Robust Prediction of Enzyme Variant Kinetics with RealKcat

Karuna Anna Sajeevan[1,2]*, Abraham Osinuga[3]*, Arunraj B[1], Sakib Ferdous[1], Nabia Shahreen[3], Mohammed Sakib Noor[1], Shashank Koneru[1], Laura Mariana Santos-Correa[1], Rahil Salehi[1], Niaz Bahar Chowdhury[3], Brisa Calderon-Lopez[1], Ankur Mali[4], Rajib Saha[3,†], and Ratul Chowdhury[1,2,†]

[1]Department of Chemical and Biological Engineering, Iowa State University, Ames, Iowa, USA
[2]The Center for Biorenewable Chemicals, Iowa State University, Ames, Iowa, USA
[3]*Department of Chemical and Biomolecular Engineering, University of Nebraska-Lincoln, Lincoln, Nebraska, USA*
[4]Department of Computer Science and Engineering, University of South Florida, Tampa, Florida, USA

[†]Email all correspondences to: rsaha2@unl.edu or ratul@iastate.edu
*Equal contribution

Author contributions:

Conceptualization: R.C., R.S., A.M., A.O., K.A.S.
Methodology: R.C., A.O., K.A.S., A.M., S.F.
Investigation: K.A.S., A.O., A.M., A.B., S.F., N.S., M.S.N., S.K., L.M.S., R.S., N.B.C., B.C.
Visualization: A.O., A.M., M.S.N., S.F., R.C.
Funding acquisition: R.C., R.S.
Project administration: R.C.
Supervision: R.C., R.S.
Writing – original draft: A.O., K.A.S, A.M., A.B.
Writing – review & editing: K.A.S., A.O., A.M., A.B., R.S., R.C.

Competing Interest Statement: Authors declare that they have no competing interests.

Classification: Biochemistry; Systems Biology; Biophysics and Computational Biology.

Keywords: Enzyme kinetics; Bio-aware Machine learning; Enzyme engineering; Biocatalysis; Database curation

This PDF file includes:

Main Text

Figures 1 to 5

Tables 1 to 2

Abstract

Accurate prediction of kinetic parameters is crucial for understanding known and tailoring novel enzymes for biocatalysis. Current models fail to capture mutation effects on catalytically essential residues, limiting their utility in enzyme design. We grid-searched through ten model architectures (25,671 hyperparameter combinations) to identify a gradient-based additive framework called RealKcat, trained on 27,176 experimental entries curated manually (KinHub-27k) by screening 2,158 articles. Clustering catalytic turnover (k_{cat}) and substrate affinity (K_M) by rational orders of magnitude, RealKcat achieves >85% test accuracy, demonstrating highest sensitivity to mutation-induced variability thus far, and is the first-of-its-kind-model to demonstrate complete loss of activity upon deletion of the catalytic apparatus. Finally, state-of-the-art k_{cat} validation accuracy (96%) on alkaline phosphatase (PafA) mutant industrial dataset confirms RealKcat's generalizability in learning per-residue catalytic relevance.

Significance Statement

Enzymes are proteins that facilitate biochemical reactions. Measuring enzyme efficiency (catalytic rates, k_{cat}) is important yet challenging and time-consuming. Predicting how short/ long range changes to catalytic site affects enzyme activity will bolster biotechnology and pharmaceutical innovations. However, existing models (a) operate at modest accuracy, and (b) fail to capture loss of activity even upon alteration of the catalytic apparatus. We introduce RealKcat machine-learning platform with a rigorously curated KinHub-27k dataset by manually screening 2,158 articles, and 17k synthetic datapoints to demonstrate state-of-the-art accuracy. This breakthrough fills important gaps in database oversight, bio-aware machine learning, and computational enzyme design.

Main Text

Introduction

Precise characterization of enzyme kinetics is foundational for synthetic biology, systems biology, and disease biomarker discovery, where tailored enzymes drive innovations in biomanufacturing, and therapeutics (*1, 2*). Predicting catalytic activity enables rapid characterization of metabolic landscape of non-model species, engineering of enzymes for specific substrates or optimized biochemical pathways, reducing the need for costly experimental trials. This is particularly valuable for rapid pathway optimization—whether for sustainable bioproduction of metabolites/ protein products and discerning disease-linked metabolic shifts—where a trial-and-error approaches are prohibitive and do not unfold mechanistic bases. Sequence-linked functionally variant enzymes are key for industrial biocatalysis, and disease prognosis markers (*3, 4*), highlighting the need for robust, mutation-sensitive predictive models. However, current computational methods often lack the accuracy required, particularly for enzymes modified at catalytically crucial residues. Moreover, enzyme kinetics assessed in controlled *in vitro* settings often fall short of capturing the dynamic biochemical environments of cells, leading to predictions that may diverge from *in vivo* behavior. This gap limits our understanding of metabolic shifts, drug responses, and pathway designs in non-model organisms and human systems. With the rising demand for precise enzyme design and characterization in synthetic biology, systems biology, and biomedicine, there is a pressing need for machine learning (ML) models that offer robust, mutation-sensitive predictions beyond the constraints of traditional datasets and experimental limitations.

Advancements in ML have introduced tools for predicting enzyme kinetics, including models such as DLKcat, TurNuP, UniKP, CatPred, and EITLEM-Kinetics, each demonstrating the potential of data-driven approaches to capture enzyme-substrate interaction kinetics (*5–9*). For instance, DLKcat utilizes convolutional neural networks (CNNs) and graph neural networks (GNNs) to predict enzyme turnover number (k_{cat}) across diverse enzyme-substrate pairs, though its performance depends heavily on dataset diversity. Building on this foundation, TurNuP employs ESM-1b encodings for enzyme features (*6, 10*) and RDKit-derived reaction fingerprints, implemented through a gradient-boosted tree algorithm to improve generalizability for enzymes with limited dataset. UniKP further advances the field by incorporating a two-layer model that

encodes enzyme sequences and substrate structures, with additional environmental variables like pH and temperature (*7*). However, its accuracy remains constrained by the quality and diversity of training data. A more recent model, CatPred, directly predicts k_{cat}, Michaelis constant (K_M), and inhibition constant (K_i) by employing advanced neural networks trained on a comprehensive dataset derived from SABIO-RK and BRENDA (*11, 12*), which are widely regarded as key resources in enzyme kinetics modeling. The model achieves state-of-the-art accuracy, with 79.4% of k_{cat} predictions and 87.6% of K_M predictions falling within one order of magnitude error of the experimental values (*8*). Additionally, EITLEM-Kinetics introduces an ensemble iterative transfer learning approach, enabling accurate kinetic parameter predictions for mutants with low sequence similarity (*13*). While these tools have advanced the field, a significant gap remains in their sensitivity to mutations in catalytic sites. For example, alanine mutations at catalytic residues yield near-identical catalytic rates across all these models, underscoring a reliance on sequence similarity without true catalytic awareness. Moreover, CatPred's method of concatenating Simplified Molecular Input Line Entry System (SMILES) strings for substrates and cofactors, focusing on reaction fingerprints, further compounds this issue by overlooking distinct substrate and cofactor effects. Hence, enzyme engineering requires a high sensitivity to mutations, along with well-curated, robust datasets and optimal machine learning frameworks to guarantee accuracy and reliability.

To address these limitations, we present RealKcat, which employs optimized gradient-boosted decision trees to deliver robust, mutation-sensitive predictions of enzyme kinetics. RealKcat is built on the carefully curated KinHub-27k dataset, containing 27,176 entries from BRENDA and SABIO-RK, and further enriched with negative data from UniProt and InterPro active site annotations (*14, 15*). This dataset addresses about 1800 data inconsistencies through a thorough, article-by-article review of the original sources. By integrating ESM-2 sequence embeddings to capture evolutionary context (*16*), and ChemBERTa embeddings for substrate representation (*17*), RealKcat constructs a rich feature set that significantly enhances predictive accuracy. Unlike previous models, RealKcat frames enzyme kinetics' prediction as a classification problem, clustering k_{cat} and K_M values by orders of magnitude with dedicated clusters for extreme values, a strategy that captures functional relevance across diverse enzyme classes. This is particularly important because industrial-scale enzyme engineering processes largely concern about a range of order of magnitude (say, $10^5 < k_{cat} < 10^7$) of a specific enzyme

variant rather than the exact numerical value. In addition, predicting a range of kinetic values enables their usage in constructing feasible metabolic models (as shown before (*18*)) allowing for differences *in vitro* to *in vivo* kinetic rates, to capture experimental phenotypes with fidelity (*19–21*). Utilizing techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) to balance class representation (*22*), RealKcat achieves high accuracy, exceeding 89% test performance for k_{cat} and 85% on K_M, and demonstrates tolerance to mutation-driven variability. Validation on the alkaline phosphatase (PafA) dataset from Markin et al. (*23*), containing 1,016 single-site mutants, confirmed RealKcat's ability to detect catalytic relevance, achieving accuracies of 53% for k_{cat} and 93% for K_M as well as e-accuracies within one order of magnitude error of 96% for k_{cat} and 100% for K_M. These results establish RealKcat as a highly robust tool for enzyme design, optimization, synthetic biology and metabolic engineering, significantly advancing the predictive accuracy of enzyme kinetics.

Results

Dataset Curation and Composition

To establish a high-quality foundation for RealKcat, we curated a dataset of 27,176 experimentally verified enzyme kinetics entries, meticulously sourced from SABIO-RK, BRENDA, and UniProt (see Fig. 1A). Each entry underwent a rigorous corroboration process, wherein we cross-checked 2,158 source articles to resolve 1804 inconsistencies in catalytic parameters ($k_{cat,}$ K_M), enzyme sequences, and substrate identity (see Table 2). The entire KinHub-27k dataset and accompanying code are available for download and inference ***upon publication*** (https://chowdhurylab.github.io/downloads.html). This intensive manual curation, addressing discrepancies that could impair the dataset's reliability, enhances prediction accuracy for machine learning applications, making RealKcat the first enzyme kinetic predictor trained on a rigorously curated dataset. The curated dataset contains 3,364 unique SMILES, 11,350 sequences, and spans 1,508 unique organisms (Fig. 1E). Sequences longer than 1024 characters were removed, and duplicate entries were consolidated by grouping identical isomeric SMILES from PubChem, retaining maximum k_{cat} and minimum K_M, resulting in a refined dataset of 23,121 unique entries with 14,637 wild-type entries and 8,484 unique sequences (Fig. 1D). Figure 1B shows the distribution of enzyme commission (EC) numbers, while Fig. 1C visualizes high-similarity

enzyme sequences by EC class, using Hamming distance on fixed-length encodings to depict structural relationships across classifications.

To further enhance RealKcat's ability to identify catalytically relevant patterns, we, for the first time, generated a negative dataset by mutating catalytic residues in the curated database (KinHub-27k) sequences (annotated as active sites in UniProt) to alanine (*Ala*). These variant sequences were assigned a new "*Cluster 0*" for (k_{cat} = 0), simulating inactive variants (i.e., variants with missing catalytic apparatus) (see Fig. 1G). The K_M values for these variants were retained, refer to Fig. 1H. Initially, this approach yielded approximately 5,000 negative data points, which we expanded to ~17,000 entries by integrating data from InterPro annotations. Customized scripts reconciled discrepancies in catalytic residue positions between PDB and UniProt by developing sequence motifs based on relative spacing, enabling precise verification and extension of annotations. This comprehensive dataset, comprising wild type (WT), mutant (MD), and synthetically inactive sequences, provides a balanced and diverse training set, enhancing RealKcat's sensitivity to catalytically relevant mutations and its capacity to distinguish active states.

Feature Embedding Strategy and Clustering for Multi-Class Enzyme Kinetics Prediction

The RealKcat framework, depicted in Fig. 2., leverages optimized gradient-based additive modeling via XGBoost architecture (Fig. 2B) to achieve high accuracy and sensitivity in predicting enzyme kinetics. XGBoost's capacity to capture complex, non-linear dependencies in structured biological data are essential for multi-class classification tasks in enzyme kinetics (*24, 25*). Its gradient-boosting structure allows RealKcat to handle diverse data distributions efficiently, enabling robust predictions of catalytic turnover (k_{cat}) and substrate affinity (K_M) across a wide range of enzyme-substrate interactions. To maximize predictive reliability, we curated a comprehensive feature set by combining two advanced embeddings: ESM-2 for enzyme sequences and ChemBERTa for substrates (Fig. 2A). ESM-2 embeddings capture evolutionary and structural nuances directly from protein sequences, offering insights into functional dependencies that emerge from sequence variations (*10*). ChemBERTa, on the other hand, encodes molecular characteristics of substrates, including functional groups and stereochemistry – captured via isomeric SMILEs, thus enhancing RealKcat's capacity to recognize critical enzyme-substrate interactions (*17*).

Unlike previous models that concatenate substrates and products into combined SMILES strings or overlook detailed catalytic site information, RealKcat generates substrate embeddings from each compound's isomeric SMILES individually as most *in vitro* kinetic data often holds true at single-molecule level (*26*). This method also preserves distinct substrate and cofactor effects, enhancing the model's sensitivity to molecular variations. Both ESM-2 and ChemBERTa embeddings were concatenated into a unified 2,048-dimensional vector (1,280 from ESM-2 and 768 from ChemBERTa, refer to Fig. 2A), creating an enriched representation that integrates evolutionary and molecular perspectives. To handle the inherent variability in enzyme kinetics data, we implemented an order-of-magnitude clustering strategy, categorizing k_{cat} and K_M values into bins based on magnitude intervals (Fig. 1F), with additional clusters for extreme values (See Materials and Methods for details). This clustering approach captures the exponential variability of kinetic parameters, which is characteristic of biological systems, enabling RealKcat to generalize across different enzyme classes and catalytic efficiencies (*13, 27*).

Moreover, we conducted a comprehensive exploration of state-of-the-art models, covering both traditional machine learning algorithms—such as linear models, random forests, decision trees, gradient boosting, and XGBoost—and a variety of deep learning architectures. Our deep learning experiments included multi-layer perceptrons, convolutional layers for feature extraction paired with fully connected layers, recurrent neural networks like LSTMs and GRUs (both with and without bidirectional connections), and two-layer transformer decoders. We also evaluated ensemble methods, which performed comparably to our top model, XGBoost. However, XGBoost consistently outperformed all other models while being significantly more computationally efficient. This extensive analysis involved testing over 25,000 model variations, tuning hyperparameters across a broad range—including learning rates, number of layers, hidden units, number of decision trees, and optimizers. As shown in Fig. 3A, model performance on our dataset varied significantly, with deep learning models exhibiting lower accuracy on test splits and random performance on mutated datasets.

Predictive Accuracy and Sensitivity in Enzyme Kinetics Modeling

RealKcat's predictive framework, strengthened by SMOTE-based data balancing and order-of-magnitude clustering, achieves exceptional accuracy and sensitivity in enzyme kinetics modeling. On the test set, RealKcat attained accuracies of 88.72% for k_{cat} and 85.08% for K_M,

with e-accuracies within one order of magnitude reaching 94.72% and 97.16%, respectively (Fig. 3B-C). High recall and F1-scores—88.72% and 88.64% for k_{cat}, and 85.08% and 84.96% for K_M (Fig. 3D-E)—demonstrate its balanced precision and recall. Confusion matrices further reveal RealKcat's capability to resolve kinetic ranges, especially within central clusters (Fig 1F, 3G-H), while maintaining low misclassification at extremes (Fig. 3G-H). Additionally, t-SNE visualizations confirm robust clustering of true and predicted classes, underscoring RealKcat's ability to capture high-dimensional parameter spaces accurately (Fig. 3F&I).

In comparison to existing enzyme kinetics prediction models, RealKcat exhibits superior predictive performance and sensitivity across catalytic parameters k_{cat} and K_M. While models such as DLKcat report a Pearson correlation coefficient (PCC) of 0.71, RealKcat achieves significantly higher accuracy on both training and test datasets, with e-accuracies reaching over 94.72% for k_{cat} and 97.16% for K_M, demonstrating enhanced precision (Fig. 3B-C). Additionally, CatPred reports R^2 values of 0.61 and 0.65 for k_{cat} and K_M predictions, respectively, with modest prediction within one order of magnitude for only 79.4% of k_{cat} predictions. EITLEM-Kinetics, which employs an iterative transfer learning strategy, shows promising improvements in R^2 values, reaching up to 0.727 for k_{cat} and 0.681 for K_M; however, RealKcat's approach of SMOTE-augmented data balancing and clustering offers a robust alternative by better handling data imbalance and variability. TurNuP, with a reported R^2 of 0.44, further underscores the limitations of existing models in capturing the nuanced effects of mutation and catalytic variability. RealKcat's robust performance, when compared with other existing models – refer to table S1, and across a diverse dataset of wild-type and mutant enzyme entries underscores its capacity to accurately predict kinetic parameters with high sensitivity, effectively robust in capturing both natural and engineered sequence variations.

Mutation-Aware Kinetic Predictions Across Catalytically Relevant Residues

RealKcat demonstrated a high degree of sensitivity to mutations affecting catalytically relevant residues, accurately predicting changes in kinetic parameters associated with point mutations in the enzyme's active site. To comprehensively validate RealKcat's mutation-aware capabilities, we tested its performance emphasizing its sensitivity to structural modifications at catalytic sites on a six enzymes dataset, across glycolytic, TCA, and pentose phosphate pathways, in *Saccharomyces cerevisiae* and *Escherichia coli*. Since this dataset is different from what the model was trained on and includes enzyme entries with known catalytic residues from

UniProt, we tested how well RealKcat predicts k_{cat} and K_M values in response to specific point mutations at these residues. As depicted in Fig. 4A, RealKcat's performance in predicting wild-type k_{cat} values were compared with the other available models such as DLKcat, CatPred, UniKP, and EITLEM-Kinetics. The plot demonstrates RealKcat's high predictive accuracy, as reflected by the close alignment of predicted k_{cat} range values with experimentally measured rates across a wide dynamic range. RealKcat's predictions remain within the same order of magnitude of the experimental values for most data points, surpassing the accuracy observed in other comparative models (as mentioned ealier). This result underscores RealKcat's capability to generalize across diverse enzymes with different catalytic efficiencies, indicating that the model effectively captures the underlying biochemical kinetics. Specific enzyme-substrate pairs, such as P00942 (dihydroxyacetone 3-phosphate), are highlighted to showcase RealKcat's ability to accurately model a few cases where other models diverge from the experimental data. Additional examples include P61889 (oxaloacetate) and P0A6L0 (2-deoxy-D-ribose), where RealKcat demonstrates consistent predictive accuracy, refer to Fig. 4 and Table 1.

Figure 4B shifts the focus to RealKcat's sensitivity to mutations, showing predicted k_{cat} values for mutant enzymes where catalytic residues, like H95 and E165 in *Saccharomyces cerevisiae* glyceraldehyde 3-phosphate dehydrogenase (P00942, refer to Table 1, were mutated to alanine. This evaluation is critical for understanding RealKcat's ability to infer catalytic relevance in response to structural changes. Notably, RealKcat accurately reflects the expected decrease in catalytic activity upon mutation of critical residues, evidenced by the shift of data points away from the parity line for several enzyme mutants, such as P0A6L0 and P42222. The performance comparison with models like DLKcat, CatPred, UniKP, and EITLEM-Kinetics reveals that RealKcat more accurately captures the impact of catalytic residue mutation on k_{cat}, thereby demonstrating a superior mutation-aware prediction capability. This mutation sensitivity is essential for applications in enzyme engineering and synthetic biology, where accurate modeling of mutation effects can accelerate the design of enzymes with tailored functions.

In assessing RealKcat's performance on wild-type K_M predictions and its sensitivity to catalytic mutations, our results demonstrate significant advances over existing models (Fig. 4). As shown in Fig. 4C, RealKcat accurately predicted wild-type K_M values across a broad range of substrate affinities, aligning closely with experimental data along the parity line. These results

were also benchmarked to CatPred, UniKP, and EITLEM-Kinetics, Fig. 4. This alignment is particularly evident for substrates associated with enzyme P00942, where RealKcat's predictions more faithfully capture substrate binding affinities, K_M, emphasizing its robustness in handling diverse biochemical contexts. Fig. 4D shows RealKcat's response to alanine mutations at catalytic residues being appropriately minimal, thus indicating its nuanced understanding of K_M dependency on substrate geometry rather than catalytic residues.

To further validate RealKcat's ability to predict the impact of mutations on enzyme kinetics, we employed the high-throughput dataset of the alkaline phosphatase (PafA) from *Elizabethkingia meningoseptica*, sourced from Markin et al. (*23*). This dataset, which explores the functional consequences of single-point mutations, introduces glycine and valine at each residue position to systematically probe catalytic and structural roles. For PafA, catalytic residues include R164, with key mutations R164A and R164G in the test set to specifically assess RealKcat's mutation-sensitivity. The PafA dataset contains kinetic measurements (k_{cat} and K_M) for 1,016 single-point mutations with the substrate carboxy 4-methylumbelliferyl phosphate ester (cMUP), a binding-limited substrate. For RealKcat's model training, we partitioned the data into training, validation, and test sets containing 554, 310, and 310 data points, respectively. Importantly, the wild-type (WT) and a catalytic mutation (T79S) were included in the training partition, while the key R164A and R164G mutations were reserved for the test set to evaluate RealKcat's inferencing power on unobserved, catalytically significant variants. This partitioning ensured that training spanned the high, median, and low ranges of observed k_{cat} values, creating a robust foundation for mutation sensitivity testing.

In the results shown in Fig. 4F&G, RealKcat demonstrates a test accuracy of 53% for k_{cat} and an exceptional e-accuracy of 96% within ±1 order of magnitude for k_{cat} predictions on the PafA dataset. For K_M, the model showed even higher e-accuracy of 93% and 100%. RealKcat's predictions closely aligned with the experimental data, particularly for catalytic mutations. The R164A mutation was predicted within an order of magnitude lower than the *in vivo* value, while R164G, which introduces a flexible glycine side chain, was predicted within an order of magnitude higher. These shifts, especially for R164G, underscore RealKcat's sensitivity to changes in side-chain properties and indicate potential areas for refinement with further training on glycine mutations at catalytic sites. Notably, the structural analysis (see fig. S1) shows that

R164 is highly interconnected, with 13 neighboring residues, indicating that mutations at this site might compromise structural stability or lead to other catalytic behavior. Nonetheless, the high accuracy within a ±1 order of magnitude underscores RealKcat's potential to inform enzyme design by predicting functional impacts of mutations across diverse kinetic landscapes.

Discussion

RealKcat provides a mutation-sensitive, catalytic-aware tool for predicting enzyme kinetics by combining thorough data curation, pretrained language models specialized in biochemical and molecular representation, and strong classification modeling. This approach addresses key limitations of previous models and pushes forward the field of computational enzyme design. The rigorous data curation and clustering approach in RealKcat enhances predictive accuracy and reliability, establishing a robust foundation for computational modeling of enzyme kinetics. While k_{cat} and K_M are continuous parameters typically approached as regression problems, practical challenges in computational frameworks necessitate a classification-based treatment for pragmatic high-fidelity predictions. Unlike regression-based models, such as DLKcat, CatPred, which vary in performance across enzyme domains, RealKcat groups kinetic parameters into functionally meaningful clusters based on affinity and catalytic turnover. For instance, according to criteria from Bar-Even et al. (*27*), K_M values can be categorized into high (>1 mM), moderate (0.1–1 mM), and low (≤0.1 mM) clusters, distinguishing enzymes that maintain metabolic flux across dynamic conditions from those requiring high substrate affinity. Similarly, k_{cat} values span high (>1000 s^{-1}), moderate (100–1000 s^{-1}), and low (≤100 s^{-1}) categories, addressing the varying catalytic efficiencies observed across enzyme families. This classification-based approach empowers RealKcat to capture biologically meaningful distinctions, offering a more nuanced alternative to traditional regression frameworks. In addition, RealKcat's implementation of order-of-magnitude clustering captures physiologically relevant distinctions in enzyme behavior, aligning closely with *in vivo* data and reducing discrepancies, including physiological, that often arise in traditional *in vitro* analyses, which fail to reflect the complexity of cellular environments (*28*). Moreover, several *in vivo* data often encounter noise from the dynamic complexity of cellular processes, complicating the interpretation of enzyme activity and regulation (*29*).

By framing enzyme kinetics as a classification task, RealKcat's approach overcomes the limitations of regression-based models, capturing a broader range of kinetic variability while maintaining catalytic awareness. When compared to existing enzyme kinetics prediction models, RealKcat offers substantial improvements in sensitivity and accuracy, especially in mutation detection. Models such as DLKcat, TurNuP, UniKP, CatPred, and EITLEM-Kinetics have demonstrated utility in enzyme-substrate interaction modeling, yet they often fall short in distinguishing activity changes resulting from specific residue alterations, a crucial capability for enzyme engineering. Employing pretrained language models tailored for biochemical and molecular representation learning, ESM-2 and ChemBERTa embeddings, RealKcat integrates evolutionary information and molecular characteristics to generate nuanced predictions that faithfully represent enzyme behavior across a wide range of biological contexts and capturing subtle sequence-residue-function relationships. With its high accuracy and sensitivity to mutation-driven changes in catalytic turnover (k_{cat}) and substrate affinity (K_M), the model achieves over 88% accuracy for k_{cat} and 85% for K_M, on test datasets. Additionally, its e-accuracies exceed 94% for k_{cat} and 97% for K_M within one order of magnitude, surpassing the performance of existing models. Notably, RealKcat accurately reflected decreased catalytic activity resulting from mutations at critical residues, while maintaining precise K_M predictions when substrate affinity was unaffected by these mutations. Further validation using a high-throughput dataset of over 1,000 single-point mutations in alkaline phosphatase (PafA) confirmed RealKcat's mutation-aware predictive capabilities, achieving high e-accuracy for both kinetic parameters. These results highlight RealKcat's robustness in modeling mutation-induced kinetic variability and its potential as a powerful tool for enzyme design and optimization. Furthermore, RealKcat accurately maintains wild-type K_M values even in the presence of non-catalytic mutations. This highlights its advanced capability to model substrate affinity independently of changes to the catalytic site, a key feature for guiding precise enzyme modifications in biotechnology applications.

The implications of RealKcat extend beyond accurate prediction; the model's capabilities present an opportunity to accelerate enzyme design and optimization in synthetic biology and biomedicine. For metabolic engineering, RealKcat's precise kinetics predictions facilitate the development of tailored enzymes for specific pathways, reducing the need for iterative laboratory screening. Additionally, the model's mutation-sensitive framework is particularly

suited for disease modeling, where understanding the functional impact of sequence variations is crucial for elucidating disease mechanisms and identifying therapeutic targets. The high predictive power of RealKcat also makes it highly suitable for virtual enzyme screening in biotechnological applications, offering an efficient, reliable method for exploring potential enzyme-substrate interactions before costly *in vitro* experiments.

Furthermore, RealKcat's predictive capabilities hold valuable applications for systems biologists, particularly in kinetic modeling, enzyme-constrained modeling, and proteome-constrained metabolic models, where reliable kinetic parameters like k_{cat} and K_M are essential. Systems biology frameworks, such as those used for enzyme-constrained or proteome-constrained modeling, rely on high-confidence kinetic data to accurately simulate metabolic flux and to understand cellular responses under varying conditions (*30–32*). RealKcat enhances parameterization for kinetic and constraint-based models by providing precise predictions of enzyme turnover rates and substrate affinities, reducing dependence on stochastic or surrogate models that often require initial uncertainty distributions, which can introduce variability (*33, 34*). Traditional kinetic surrogate ML/DL models struggle with managing these distributions, leading to compromised prediction reliability and limited utility in precision-demanding simulations (*33–35*). RealKcat's classification-based framework offers a robust solution for capturing enzyme kinetics within order-of-magnitude clusters, providing reliable predictions that significantly reduce the uncertainties typically introduced by stochastic modeling approaches.

RealKcat's mutation-sensitive framework also enhances its utility in tracking sequence variations that drive functional innovations but can also lead to compromised enzyme function and contribute to disease pathology. These predictive capabilities are essential for biomedical research, where understanding the specific effects of mutations on enzyme function can inform studies on metabolic disorders, drug responses, and potential therapeutic targets (*36*). Sequence variations can either augment functional capabilities or, conversely, compromise enzyme activity in ways that contribute to disease, highlighting the need for mutation-sensitive models like RealKcat that can discern these changes with high fidelity (*37*). By capturing these subtle sequence-function relationships, RealKcat provides insights that are crucial for both biotechnological and biomedical applications, enabling the rational design of enzymes for

synthetic biology and offering a predictive platform for identifying critical mutations that may lead to pathological states in human health.

While RealKcat sets a new standard in enzyme kinetics prediction, the study is not without limitations. The dataset, though large and meticulously curated, is biased toward certain enzyme types, and underrepresented classes, such as translocases (refer to Fig. 1), may not achieve the same predictive accuracy. This bias is, however, dictated by the heterogeneity in experimental data that has been reported thus far. Additionally, the reliance on Isomeric SMILES for molecular representation, though effective, may fall short in capturing complex 3D structural nuances, which could affect predictions for stereochemically-rich molecules. Predicting kinetics for enzymes with limited representation remains a challenge, and data curation choices, while aimed at maximizing fidelity, may introduce subtle biases. Future work could address these limitations by expanding the dataset to include a wider range of enzyme classes and exploring advanced transfer learning techniques to enhance model adaptability for less-represented enzymes. These efforts would likely enhance RealKcat's robustness, broadening its applicability across diverse enzyme functions and unlocking new possibilities in biotechnological innovation, drug discovery, and metabolic engineering.

Materials and Methods

Database Building

In this study, enzyme kinetic parameters, specifically k_{cat} and K_M, were systematically collected and curated from BRENDA [release 2023_1] (*11*) and SABIO-RK [as of May 2024] (*12*), using custom Python scripts [steps in Fig. 5]. Initially, 237,487 entries from BRENDA and 77,663 from SABIO-RK were aggregated, parsed for annotations (UniProt ID, organism name, EC number, substrate name, mutant site information and kinetic values), and filtered to retain entries with the maximum k_{cat} and minimum K_M values. The maximum k_{cat} represents the highest observed catalytic turnover rate, which is critical for identifying the peak catalytic efficiency of the enzyme. Conversely, the minimum K_M reflects the strongest substrate affinity, indicating the lowest substrate concentration required for effective catalysis. Hence, this resulted in 96,782 entries from BRENDA and 54,906 from SABIO-RK. Substrate isomeric SMILES keys were obtained from PubChem database (*38*) using IUPAC name of substrates, ensuring consistency across databases to eliminate redundancies. After extensive data cleaning, entries

from both databases were merged, incorporating sequence data from UniProt (*14*) and BRENDA. Sequences with UniProt IDs were directly fetched, while those without UniProtIDs were obtained by querying BRENDA and UniProt using EC number and organism name, followed by applying mutation annotations to adjust sequences accordingly. This process resulted in a database containing 30,442 k_{cat} and 44,615 K_M entries, and a merged dataset of 26,244 entries with both k_{cat} and K_M values, including 10,244 mutant entries spanning 2,158 scientific articles with unique PubMed IDs (PMIDs).

Recognizing that a significant proportion of these entries might have discrepancies compared to the original references due to data entry errors (*27*), a manual curation process was conducted. Each article associated with the respective PMIDs for the mutant entries was meticulously reviewed. This rigorous curation process resulted in 11,175 collected mutant entries with k_{cat} and K_M values, resolving inconsistencies in ~ 1800 entries and adding ~ 1,072 new entries (Table 2). For example, the updated k_{cat} and K_M values for human m-NAD-malate dehydrogenase (D102E) in the conversion of fumarate to pyruvate are 60.61 s□¹ and 0.00091 M respectively. These values correct the earlier erroneous values of 44.14 s□¹ and 0.00255 M. On the other hand, the corrected k_{cat} and K_M values for catalysing 2-ketocaproate by *Lactobacillus pentosus* D-lactate dehydrogenase (Y52A) are 102 s□¹ and 0.0007 M, respectively, correcting the previous incorrect values of 0.7 s□¹ and 0.102 M. We identified and corrected inaccuracies in reported K_M (788 instances), k_{cat} (618 instances), mutants (18 instances), and substrates (240 instances). This curated mutant database, with 16,001 wildtype (WT) and 11,175 mutant (MD) entries with both k_{cat} and K_M now forms the curated KinHub-27k dataset that serves as input for preprocessing which then goes into training ML models to predict kinetic parameters.

To capture sequence similarity among enzyme classes, we constructed a network graph. Protein sequences were standardized to a fixed length of 1024 characters through truncation or padding and encoded by their character ordinal values. A fixed length of 1024 characters was used, as ESM-2 does not accept sequences longer than this length. Sequence similarity was quantified using Hamming distance, producing a similarity matrix. We applied a threshold of 0.8, creating edges between sequences with high similarity. Each sequence node was colored according to its primary EC class and visualized in a force-directed layout, as shown in Fig 1C. This method effectively highlights inter- and intra-class sequence relationships among diverse enzyme classes.

Data Preprocessing and Feature Engineering

In the RealKcat framework, enzyme kinetic parameters k_{cat} and K_M are treated as multi-class classification tasks, necessitated by practical constraints in computational frameworks when dealing with continuous values. The continuous nature of enzyme kinetics, traditionally a regression problem, is pragmatically addressed through a classification approach that leverages clustering of kinetic parameters by order of magnitude. This binning process categorizes k_{cat} values into ranges, each representing an order of magnitude difference, while reserving specific clusters for extreme values (refer to Figure 1F, G&H). This binning approach involved defining specific cluster edges: k_{cat} values were categorized with boundaries set at $[0,10^{-8},10^{-2},10^{-1},10^{0},10^{1},10^{2},10^{3},10^{8}]$ s^{-1} and K_M values at $[10^{-10},10^{-5},10^{-4},10^{-3},10^{-2},10^{-1},10^{4}]$ M. The classification approach provides a structured prediction strategy, allowing the model to interpret parameter ranges with high fidelity and minimizing the impact of noise from imprecise continuous predictions.

To complement this clustering approach, we incorporated a carefully constructed negative dataset to bolster RealKcat's robustness in distinguishing catalytically relevant modifications. Utilizing annotations from UniProt and InterPro, we systematically generated negative samples by altering active site residues in wild-type entries. Specifically, we focused on entries annotated with active site residues, using UniProt and InterPro data to guide the identification of catalytically important residues. In each wild-type sequence, we substituted these residues with alanine (*Ala*), creating mutations likely to ablate catalytic activity. These alterations targeted catalytic, substrate, and cofactor-binding residues, where applicable, to ensure that RealKcat could learn the nuanced impacts of such changes. Importantly, this substitution to alanine allowed for a consistent baseline across diverse enzymes, given alanine's minimal side-chain interference and widespread use in functional assays to assess catalytic relevance.

For enzymes annotated with multiple catalytic residues, we performed single-point mutations to alanine on each residue individually, maximizing the diversity of functional disruptions within our negative set. By this approach, we generated approximately 5,000 primary negative data points. To expand this set and cover additional enzyme classes and structural contexts, we leveraged InterPro annotations, developing custom scripts to align sequence motifs indicative of catalytic functions across enzyme homologs. This alignment facilitated the

identification of catalytic motifs, which we used to introduce systematic alanine substitutions in homologous regions. Furthermore, we resolved annotation discrepancies and alignment mismatches between UniProt and PDB using a motif-based approach, aligning catalytic residues within each enzyme's native sequence and positional spacing. This comprehensive curation process yielded a final set of approximately 17,000 high-confidence negative samples.

The negative samples generated by catalytic residue alanine substitutions were assigned to the lowest kinetic parameter "cluster 0" for the k_{cat} parameters, effectively labeling them as catalytically inactive. This cluster was intentionally defined to capture the absence or severe reduction in catalytic turnover due to disrupted active sites. For K_M, since the focus of these mutations was to disable catalytic function rather than alter substrate binding affinity, these entries were allowed to retain the K_M values for the corresponding WT entries. This expanded negative dataset serves to improve RealKcat's robustness by training the model to recognize non-catalytic modifications, enhancing predictive accuracy in both wild-type and mutant contexts.

To address class imbalance in our dataset, particularly in underrepresented clusters, we applied the Synthetic Minority Over-sampling Technique (SMOTE), following the method outlined by Chawla et al. (*22*). SMOTE generates synthetic instances by interpolating between a minority class instance and its k-nearest neighbors, effectively enhancing the representation of underrepresented kinetic parameter clusters without altering class labels. We employed the default parameter k=5, which adequately balanced the dataset, ensuring even distribution across kinetic bins. For consistency and reproducibility in applying SMOTE, we set a random seed of 42. This ensures that the synthetic samples generated during oversampling remain consistent across model runs. Applying SMOTE post-clustering allowed for targeted correction within each order-of-magnitude bin, which was essential given the naturally skewed distribution of enzyme kinetics data. This strategy provided a robust dataset foundation for RealKcat, improving model reliability in predicting diverse kinetic profiles, including mutation-sensitive responses often underrepresented in biological datasets.

The final dataset was partitioned into training, validation, and test subsets, adhering to an 80/10/10 split. This allocation provides a substantial training set while maintaining ample data for validation and unbiased testing. The distribution across these subsets was balanced to reflect the diversity of kinetic parameter clusters, thus allowing the model to generalize effectively

across the spectrum of k_{cat} and K_M classes. Each subset maintained proportional representation of classes, including both wild-type and mutant data, to ensure that the model could capture nuances across catalytic modifications as well as substrate affinities.

In RealKcat's feature extraction pipeline, each sample, from the established dataset, comprises an enzyme sequence, isomeric SMILES representation of the substrate, and assigned clusters for k_{cat} and K_M values. We employed ESM-2 embeddings, for evolutionary insights, where each enzyme sequence is encoded into a 1,280-dimensional vector by averaging over residues, yielding a consistent 1×1280 representation. ESM-2 is a transformer-based language model, encodes enzyme sequences by leveraging extensive protein databases to capture evolutionary motifs and structural dependencies (*10*). This method captures conserved functional and structural dependencies, irrespective of sequence length. Substrate features were extracted using ChemBERTa, which processes isomeric SMILES strings to produce a 768-dimensional vector. ChemBERTa encodes isomeric SMILES strings representing substrates, extracting information about functional groups, stereochemistry, and other molecular attributes crucial for enzyme-substrate interactions (*17*). The combined embeddings from ESM-2 and ChemBERTa were concatenated to form a unified 2,048-dimensional feature vector per sample. This enriched representation of evolutionary and molecular perspectives constitutes RealKcat's input, providing the model with comprehensive structural and chemical contexts necessary for accurate kinetic parameter classification across a wide array of enzyme-substrate pairs.

Model Architecture and Training

The architecture of the RealKcat model centers on a highly optimized gradient-boosting framework, specifically utilizing the Extreme Gradient Boosting (XGBoost) algorithm. XGBoost is a robust machine learning model, particularly well-suited for handling high-dimensional, non-linear data and yielding high accuracy in multi-class classification contexts (*24*), which aligns with the complexities of enzyme kinetics prediction. In RealKcat, XGBoost constructs an ensemble of decision trees, where each tree $\boldsymbol{f_k(X, \theta_k)}$ is sequentially added to the model, aiming to minimize the residual errors left by the previous trees. This process optimizes a compound objective function:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{K} \Omega(f_{(k)})$$

where $l(y_i, \hat{y}_i)$ denotes the log-loss for multi-class classification, specifically capturing the errors between true labels y_i and predictions $\hat{y}_i$. The second term, $\Omega(f_{(k)}) = (\gamma \times T) + (0.5 \times \lambda \times ||w||^2)$, is a regularization penalty that controls model complexity by penalizing each tree's structural complexity, with T representing the number of leaves in a tree and w the leaf weights. This regularization ensures that the model balances the trade-offs between bias and variance, which is crucial in kinetic prediction tasks where enzyme-substrate interactions are diverse. **γ (gamma)** parameter controls the structural complexity of each tree by penalizing the addition of new leaves. Specifically, γ determines the minimum gain a node must produce to be further split. Higher values of γ make the model more conservative, favoring trees with fewer leaves by only allowing splits that significantly improve the model. This prevents the formation of overly detailed trees, which could lead to overfitting. However, increasing **T** can lead to overfitting, as it allows the tree to become more tailored to the training set. The term **γ × T** thus penalizes trees with large numbers of leaves, encouraging simpler structures that are less likely to overfit. Defining other components:

λ (lambda): This is the **L2 regularization parameter** on leaf weights, and it helps to control the size of the weights associated with each leaf. A higher value of **λ** increases the penalty on large weights, which discourages the model from relying too heavily on any single leaf's prediction. By controlling the magnitude of leaf weights, **λ** helps improve the model's robustness, making it less sensitive to variations in the data.

$||w||^2$: This term represents the **squared L2 norm of the leaf weights**, where **w** is a vector containing the prediction values at each leaf node in the tree. The squared norm $||w||^2$ calculates the sum of the squares of these weights, which quantifies the overall "strength" of the predictions at the leaves. By penalizing larger norms, the regularization term helps to distribute the prediction values more evenly across leaves, thus reducing the likelihood of overfitting.

$0.5 \times \lambda \times ||w||^2$: This part of the regularization term penalizes large leaf weights while controlling the model's complexity by discouraging the creation of overly powerful leaves. The

factor of 0.5 is used to scale this penalty, allowing XGBoost to balance the model's accuracy with its simplicity more effectively.

Together, these components of $\Omega(f_{(k)})$ provide nuanced control over the model's complexity. By adjusting γ and λ, the user can fine-tune how aggressively the model penalizes complexity in terms of both the number of leaves and the magnitude of leaf predictions, ultimately promoting a balance between model accuracy and generalizability.

XGBoost further improves predictive accuracy by utilizing gradient-based $\boldsymbol{pseudo-residuals}$, computed as the gradient of the loss function with respect to current predictions, thereby emphasizing instances with high residual error from previous rounds. This refinement process, regulated by a learning rate $\boldsymbol{\eta}$, enables RealKcat to iteratively correct errors while stabilizing updates, effectively capturing subtle variations in kinetic parameters across different enzyme classes. Mathematically, the **additive** update for each tree in the boosting process can be described by:

$$f_{(k+1)}(X) = f_{(k)}(X) + \eta \times PseudoResiduals$$

where η is tuned to gradually refine predictions without over-adjusting for noise in the data.

Hyperparameter Optimization*:* To ensure robust model performance, we conducted rigorous hyperparameter tuning using a 5-fold StratifiedKFold cross-validation approach. This method preserved class distribution across folds, which was essential given the multi-cluster nature of enzyme kinetics data. We optimized key hyperparameters, which were finalized as follows:

- **Number of Estimators**: 318
- **Maximum Depth**: 50
- **Learning Rate (η)**: 0.0753
- **Maximum Delta Step (δ)**: 4
- **L1 Regularization (α)**: 1.23

These hyperparameters were selected based on their impact on both accuracy and stability, carefully balancing model complexity and avoiding overfitting. To prevent overtraining, we

introduced early stopping with a patience threshold of 2 rounds on validation metrics, which helped in achieving an optimal balance between accuracy and generalization.

***Feature Standardization**:* Given the heterogeneous nature of the input features, each embedding group—ESM-2 and ChemBERTa—was independently standardized using the mean and standard deviation from the training set. Specifically, each feature $\boldsymbol{x}$ in the embedding was transformed as follows:

$$\boldsymbol{x}' = \frac{(\boldsymbol{x} - \boldsymbol{\mu})}{\boldsymbol{\sigma}}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the mean and standard deviation of the feature in the training set. This standardization was applied consistently across validation and test sets, ensuring a stable feature distribution across datasets, aiding in model convergence, and enhancing generalization across diverse kinetic parameter classes

Performance and Evaluation Metrics

RealKcat's predictive accuracy and robustness were evaluated using a suite of metrics tailored to multi-class classification, capturing various dimensions of model performance across different kinetic parameter classes. The primary metric, **accuracy**, calculated as: $Accuracy = \frac{\sum_i^N (\hat{y}_i = y_i)}{N}$, $\hat{y}_i$ and y_i are the predicted and true labels, respectively, measures the overall proportion of correctly classified instances across all kinetic clusters. For more detailed class-specific performance, **precision** and **F1-score** were calculated per class.

An essential additional metric, effectively used in this study, for enzyme kinetics was the **e-accuracy** (error-tolerance accuracy), which calculates the percentage of predictions within one order of magnitude of the true kinetic parameter cluster: $e\_Accuracy = \frac{\sum_i^N (|\hat{y}_i - y_i| \leq 1)}{N}$. Additionally, **confusion matrices** and **t-SNE visualizations** were used to obtain a visual assessment of RealKcat's performance across clusters. Confusion matrices displayed accuracies of predictions per class, while t-SNE plots revealed the clustering quality and distinct separation of the samples concatenated feature representations overlayed with class labels, reflecting RealKcat's proficiency in managing complex, multi-class kinetic data. These metrics, including

accuracy, precision, F1-score, e-accuracy, and confusion matrices, were implemented in python using the scikit-learn (sklearn) library.

Catalytic Relevance Validation with External Datasets

To validate RealKcat's capacity for mutation-aware predictions, we evaluated its performance on two external datasets: an independent enzyme dataset spanning six enzymes from *Saccharomyces cerevisiae* and *Escherichia coli*, and the high-throughput alkaline phosphatase dataset (PafA) from *Elizabethkingia meningoseptica* (Markin et al., 2021) (*23*). These datasets were selected to test RealKcat's predictive accuracy and sensitivity to point mutations at catalytic residues.

For the independent enzyme dataset, meaning we excluded these data points from the established training set, hence "out of distribution". Here, we define "out of distribution" as having zero sequences with ≥60% identity and no sequences with 100% identity to the training database. This criterion holds for each of the six enzymes except for P00942, which has neighboring sequences at ≤95% identity but no exact (100%) matches in the training set, see fig. S2. We curated enzymes involved in central metabolic pathways, including glycolysis, the tricarboxylic acid (TCA) cycle, and the pentose phosphate pathway. Point mutations were introduced at critical catalytic residues, as annotated by UniProt, primarily substituting with alanine to examine changes in kinetic behavior. RealKcat's cluster predictions of k_{cat} and K_M were evaluated against experimental data and benchmarked against models including DLKcat, UniKP, CatPred, and EITLEM-Kinetics. The primary focus was on RealKcat's ability to detect decreases in catalytic activity and to accurately predict substrate affinities post-mutation.

The second validation dataset involved the alkaline phosphatase PafA from *Elizabethkingia meningoseptica*, provided a large-scale validation with 1,016 single-point mutations introduced at each residue, to probe functional shifts. Using the substrate carboxy 4-methylumbelliferyl phosphate ester (cMUP), this dataset allowed for detailed assessment of RealKcat's sensitivity to catalytic and residue-positional modifications. A carefully controlled partitioning strategy was employed to ensure the validity of RealKcat's mutation-aware predictions. The data was divided into training, validation, and test sets with 554, 310, and 310 data points, respectively. Within the training set, we included the wild-type (WT) sequence and a single catalytic site mutation (T79S), enabling the model to learn baseline catalytic information

without complete exposure to all critical sites. Conversely, the test set contained mutations at a second catalytic residue, specifically R164A and R164G in test and R164V in validation, which were withheld during training to rigorously assess RealKcat's mutation-aware predictions on unseen, catalytically impactful variants.

Computational Resources

This study leveraged the high-performance computational resources of the University of Nebraska-Lincoln's SWAN HCC cluster, operating on Ubuntu Linux. Our primary system setup featured dual Intel(R) Xeon(R) Gold 6248R CPUs, totaling 48 cores (used only 16 cores) and 187 GiB of RAM, of which approximately 143 GiB was available for handling large datasets. For GPU-accelerated tasks, we employed an NVIDIA Tesla V100S-PCIE-32GB, running CUDA 12.4 with NVIDIA driver version 550.127.05, which provided 32 GiB of VRAM—around 17.9 GiB of which was dedicated to embeddings and deep learning computations. Additionally, the Iowa State University high-performance computer NOVA contributed to data cleaning, embeddings, and model training, utilizing dual 18-core Intel Xeon Gold 6140 CPUs (36 cores total), 192 GiB RAM, and dual NVIDIA Tesla V100-32GB GPUs with CUDA 11.8 for GPU-accelerated processing.

Several components of this implementation were managed using Python 3.12.5, integrating a robust suite of libraries for machine learning and biochemical data processing. The primary libraries included XGBoost (version 2.1.1) for model training and prediction, and PyTorch (version 2.4.0) for handling tensor operations and neural network computations. To balance class distributions, we used the imbalanced-learn library (version 0.12.3) for SMOTE applications, while Scikit-learn (version 1.5.1) provided tools for calculating evaluation metrics and model performance. For reproducibility, we set a random seed of 42 across all stochastic processes, ensuring consistency in SMOTE applications, train-validation-test splits, and model initialization.

Code availability: An open-source notebook with code is made available in a public repository to foster use easy, through the link: https://github.com/TKAI-LAB-Mali/RealKcat. Additionally, we are committed to making all scripts, notebooks, and relevant code available through a public repository upon publication, supporting reproducibility and enabling further exploration by the research community.

Acknowledgments

A.O. wishes to express gratitude for the financial support provided for this research. This includes the NIH MIRA Award (5R35GM143009) awarded to R.S. This work is also partially supported by the Iowa State University Startup Grant, and NSF EPSCoR RII Track-1, Award Number DQDBM7FGJPC5 to R.C., and Iowa Economic Development Authority Award Number: 24IEC006 to R.C. and R.S.

Funding:

National Institutes of Health grant MIRA Award 5R35GM143009 (R.S.)

Iowa State Startup Grant; Building A World of Difference Faculty Fellowship (R.C.)

NSF EPSCoR RII Track-1, Award Number DQDBM7FGJPC5 (R.C.)

Iowa Economic Development Authority Award Number: 24IEC006 (R.C. and R.S.)

Data and materials availability: The entire KinHub-27k dataset and accompanying code will be made available for download and inference ***upon publication*** at the link: https://chowdhurylab.github.io/downloads.html. However, an open-source notebook with code is made available in a public repository to foster use easy, through the link: https://github.com/TKAI-LAB-Mali/RealKcat. Additionally, we are committed to making all scripts, notebooks, and relevant code available through a public repository upon publication, supporting reproducibility and enabling further exploration by the research community.

References

1. D. Davidi, R. Milo, Lessons on enzyme kinetics from quantitative proteomics. *Curr Opin Biotechnol* **46**, 81–89 (2017).
2. G. A. Holdgate, T. D. Meek, R. L. Grimley, Mechanistic enzymology in drug discovery: a fresh perspective. *Nature Reviews Drug Discovery 2017 17:2* **17**, 115–132 (2017).
3. N. J. Turner, Directed evolution drives the next generation of biocatalysts. *Nature Chemical Biology 2009 5:8* **5**, 567–573 (2009).
4. Y. Okada, Q. S. Wang, A massive effort links protein-coding gene variants to health. *Nature 2021 599:7886* **599**, 561–563 (2021).
5. F. Li, L. Yuan, H. Lu, G. Li, Y. Chen, M. K. M. Engqvist, E. J. Kerkhoven, J. Nielsen, Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis 2022 5:8* **5**, 662–672 (2022).
6. A. Kroll, Y. Rousset, X. P. Hu, N. A. Liebrand, M. J. Lercher, Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. *Nature Communications 2023 14:1* **14**, 1–14 (2023).

7. H. Yu, H. Deng, J. He, J. D. Keasling, X. Luo, UniKP: a unified framework for the prediction of enzyme kinetic parameters. *Nature Communications 2023 14:1* **14**, 1–13 (2023).
8. V. S. Boorla, C. D. Maranas, CatPred: A comprehensive framework for deep learning in vitro enzyme kinetic parameters kcat, Km and Ki. *bioRxiv*, 2024.03.10.584340 (2024).
9. D. Heckmann, C. J. Lloyd, N. Mih, Y. Ha, D. C. Zielinski, Z. B. Haiman, A. A. Desouki, M. J. Lercher, B. O. Palsson, Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nature Communications 2018 9:1* **9**, 1–10 (2018).
10. A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* **118**, e2016239118 (2021).
11. A. Chang, L. Jeske, S. Ulbrich, J. Hofmann, J. Koblitz, I. Schomburg, M. Neumann-Schaal, D. Jahn, D. Schomburg, BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res* **49**, D498–D508 (2021).
12. D. Dudaš, U. Wittig, M. Rey, A. Weidemann, W. Müller, Improved insights into the SABIO-RK database via visualization. *Database* **2023**, 1–9 (2023).
13. X. Shen, Z. Cui, J. Long, S. Zhang, B. Chen, T. Tan, EITLEM-Kinetics: A deep-learning framework for kinetic parameter prediction of mutant enzymes. *Chem Catalysis* **4**, 101094 (2024).
14. A. Bateman, M. J. Martin, S. Orchard, M. Magrane, S. Ahmad, E. Alpi, E. H. Bowler-Barnett, R. Britto, H. Bye-A-Jee, A. Cukura, P. Denny, T. Dogan, T. G. Ebenezer, J. Fan, P. Garmiri, L. J. da Costa Gonzales, E. Hatton-Ellis, A. Hussein, A. Ignatchenko, G. Insana, R. Ishtiaq, V. Joshi, D. Jyothi, S. Kandasaamy, A. Lock, A. Luciani, M. Lugaric, J. Luo, Y. Lussi, A. MacDougall, F. Madeira, M. Mahmoudy, A. Mishra, K. Moulang, A. Nightingale, S. Pundir, G. Qi, S. Raj, P. Raposo, D. L. Rice, R. Saidi, R. Santos, E. Speretta, J. Stephenson, P. Totoo, E. Turner, N. Tyagi, P. Vasudev, K. Warner, X. Watkins, R. Zaru, H. Zellner, A. J. Bridge, L. Aimo, G. Argoud-Puy, A. H. Auchincloss, K. B. Axelsen, P. Bansal, D. Baratin, T. M. Batista Neto, M. C. Blatter, J. T. Bolleman, E. Boutet, L. Breuza, B. C. Gil, C. Casals-Casas, K. C. Echioukh, E. Coudert, B. Cuche, E. de Castro, A. Estreicher, M. L. Famiglietti, M. Feuermann, E. Gasteiger, P. Gaudet, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz, C. Hulo, N. Hyka-Nouspikel, F. Jungo, A. Kerhornou, P. Le Mercier, D. Lieberherr, P. Masson, A. Morgat, V. Muthukrishnan, S. Paesano, I. Pedruzzi, S. Pilbout, L. Pourcel, S. Poux, M. Pozzato, M. Pruess, N. Redaschi, C. Rivoire, C. J. A. Sigrist, K. Sonesson, S. Sundaram, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, J. Zhang, UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **51**, D523–D531 (2023).
15. T. Paysan-Lafosse, M. Blum, S. Chuguransky, T. Grego, B. L. Pinto, G. A. Salazar, M. L. Bileschi, P. Bork, A. Bridge, L. Colwell, J. Gough, D. H. Haft, I. Letunić, A. Marchler-Bauer, H. Mi, D. A. Natale, C. A. Orengo, A. P. Pandurangan, C. Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, A. Bateman, InterPro in 2022. *Nucleic Acids Res* **51**, D418–D427 (2023).
16. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives,

Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science (1979)* **379**, 1123–1130 (2023).

17. S. Chithrananda, G. Grand, B. R. Deepchem, ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. (2020).
18. A. Khodayari, C. D. Maranas, A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. *Nat Commun* **7** (2016).
19. Y. Chen, F. Li, J. Nielsen, Genome-scale modeling of yeast metabolism: retrospectives and perspectives. *FEMS Yeast Res* **22** (2022).
20. H. Lu, F. Li, B. J. Sánchez, Z. Zhu, G. Li, I. Domenzain, S. Marcišauskas, P. M. Anton, D. Lappa, C. Lieven, M. E. Beber, N. Sonnenschein, E. J. Kerkhoven, J. Nielsen, A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nature Communications 2019 10:1* **10**, 1–13 (2019).
21. A. Nilsson, J. Nielsen, B. O. Palsson, Metabolic Models of Protein Allocation Call for the Kinetome. *Cell Syst* **5**, 538–541 (2017).
22. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **16**, 321–357 (2002).
23. C. J. Markin, D. A. Mokhtari, F. Sunden, M. J. Appel, E. Akiva, S. A. Longwell, C. Sabatti, D. Herschlag, P. M. Fordyce, Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. *Science (1979)* **373** (2021).
24. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **13-17-August-2016**, 785–794 (2016).
25. A. Abdel-Rehim, O. Orhobor, L. Hang, H. Ni, R. D. King, Protein-ligand binding affinity prediction exploiting sequence constituent homology. *Bioinformatics* **39** (2023).
26. X. S. Xie, Enzyme kinetics, past and present. *Science (1979)* **342**, 1457–1459 (2013).
27. A. Bar-Even, E. Noor, Y. Savir, W. Liebermeister, D. Davidi, D. S. Tawfik, R. Milo, The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* **50**, 4402–4410 (2011).
28. J. M. Woodley, Accelerating the implementation of biocatalysis in industry. *Appl Microbiol Biotechnol* **103**, 4733–4739 (2019).
29. A. Osinuga, A. G. Solís, R. E. Cahoon, A. Al-Siyabi, E. B. Cahoon, R. Saha, Deciphering Sphingolipid Biosynthesis Dynamics in Arabidopsis thaliana cell cultures: Quantitative Analysis Amidst Data Variability. *iScience* **0**, 110675 (2024).
30. E. J. O'Brien, J. A. Lerman, R. L. Chang, D. R. Hyduke, B. Palsson, Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol Syst Biol* **9** (2013).
31. Z. Mao, X. Zhao, X. Yang, P. Zhang, J. Du, Q. Yuan, H. Ma, ECMpy, a Simplified Workflow for Constructing Enzymatic Constrained Metabolic Network Model. *Biomolecules* **12**, 65 (2022).
32. B. J. Sánchez, C. Zhang, A. Nilsson, P. Lahtvee, E. J. Kerkhoven, J. Nielsen, Improving the phenotype predictions of a yeast genome□scale metabolic model by incorporating enzymatic constraints. *Mol Syst Biol* **13** (2017).
33. D. M. Blei, A. Kucukelbir, J. D. McAuliffe, Variational Inference: A Review for Statisticians. *J Am Stat Assoc* **112**, 859–877 (2017).

34. S. Choudhury, B. Narayanan, M. Moret, V. Hatzimanikatis, L. Miskovic, Generative machine learning produces kinetic models that accurately characterize intracellular metabolic states. *Nature Catalysis 2024 7:10* **7**, 1086–1098 (2024).
35. S. Choudhury, M. Moret, P. Salvy, D. Weilandt, V. Hatzimanikatis, L. Miskovic, Reconstructing Kinetic Models for Dynamical Studies of Metabolism using Generative Adversarial Networks. *Nature Machine Intelligence 2022 4:8* **4**, 710–719 (2022).
36. O. Sakamoto, Y. Suzuki, X. Li, Y. Aoki, M. Hiratsuka, T. Suormala, E. R. Baumgartner, K. M. Gibson, K. Narisawa, Relationship between Kinetic Properties of Mutant Enzyme and Biochemical and Clinical Responsiveness to Biotin in Holocarboxylase Synthetase Deficiency. *Pediatric Research 1999 46:6* **46**, 671–671 (1999).
37. M. A. A. Adam, C. D. Sohl, Probing altered enzyme activity in the biochemical characterization of cancer. *Biosci Rep* **42**, BSR20212002 (2022).
38. S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E. E. Bolton, PubChem 2023 update. *Nucleic Acids Res* **51**, D1373–D1380 (2023).

Figure and Tables

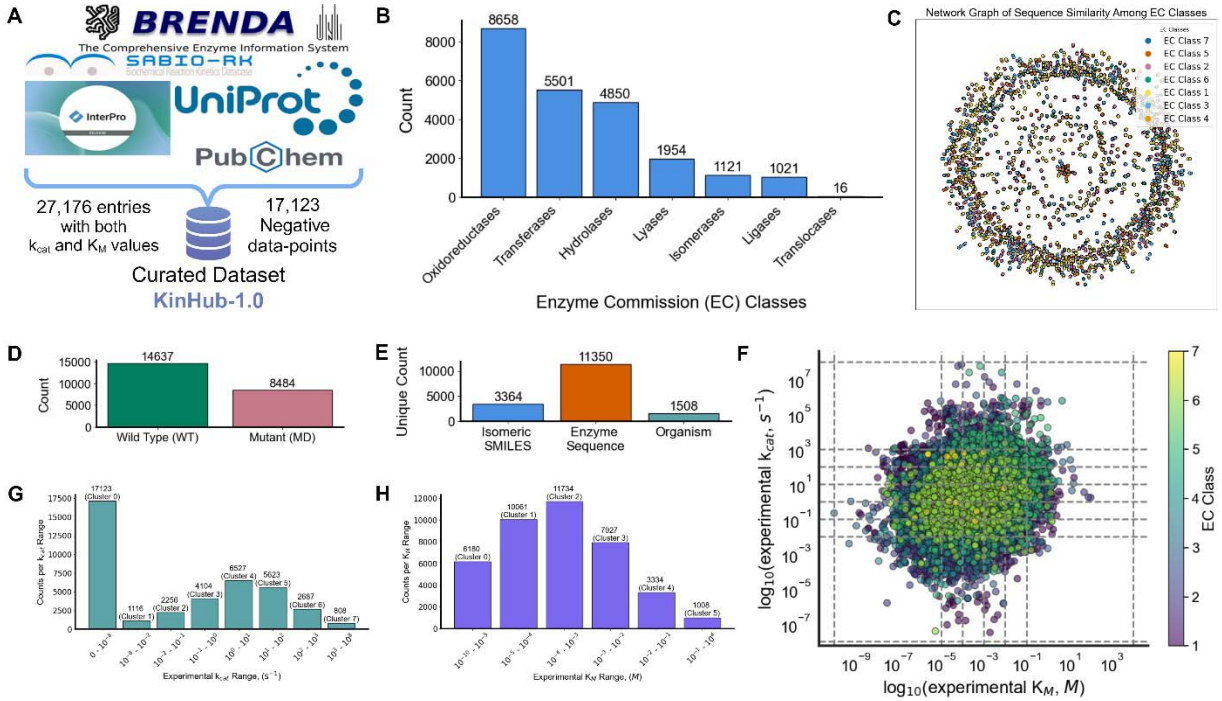


Fig. 1. Comprehensive dataset curation, classification, and distribution for RealKcat model training. (A) Raw data for KinHub-27k was collated from multiple databases, including BRENDA, SABIO-RK, UniProt, PubChem, and InterPro, resulting in a curated dataset with 27,176 entries containing both k_{cat} and K_M values, and an additional 17,123 negative data points.

(B) Distribution of enzyme entries across Enzyme Commission (EC) classes, highlighting a broad representation of enzymatic functions. **(C)** Network graph showing sequence similarity among enzymes within different EC classes, indicating clustering by sequence homology. **(D)** Counts of wild-type (WT) versus mutant (MD) entries, reflecting the dataset's inclusion of both naturally occurring and engineered enzyme variants. **(E)** Unique counts of isomeric SMILES for substrates and enzyme sequences, indicating the dataset's structural diversity. **(F)** Log-log scatter plot of experimental K_M and k_{cat} values, depicting the core dynamic range of kinetic parameters. **(G & H)** Distribution of kinetic parameter clusters for k_{cat} and K_M, respectively, categorized by order of magnitude to facilitate robust multiclass classification.

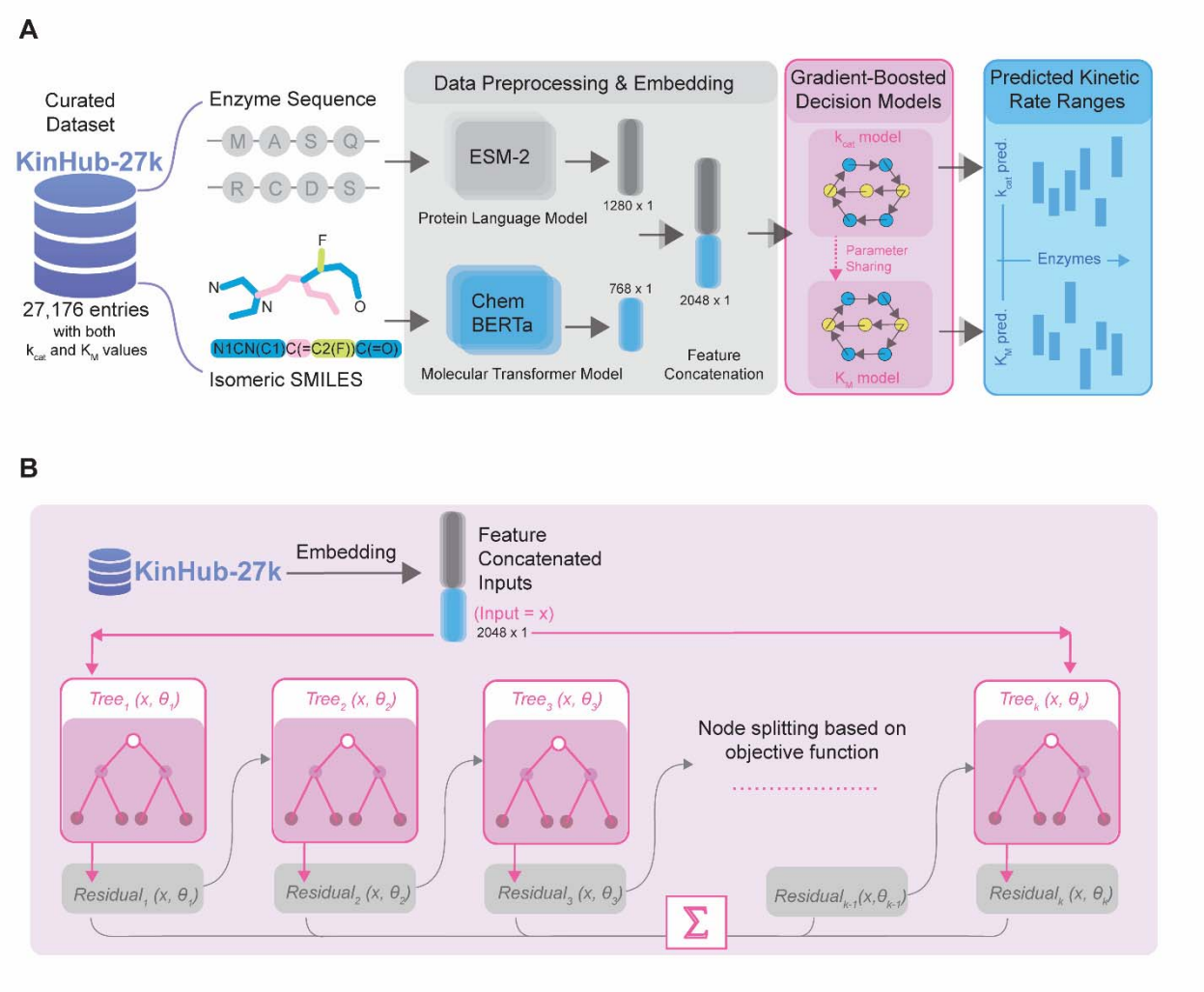


Fig. 2. RealKcat model workflow and architecture for enzyme kinetics prediction. (A) The KinHub-27k dataset with 27,176 entries provides enzyme sequence and substrate information, embedded with ESM-2 (1280-dimensional) and ChemBERTa (768-dimensional) models, respectively. The combined 2048-dimensional feature vectors are used to predict k_{cat} and K_M values through separate gradient-boosted models, which share hyperparameters. Predictions are output as ranges, enhancing model stability across diverse enzyme-substrate pairs. **(B)** Gradient-boosted decision tree model architecture. Feature inputs are iteratively refined through an additive process of decision trees, each learning from the previous residuals to improve classification accuracy for predict k_{cat} and K_M ranges

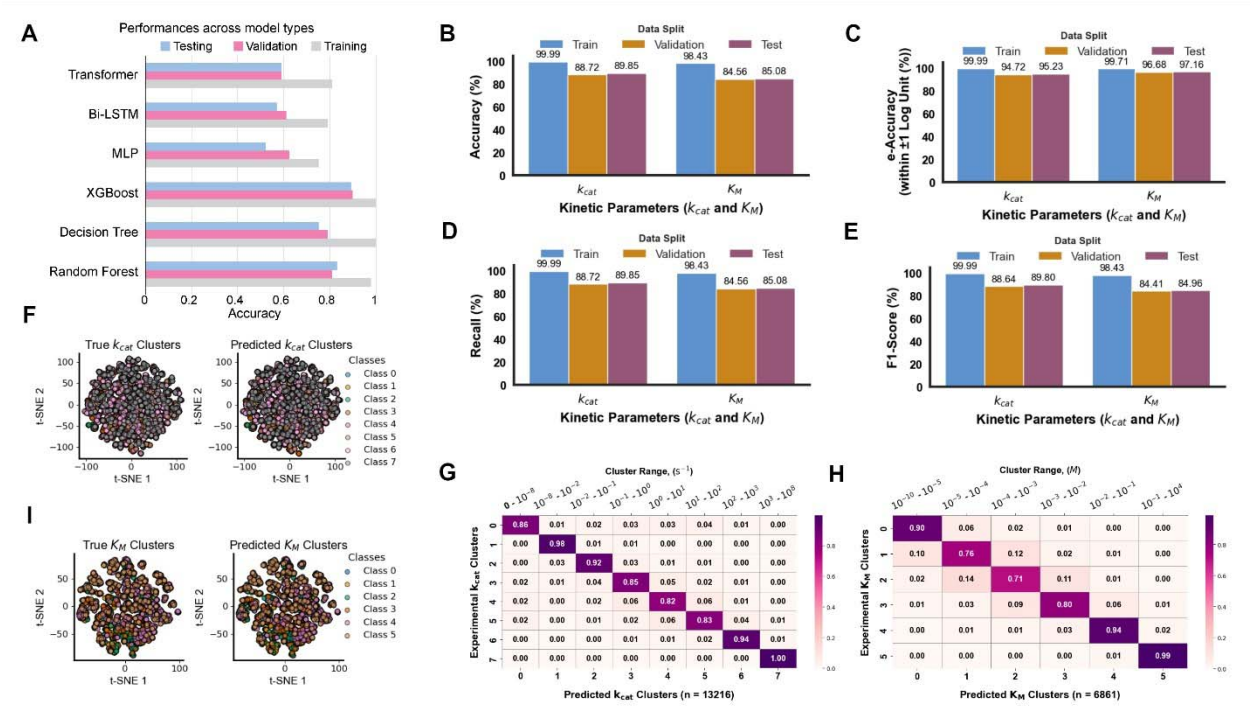


Fig. 3. