number	PDB ID
Kinetics	BRENDA	Kinetics	Yes	Yes	Part
	Sabio-RK	Kinetics	Yes	Yes	No
	STRENDA DB	Kinetics with uniform data standard	Yes	Yes	No
Structure	PDB	PDB Structure	Yes	Yes	Yes
	AlphaFold DB	Predicted Structure	Yes	No	No
	UniProt	Sequence with Functional annotation	Yes	Yes	Yes
Kinetics and structure data for designed enzymes	ProtaBank	Kinetcs/Structure	Part	Part	Part
	Design2Data	Kinetcs/Structure	Yes	Yes	Yes

To develop holistic, predictive models for enzyme catalysis, an integrated database is needed that merges related enzyme sequence, structure, and function data in one place. However, three challenges are identified. First, collecting data of various sources is difficult because databases involve different design (e.g., relational, object-oriented, or hybrid), storage hierarchy, query mechanism, and API protocol. As such, curating enzyme features consumes significant efforts. Second, data cleaning is tricky due to various data standards adopted by different databases. Although unified data reporting standards have been reported (e.g., STRENDA[30] and EnzymeML[31]), existing enzyme data entries still involve missing or inaccurate mutational spot labels, experimental conditions, or other information. Additionally, manual typos and rounding errors are not uncommon, leading to obstacles for data validation. Third, data joining between enzyme structure and kinetics is challenging because they do not have consistently shared keys.

Enzyme kinetics databases store data entries by EC number and do not always have PDB ID for mapping with the structure database (Table 1). Although UniProt is used across databases, one-to-one mapping between structure and kinetics is difficult because one UniProt may correspond to tens of PDB IDs.

Here, we developed an integrated structure-kinetics enzymology database, IntEnzyDB, for facile data-driven modeling and machine learning. We have previously reported the beta-version of IntEnzyDB as a hydrolase database.[32] In this work, we expanded IntEnzyDB to incorporate data from six enzyme commission classes. IntEnzyDB allows fast operation of large amount of enzyme structure data and enables mapping between enzyme kinetics and structure. Using these data, we analyzed the propensity of catalytic efficiency enhancement, neutrality, and deletion for mutations that are close or distal to the active site. Finally, we introduced the web-interface for IntEnzyDB that allows public users to freely access and analyze the data.

2. Computational Methods

Database Construction. IntEnzyDB is a relational database with flattened data structure. IntEnzyDB adopts one data table to store all enzyme records of the same structural hierarchy (i.e., chain, residue, or atom) or property (i.e., kinetics). The current version of IntEnzyDB consists of five data tables, including: one table storing enzyme kinetic parameters such as Michaelis constant (K_M) and apparent turnover number (k_{cat}); three tables storing enzyme chain-level, amino acid-level, and atom-level structural information; and one table for one-to-one mapping of enzyme structure, substrate, and kinetics. Notably, the number of data tables can be easily expanded as we further develop IntEnzyDB to incorporate more enzyme properties (e.g., stability, mechanism, etc.).

Data Collection. The kinetics data in IntEnzyDB were extracted from BRENDA,[12] SABIO-RK,[13] ProtaBank[29] and Design2Data[4] databases; the structure data from RCSB Protein Databank (PDB)[11] and the sequence data from UniProt[10]. The enzyme kinetics table contains EC number, UniProtKB, organism, substrate, experimental temperature, mutational information. Using UniProt Retrieve/ID mapping tool and PDB Data API, we collected 8086 protein structures associated with the PDB IDs under UniProtKBs in the kinetics table.

The PDB structure data are stored in three tables. The enzyme chain table stores the general information of a PDB structure, including PDB ID, EC number, enzyme type, enzyme name, mutation, organism, chain ID, resolution, FASTA sequence, active site location, number of residues, and missing residues. The enzyme amino acid table stores the amino acid level structural information, including PDB ID, chain ID, amino acid name, amino acid index, and center-of-mass spatial coordinates of amino acid. The enzyme atom table stores the atom level structural information, including PDB ID, chain ID, atom name, atom index, amino acid name, amino acid index, atom coordinates. The database is open to the public and can be accessed through website interface (http://ec2-18-117-226-14.us-east-2.compute.amazonaws.com/). Any change will be posted on the website interface.

Data Curation. The kinetics data are curated based on the following criteria: (1) at least one wild-type kinetics parameter (k_{cat} and K_M) exists under one UniProtKB, (2) at least one PDB structure exists under one UniProtKB, (3) substrate information exists for each kinetic parameter, (4) experimental temperature is known for each kinetic parameter, (5) mutation is known for each kinetic parameter, and (6) mutations are single amino acid substitution. The curation yields 4037 k_{cat}/K_M values derived from 686 enzymes and 2540 enzyme mutants (i.e., single amino acid substitution) combined with 929 substrates. The experimental temperature of the kinetic

parameters ranges from 295.15 to 343.15 K (Supporting Information, Figure S1). These enzymes span over six types of enzyme commission (EC) classes, including: Oxidoreductases (EC 1), Transferases (EC 2), Hydrolases (EC 3), Lyases (EC 4), Isomerases (EC 5), and Ligases (EC 6).

To conduct one-to-one mapping of enzyme kinetics to structure, we adopted a three-step curation workflow. Step-1, we extracted PDB IDs from the research articles associated with the enzyme kinetics using text mining method (Supporting Information .zip file). Step-2, for the kinetic values where the PDB IDs are not available from the research paper, we manually identified the PDB structure by aligning the mutation spot annotation (taken from PDB file). Step-3, under each UniProtKB, we selected the PDB structures with active-site annotation, top resolution, and the least number of missing residues. This three-step approach allowed us to one-to-one map the kinetic data with the PDB structure through UniProtKB, yielding 159 PDB structures precisely paired with 1019 k_{cat}/K_M values. For the curated dataset, we evaluated the distribution of structure resolution and number of unresolved residues (Supporting Information, Figure S2). These data allow in-depth analysis of enzyme structure-function relationship. Data collection and curation are performed in Python and R software, and all statistical analysis are performed in R software. The curated kinetic and structural data tables, and data curation codes can be found in the Supporting Information zip file.

3. Results and Discussion

3a. Design Architecture and Data Processing Efficiency of IntEnzyDB. Unlike object-oriented databases that store each enzyme record in an individual data table (or file),[11] IntEnzyDB adopts a relational database architecture with a flattened data structure (detailed in the Computational Methods section). This allows IntEnzyDB to be expandable to incorporate other types of enzyme function data such as stability[33] and solubility[34]. The database employs five tables to store enzyme

kinetics and structure information (*top*, Figure 1), including three tables for cleaned enzyme structure data derived from RCSB PDB (i.e., ① chain, ② amino acid, and ③ atom table), one table for kinetics derived from BRENDA and Sabio-RK (labeled as ④), and one reference table (labeled as ⑤). The chain, amino acid, and atom tables share PDB ID and Chain ID as foreign keys. The chain table contains general protein structure information, including enzyme name, organism, gene, FASTA sequence, active site, and resolution; the amino acid table stores amino acid attributes, properties, and physiochemical parameters, including residue name, residue sequence number, amino acid weight, center of mass coordinate; the atom structure table stores the atom types and coordinates, including atom name, atom sequence number, residue name, residual sequence number, atomic weight, and atom Cartesian coordinates.

The kinetics table contains kinetic parameters, enzymology assay information, and sequence data, including UniProtKB, EC number, organism, substrate, mutation, experimental temperature, apparent turnover number (k_{cat}), Michaelis constant (K_M), enzyme efficiency (k_{cat}/K_M), and change of free energy barriers for a mutant compared to the wild-type enzyme ($\Delta\Delta G^{\ddagger}$, converted from k_{cat}/K_M according to eq 1). Kinetic table uses UniProtKB (sequence ID) as the foreign key. The reference table (table ⑤, Figure 1) contains one-to-one mapping relationship between kinetics and PDB based on foreign keys PDB ID, Chain ID, and UniProtKB (detailed in Computational Methods section). This table can be used to identify PDB structure for a given kinetic data of interest. The data from the table can also be used to investigate structure-kinetics relationship.

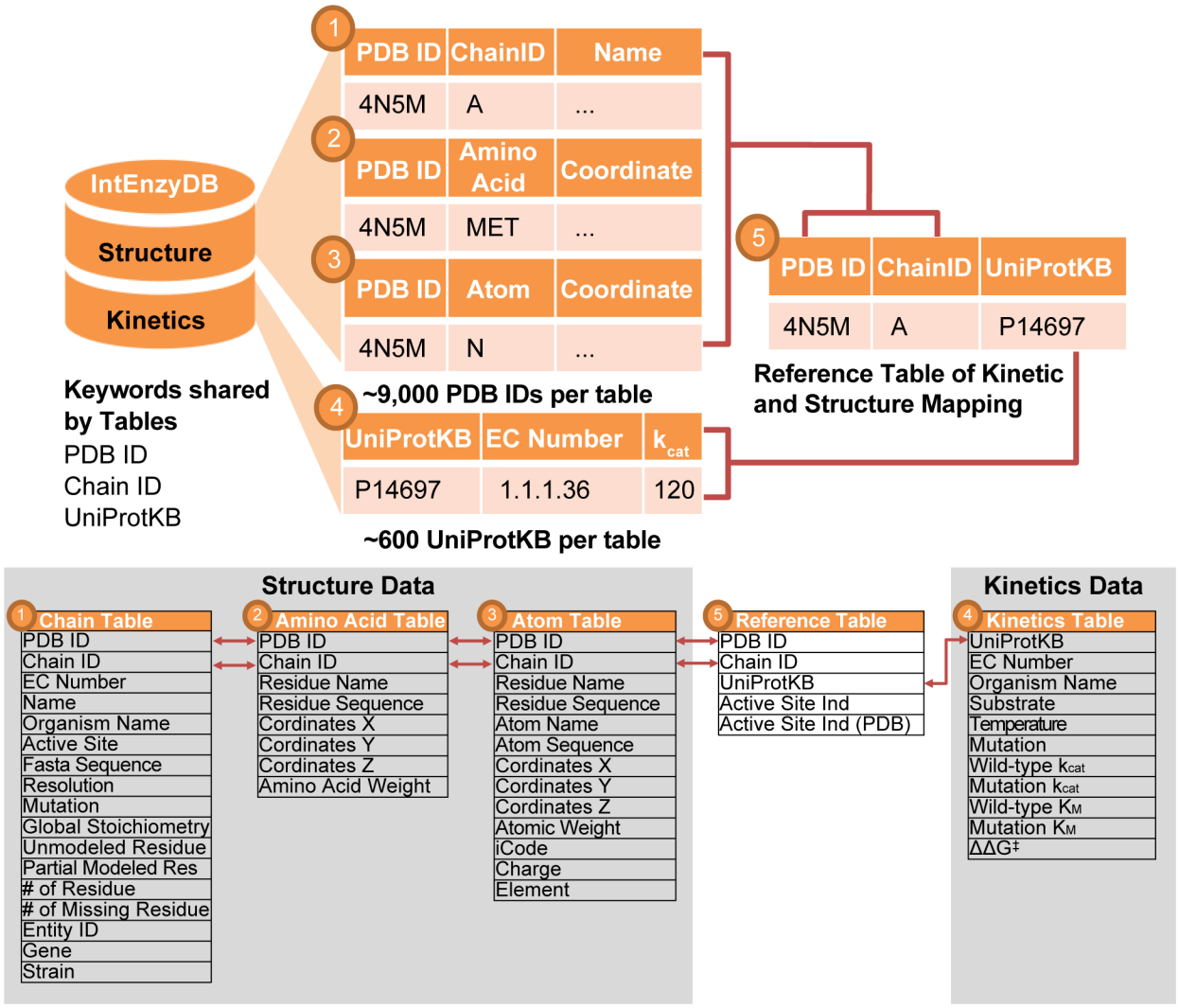

Figure 1. The architecture and relation map for IntEnzyDB. (Top) The database architecture involves five tables, including: three for enzyme structure tables (i.e., chain-level, amino acid-level, and atom-level), one for enzyme kinetics, and one reference table with foreign keys from structure and kinetics tables. The tables are mapped by the following keys: PDB ID, Chain ID, and UniProtKB. (Bottom) The mapping relationship between variables of different tables.

We benchmarked the time of pulling enzyme structure data using IntEnzyDB against a manual curation strategy (Figure 2). In contrast, using IntEnzyDB, a user can directly filter and download cleaned and tabulated structural data using SQL language; for the manual curation

strategy, a user needs to first download data from PDB, then read and reformat data by entry, and eventually combine them to one table on local computer. Figure 2 shows that IntEnzyDB is ~2 times faster than the traditional approach for 200 enzymes (80 s vs 173 s) and ~6 times faster for 1000 enzymes (151 s vs 905 s). The results indicate that the operating time by using IntEnzyDB is nearly independent of data size, which largely outperforms the manual operation strategy when operating on large amount of structural data (i.e., thousands or more).

The high data processing efficiency of IntEnzyDB likely results from its flattened data structure. Comparing to the traditional approach where data tables and files are accessed serially, IntEnzyDB loads all data entries at one time. This approach makes IntEnzyDB slower when processing smaller amount of data (e.g., for one enzyme structure, 86 s vs 1.9 s), but can save tremendous amount of time for repeatedly opening and reading files when handling large amount of structure data (e.g., 3.5 mins for 5000 structures). Therefore, IntEnzyDB provides an efficient solution for extracting enzyme structural features for statistical analysis or machine learning.

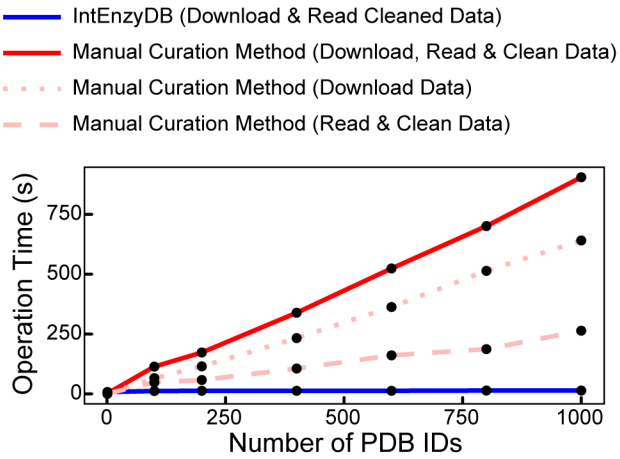

Figure 2. Operation time versus the number of PDB IDs by IntEnzyDB (blue line) and manual curation method (red line). The operation time of downloading, reading, and cleaning data in a tabulated form is measured for the tasks of processing 1, 100, 200, 400, 600, 800, and 1000 PDB structures. The data downloading and reading/cleaning are represented by the dotted and dashed

lines (in light red), respectively. The total operation time for manual curation method is shown by the red solid line. All operation times are measured in seconds.

3b. Statistical Analysis of Kinetic Parameters in IntEnzyDB. From IntEnzyDB, we curated 4037 k_{cat}/K_M values for enzymes with single amino acid substitution. The dataset consists of 686 wild-type enzymes, 2540 enzyme mutants, and 929 substrates (detailed in Computational Methods section). The number of k_{cat}/K_M values has tripled the size of hydrolase kinetics data we reported in the prior work (i.e., 1240).[32] Among the 4037 k_{cat}/K_M values, 29.0% are oxidoreductases (EC 1), 20.9% are transferases (EC 2), 30.7% are hydrolases (EC 3), 9.3% are ligases (EC 4), 5.2% are isomerases (EC 5), and 4.9% are lyases (EC 6) (*left*, Figure 3). To evaluate the impact of mutation on enzyme catalysis, we investigated the distribution of $\Delta\Delta G^{\ddagger}$ values derived from 2540 enzyme mutants, where $\Delta\Delta G^{\ddagger}$ is converted from the ratio of catalytic efficiency in the mutant to that in the wild-type enzyme (eq1):

$$\Delta\Delta G^{\ddagger} = -RT \ln \frac{k_{cat}^{mutant}/K_M^{mutant}}{k_{cat}^{wild-type}/K_M^{wild-type}} \qquad \text{eq1}$$

where R, T, k_{cat}, and K_M refer to the gas constant, experimental temperature, turnover number, and Michaelis constant, respectively. The distribution of $\Delta\Delta G^{\ddagger}$ follows a right-skewed Gaussian that ranges from -5.5 to 11.2 kcal/mol with a mean of 1.3 kcal/mol (*right*, Figure 3). The breadth of the distribution is wider than that of hydrolases (i.e., −4.2 to 9.4 kcal/mol), but the mean value is similar (i.e., 1.2 kcal/mol).[32] We categorized the mutants to be efficiency-enhancing ($\Delta\Delta G^{\ddagger} \leq$ -0.5 kcal/mol), -neutral ($\Delta\Delta G^{\ddagger} >$ -0.5 and $\leq$ 0.5kcal/mol, and -deleterious ($\Delta\Delta G^{\ddagger} >$ 0.5 kcal/mol). We observed 11.2% of the mutants to be efficiency-enhancing, 29.5% neutral, and 59.3% deleterious. As expected, the mutations that slow down catalytic rate are much more populated than those that are neutral or beneficial to catalysis. The efficiency-enhancing mutations appear to be more abundant in the database than their natural abundance.[35, 36] This phenomena might be caused by

the observational bias (e.g., researchers are more likely to report beneficial mutants) or the lack of deleterious mutations whose kinetic parameters are beyond the detection limit of biochemical assays.

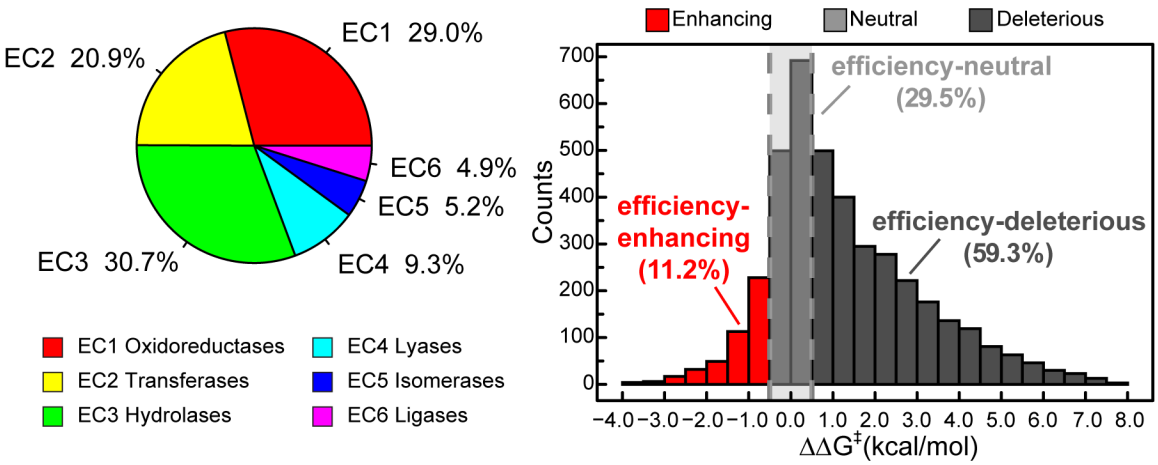

Figure 3. Statistics of kinetics data for enzymes mutants with single amino acid substitution in IntEnzyDB. (Left) The distribution of kinetics data for six EC classes. (Right) The distribution of $\Delta\Delta G^{\ddagger}$ values for 2540 enzyme variants-catalyzed reactions with a bin size of 0.5 kcal/mol. Efficiency-enhancing mutant is defined as $\Delta\Delta G^{\ddagger}$ less or equal to -0.5 kcal/mol (red), efficiency-neutral mutant is defined as $\Delta\Delta G^{\ddagger}$ greater than -0.5 kcal/mol and less or equal to 0.5 kcal/mol (light grey), efficiency-deleterious mutant is defined as $\Delta\Delta G^{\ddagger}$ greater than 0.5 kcal/mol (dark grey).

3b. Mutation Effects for Close versus Remote Mutations. After joining enzyme kinetics with structure data using the reference table (Figure 1), we obtained 1019 one-to-one mapped enzyme structure-kinetics pairs, including 376 oxidoreductases, 95 transferases, 313 hydrolases, 119 ligases, 76 isomerases, and 40 lyases. Noticeably, the data entries for hydrolase (313) is less than the amount of data (i.e., 403) curated in our prior work [32]. This is because in this work, we applied a stricter filtration condition that traces every kinetic entry to the corresponding structure in literature using text mining (detailed in the Computational Method section) rather than simply relied on UniprotKB to map kinetic entry with the best-resolved structure as done previously. In

addition, there are 3018 k_{cat}/K_M values from the kinetics table whose corresponding enzyme structures (either wild-type or mutant) or the active-site annotation is not known. To address this, we will obtain the missing structures using enzyme structure prediction tools (e.g., AlphaFold2[37] and RoseTTAFold[38]); we will also curate the active site annotation from M-CSA database[39] or label them manually.

Using 1019 structure-kinetics pairs, we investigated the difference in the efficiency-perturbing propensity for mutations that are spatially close versus distal to the active site residues (Figure 4). This analysis has been conducted for hydrolases in our prior work.[32] In contrast, the current dataset involves a greater number of enzymes with a wider converge of enzyme types. As such, the statistical study can potentially inform a more holistic trend for the spatial dependence of efficiency-perturbing mutations. The distance between a mutation spot and active site was measured between the mutation residue's Cα coordinate and the geometric center of the active-site residues' Cα coordinates. Using 15 Å as an empirical cutoff, the efficiency-enhancing propensity of the close mutations (8.0%) is found to resemble that of the distal mutations (8.6%). However, the efficiency-deleterious mutations are much more populated for the close (72.0%) than the distal mutations (46.0%). As a compensation, the efficiency-neutral mutations are about 26% more observed for distal mutations.

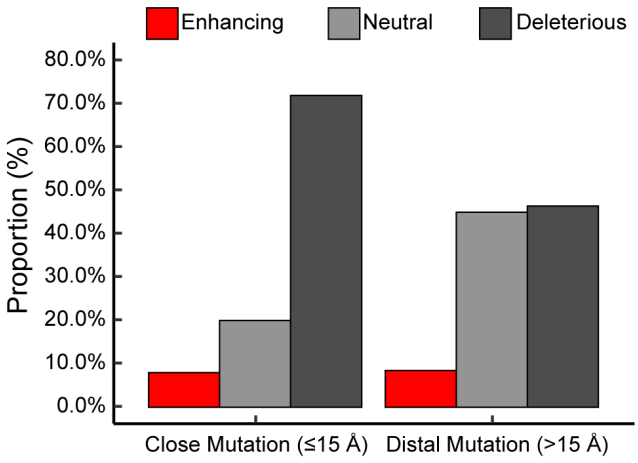

Figure 4. The proportion of efficiency-enhancing (red), -neutral (light grey), and -deleterious (dark grey) mutations for close mutation (≤15 Å) and distal mutation (>15 Å). The distance is defined based on the distance of the mutation residue Cα coordinate to the geometric center of the active-site residues Cα coordinates. Efficiency-enhancing mutant is defined as $\Delta\Delta G^{\ddagger}$ less or equal to -0.5 kcal/mol (red), efficiency-neutral mutant is defined as $\Delta\Delta G^{\ddagger}$ greater than -0.5 kcal/mol and less or equal to 0.5 kcal/mol (light grey), efficiency-deleterious mutant is defined as $\Delta\Delta G^{\ddagger}$ greater than 0.5 kcal/mol (dark grey).

The efficiency-perturbing propensity may be dependent on the choice of the spatial cutoff values. To reduce arbitrariness, we evaluated the proportions for the close versus distal mutations using different spatial cutoffs sampled from 10 to 20 Å with 1 Å interval (Supporting Information, Table S2 and Figure S3). The cutoff values below 10 Å were not tested because of the scarcity of mutations falling into the category of close mutation (especially beneficial mutations). The efficiency-enhancing propensity is estimated to be 7.3 ± 1.5% for the close mutations and 8.6 ± 1.1% for the distal mutations – they remain highly similar. Despite the fluctuation, the propensity of rate deletion is still much higher for the close mutations (72.9 ± 5.7%) than for the distal mutations (45.5 ± 4.8%). This trend remains to be compensated by the efficiency neutral mutations (19.8 ± 4.2% for close mutations and 45.9 ± 5.7% for remote mutations). Notably, the same trend still exists when separately analyzing the data for the three major enzyme classes: oxidoreductases, transferases, and hydrolases (Supporting Information, Figure S4-S6).

The statistical studies show that close mutations are equally probable in inducing efficiency enhancement as distal mutations, indicating that efficiency-enhancing mutations are globally distributed. This result is consistent with the observation that the Whitehead group reported for *E. coli*-expressed amidases;[36] and supports prior statistical study by the Kazlauskas group that both

close and distal mutations can improve activity (based on 55 rate-enhancing enzyme variants) [40]. For enzyme engineering, given the smaller number of residues in the active site than the distal spots, strategies that emphasize the mutagenesis of active site residues are likely to be more statistically productive, such as the combinatorial active site saturation test (i.e., CASTing[41]). In addition, our statistical results show that distal mutations are much less likely to induce efficiency deletion than close mutations. This illustrates the important roles of distal mutations in avoiding rate-deletion and induce neutral drift on the fitness landscape, explaining the broadly reported observation of distal mutation in beneficial mutants during directed evolution.[42]

3d. IntEnzyDB web interface. To allow IntEnzyDB accessible by public users, we developed a web interface that has a backend link to IntEnzyDB on MongoDB (Figure 5). The web interface allows users to dynamically connect to MongoDB and generate data tables based on search queries. The dynamic connection scheme also makes it easy for users to obtain the most updated data as we continue expanding the database. The website contains general information about the database architecture and scope under the "Home" and "Research" page. Under the "Database" page, a user can find kinetics data (i.e., "Kinetics Data" tab), structural data (i.e., "Structure Data" tab), and mapped structure-kinetics data (i.e., "Kinetics-Structure Reference" tab). Under the "Kinetics Data" tab, the user can find 4037 curated kinetics data for enzymes with single amino acid substitution where both k_{cat} and K_M are available. The data table contains variables including EC number (e.g., 3.1.1.2), UniProtKB (e.g., P27169), organism (e.g., *Homo sapiens*), substrate (e.g., phenylacetate), mutation (e.g., H115W), experimental temperature (e.g., 298.15K), and change of free energy barrier $\Delta\Delta G^{\ddagger}$ (e.g., 1.7 kcal/mol, converted from eq 1). Under the "Structure Data" tab, the user can find general structural information, including the PDB ID (e.g., 1V04), enzyme name (e.g., arylesterase), active site index (e.g., 115), and resolution (e.g., 2.2 Å). On the "Kinetics-Structure

Reference" tab, the mapped kinetics-structure pairs are shown. For each entry in this reference table, the UniProtKB matches an entry in the kinetics table and PDB ID in the structure table. Under this tab, a user can click on the UniProtKB or PDB ID hyperlinks to directly access UniProt and PDB website with more detailed structure and functional information.

Besides the data tables, the user can access the "Search" tab and find specific enzyme data entries in the data tables using UniProtKB, PDB ID, or EC number as search queries. The user can also visualize the statistical analysis of enzyme kinetics data under the "Statistics", including the number of enzymes in each EC class, the distribution of $\Delta\Delta G^{\ddagger}$, and the frequency of mutations in IntEnzyDB. On the "Database Access" tab, the user can find instructions to directly access IntEnzyDB on MongoDB. This way, the user can access to the full database with 5 tables shown in Figure 1 and query enzymes of interest.

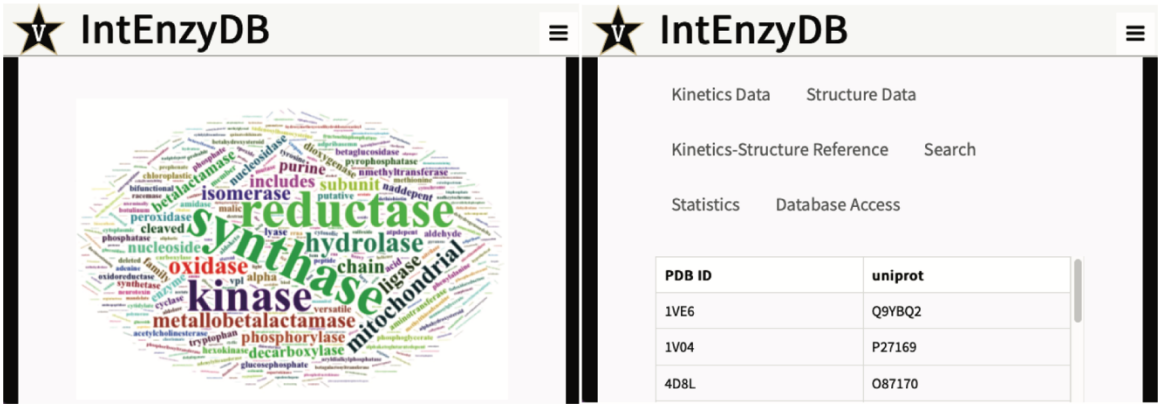

Figure 5. Screenshots of IntEnzyDB web interface. (Left) The homepage for the IntEnzyDB website. (Right) The database tabs for the website.

4. Conclusion

Here we reported IntEnzyDB as an integrated structure-kinetics enzymology database. IntEnzyDB adopts a relational architecture with flattened data structure. The database consists of

five data tables, including one kinetics table, three structure tables, and one structure-kinetics reference table. In the benchmark for processing 1000 protein structures, IntEnzyDB is six times faster than manual curation approach that relies on direct downloading from the PDB website and accessing from local directory. The high efficiency of IntEnzyDB is due to its flattened data structure: with all structure/kinetics data entries read into computer memory in the form of giant data tables, the time for repetitive file input/output operations can be saved.

From IntEnzyDB, we curated 4037 data entries where both k_{cat} and K_M are known for enzyme mutants with single amino acid substitution. These data are primarily derived from three enzyme commission classes, including: oxidoreductases (29.0%), transferases (20.9%), and hydrolases (30.7%). Ligases, isomerases, and lyases are observed to occupy 9.3%, 5.2%, and 4.9% of the population, respectively. Through analyzing mutation effects, we observed 11.2% of the mutants to be efficiency-enhancing, 29.5% neutral, and 59.3% deleterious.

Using 1019 enzyme structure-kinetics pairs, we investigated the spatial dependence of efficiency-perturbing propensity for mutations. Specifically, we categorized mutations to either close or distal to active site residues using various spatial cutoff values ranging between 10 and 20 Å with 1 Å interval; under each cutoff value, we tested the proportion for efficiency-enhancing, -neutral, and -deleterious mutations for both "close" and "distal" mutations. The efficiency-enhancing propensity is estimated to be 7.3 ± 1.5% for the close mutations and 8.6 ± 1.1% for the distal mutations – they remain highly similar. Despite the fluctuation, the propensity of rate deletion is consistently higher for the close mutations (72.9 ± 5.7%) than for the distal mutations (45.5 ± 4.8%). This trend is compensated by the efficiency neutral mutations (19.8 ± 4.2% for close mutations and 45.9 ± 5.7% for remote mutations).

Finally, we described the web interface for IntEnzyDB, which employs a backend link to MongoDB. The web interface allows public users to dynamically access and query data based on their need. Besides the kinetics, structure, and reference data tables, the web interface also contains instructions for users to directly access data tables on IntEnzyDB.

As the next steps for developing IntEnzyDB, we will further expand the mapped structure-kinetics data table by using predicted structures and active site annotation. Text mining strategies will be implemented to enable more comprehensive data validation and expansion. We will incorporate more types of enzymology data to IntEnzyDB, including stability, solubility, expressibility, and even molecular modeling data derived from high-throughput simulations[43].

ASSOCIATED CONTENT

Supporting Information. Summary of enzymology database; distribution of operating temperature for enzymatic reactions, the number of the missing residue, and resolution; data table and distributions for the proportion of rate-perturbing mutations under various spatial cutoff. (PDF) Combined R and Python code for data collection and cleaning; Python code for text mining workflow; table for mapped enzyme structure-kinetics data. (ZIP)

Data Availability. The data can be accessed from the web interface for IntEnzyDB: http://ec2-18-117-226-14.us-east-2.compute.amazonaws.com.

AUTHOR INFORMATION

Corresponding Author

*email: zhongyue.yang@vanderbilt.edu phone: 615-343-9849

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This research was supported by the startup grant from Vanderbilt University and the scholarship from the Vanderbilt Institue of Chemical Biology. This work was carried out in part using computational resources from the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number TG-BIO200057.[44]

References

1. Yang, Y.; Arnold, F. H., Navigating the Unnatural Reaction Space: Directed Evolution of Heme Proteins for Selective Carbene and Nitrene Transfer. *Accounts of Chemical Research* **2021,** *54* (5), 1209-1225.
2. Tang, Q.; Grathwol, C. W.; Aslan-Üzel, A. S.; Wu, S.; Link, A.; Pavlidis, I. V.; Badenhorst, C. P. S.; Bornscheuer, U. T., Directed evolution of a halide methyltransferase enables biocatalytic synthesis of diverse SAM analogs. *Angewandte Chemie International Edition* **2021,** *60* (3), 1524-1527.
3. Reetz, M. T., Laboratory evolution of stereoselective enzymes: a prolific source of catalysts for asymmetric reactions. *Angewandte Chemie International Edition* **2011,** *50* (1), 138-174.
4. Huang, X.; Kim, D.; Huang, P.; Vater, A.; Siegel, J. B., Design to Data for mutants of β-glucosidase B from Paenibacillus polymyxa: L171M, H178M, M221L, E406W, N160E, F415M. *bioRxiv* **2020**, 2020.11.17.387829.
5. Yoshida, S.; Hiraga, K.; Takehana, T.; Taniguchi, I.; Yamaji, H.; Maeda, Y.; Toyohara, K.; Miyamoto, K.; Kimura, Y.; Oda, K., A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science* **2016,** *351* (6278), 1196.
6. Bonk, B. M.; Weis, J. W.; Tidor, B., Machine Learning Identifies Chemical Characteristics That Promote Enzyme Catalysis. *J Am Chem Soc* **2019,** *141* (9), 4108-4118.
7. Li, F.; Yuan, L.; Lu, H.; Li, G.; Chen, Y.; Engqvist, M. K. M.; Kerkhoven, E. J.; Nielsen, J., Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis* **2022**.
8. Heckmann, D.; Lloyd, C. J.; Mih, N.; Ha, Y.; Zielinski, D. C.; Haiman, Z. B.; Desouki, A. A.; Lercher, M. J.; Palsson, B. O., Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nature Communications* **2018,** *9* (1), 5252.
9. Wu, Z.; Johnston, K. E.; Arnold, F. H.; Yang, K. K., Protein sequence design with deep generative models. *Current Opinion in Chemical Biology* **2021,** *65*, 18-27.
10. The UniProt Consortium, UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **2017,** *45* (D1), D158-D169.
11. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Research* **2000,** *28* (1), 235-242.

12. Chang, A.; Jeske, L.; Ulbrich, S.; Hofmann, J.; Koblitz, J.; Schomburg, I.; Neumann-Schaal, M.; Jahn, D.; Schomburg, D., BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Research* **2021,** *49* (D1), D498-D508.
13. Wittig, U.; Kania, R.; Golebiewski, M.; Rey, M.; Shi, L.; Jong, L.; Algaa, E.; Weidemann, A.; Sauer-Danzwith, H.; Mir, S.; Krebs, O.; Bittkowski, M.; Wetsch, E.; Rojas, I.; Müller, W., SABIO-RK—database for biochemical reaction kinetics. *Nucleic Acids Research* **2012,** *40* (D1), D790-D796.
14. Nagano, N., EzCatDB: the Enzyme Catalytic-mechanism Database. *Nucleic Acids Research* **2005,** *33* (suppl_1), D407-D412.
15. Holliday, G. L.; Almonacid, D. E.; Bartlett, G. J.; O'Boyle, N. M.; Torrance, J. W.; Murray-Rust, P.; Mitchell, J. B. O.; Thornton, J. M., MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. *Nucleic Acids Research* **2007,** *35* (suppl_1), D515-D520.
16. Kanehisa, M.; Goto, S., KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **2000,** *28* (1), 27-30.
17. Ruepp, A.; Zollner, A.; Maier, D.; Albermann, K.; Hani, J.; Mokrejs, M.; Tetko, I.; Güldener, U.; Mannhaupt, G.; Münsterkötter, M.; Mewes, H. W., The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research* **2004,** *32* (18), 5539-5545.
18. Gillespie, M.; Jassal, B.; Stephan, R.; Milacic, M.; Rothfels, K.; Senff-Ribeiro, A.; Griss, J.; Sevilla, C.; Matthews, L.; Gong, C.; Deng, C.; Varusai, T.; Ragueneau, E.; Haider, Y.; May, B.; Shamovsky, V.; Weiser, J.; Brunson, T.; Sanati, N.; Beckman, L.; Shao, X.; Fabregat, A.; Sidiropoulos, K.; Murillo, J.; Viteri, G.; Cook, J.; Shorser, S.; Bader, G.; Demir, E.; Sander, C.; Haw, R.; Wu, G.; Stein, L.; Hermjakob, H.; D'Eustachio, P., The reactome pathway knowledgebase 2022. *Nucleic Acids Research* **2022,** *50* (D1), D687-D692.
19. Caspi, R.; Altman, T.; Billington, R.; Dreher, K.; Foerster, H.; Fulcher, C. A.; Holland, T. A.; Keseler, I. M.; Kothari, A.; Kubo, A.; Krummenacker, M.; Latendresse, M.; Mueller, L. A.; Ong, Q.; Paley, S.; Subhraveti, P.; Weaver, D. S.; Weerasinghe, D.; Zhang, P.; Karp, P. D., The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research* **2014,** *42* (D1), D459-D471.
20. Martin, A. C. R., Mapping PDB chains to UniProtKB entries. *Bioinformatics* **2005,** *21* (23), 4297-4301.
21. Akiva, E.; Brown, S.; Almonacid, D. E.; Barber, A. E., 2nd; Custer, A. F.; Hicks, M. A.; Huang, C. C.; Lauck, F.; Mashiyama, S. T.; Meng, E. C.; Mischel, D.; Morris, J. H.; Ojha, S.; Schnoes, A. M.; Stryke, D.; Yunes, J. M.; Ferrin, T. E.; Holliday, G. L.; Babbitt, P. C., The Structure–Function Linkage Database. *Nucleic Acids Research* **2014,** *42* (D1), D521-D530.
22. Furnham, N.; Sillitoe, I.; Holliday, G. L.; Cuff, A. L.; Rahman, S. A.; Laskowski, R. A.; Orengo, C. A.; Thornton, J. M., FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies. *Nucleic Acids Research* **2012,** *40* (D1), D776-D782.
23. Fleischmann, A.; Darsow, M.; Degtyarenko, K.; Fleischmann, W.; Boyce, S.; Axelsen, K. B.; Bairoch, A.; Schomburg, D.; Tipton, K. F.; Apweiler, R., IntEnz, the integrated relational enzyme database. *Nucleic Acids Research* **2004,** *32* (suppl_1), D434-D437.
24. McDonald, A. G.; Boyce, S.; Tipton, K. F., ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Research* **2009,** *37* (suppl_1), D593-D597.

25. Duvaud, S.; Gabella, C.; Lisacek, F.; Stockinger, H.; Ioannidis, V.; Durinx, C., Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Research* **2021,** *49* (W1), W216-W227.
26. Knudsen, M.; Wiuf, C., The CATH database. *Human Genomics* **2010,** *4* (3), 207.
27. Andreeva, A.; Kulesha, E.; Gough, J.; Murzin, A. G., The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research* **2020,** *48* (D1), D376-D382.
28. Andreeva, A.; Howorth, D.; Chothia, C.; Kulesha, E.; Murzin, A. G., SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Research* **2014,** *42* (D1), D310-D314.
29. Wang, C. Y.; Chang, P. M.; Ary, M. L.; Allen, B. D.; Chica, R. A.; Mayo, S. L.; Olafson, B. D., ProtaBank: A repository for protein design and engineering data. *Protein Science* **2018,** *27* (6), 1113-1124.
30. Swainston, N.; Baici, A.; Bakker, B. M.; Cornish-Bowden, A.; Fitzpatrick, P. F.; Halling, P.; Leyh, T. S.; O'Donovan, C.; Raushel, F. M.; Reschel, U.; Rohwer, J. M.; Schnell, S.; Schomburg, D.; Tipton, K. F.; Tsai, M.-D.; Westerhoff, H. V.; Wittig, U.; Wohlgemuth, R.; Kettner, C., STRENDA DB: enabling the validation and sharing of enzyme kinetics data. *The FEBS Journal* **2018,** *285* (12), 2193-2204.
31. Range, J.; Halupczok, C.; Lohmann, J.; Swainston, N.; Kettner, C.; Bergmann, F. T.; Weidemann, A.; Wittig, U.; Schnell, S.; Pleiss, J., EnzymeML—a data exchange format for biocatalysis and enzymology. *The FEBS Journal* **2021,** *n/a* (n/a).
32. Yan, B.; Ran, X.; Jiang, Y.; Torrence, S. K.; Yuan, L.; Shao, Q.; Yang, Z. J., Rate-Perturbing Single Amino Acid Mutation for Hydrolases: A Statistical Profiling. *The Journal of Physical Chemistry B* **2021,** *125* (38), 10682-10691.
33. Stourac, J.; Dubrava, J.; Musil, M.; Horackova, J.; Damborsky, J.; Mazurenko, S.; Bednar, D., FireProtDB: database of manually curated protein stability data. *Nucleic Acids Research* **2021,** *49* (D1), D319-D324.
34. Niwa, T.; Kanamori, T.; Ueda, T.; Taguchi, H., Global analysis of chaperone effects using a reconstituted cell-free translation system. *Proceedings of the National Academy of Sciences* **2012,** *109* (23), 8937-8942.
35. Romero, P. A.; Arnold, F. H., Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology* **2009,** *10* (12), 866-876.
36. Wrenbeck, E. E.; Azouz, L. R.; Whitehead, T. A., Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nature Communications* **2017,** *8* (1), 15695.
37. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D., Highly accurate protein structure prediction with AlphaFold. *Nature* **2021,** *596* (7873), 583-589.
38. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee Gyu, R.; Wang, J.; Cong, Q.; Kinch Lisa, N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman Caleb, R.; DeGiovanni, A.; Pereira Jose, H.; Rodrigues Andria, V.; van Dijk Alberdina, A.; Ebrecht Ana, C.; Opperman Diederik, J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy Manoj, K.; Dalwadi, U.; Yip Calvin, K.; Burke John, E.; Garcia, K. C.; Grishin

Nick, V.; Adams Paul, D.; Read Randy, J.; Baker, D., Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021,** *373* (6557), 871-876.
39. Ribeiro, António J M.; Holliday, G. L.; Furnham, N.; Tyzack, J. D.; Ferris, K.; Thornton, J. M., Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Research* **2018,** *46* (D1), D618-D623.
40. Morley, K.; Kazlauskas, R., Improving enzyme properties: when are closer mutations better? *Trends in biotechnology* **2005,** *23*, 231-7.
41. Clouthier, C. M.; Kayser, M. M.; Reetz, M. T., Designing New Baeyer−Villiger Monooxygenases Using Restricted CASTing. *The Journal of Organic Chemistry* **2006,** *71* (22), 8431-8437.
42. Wilding, M.; Hong, N.; Spence, M.; Buckle, A. M.; Jackson, C. J., Protein engineering: the potential of remote mutations. *Biochemical Society Transactions* **2019,** *47* (2), 701-711.
43. Shao, Q.; Jiang, Y.; Yang, Z. J., EnzyHTP: A High-Throughput Computational Platform for Enzyme Modeling. *Journal of Chemical Information and Modeling* **2022,** *62* (3), 647-655.
44. Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D.; Roskies, R.; Scott, J. R.; Wilkins-Diehr, N., XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering* **2014,** *16* (5), 62-74.

TOC Abstract: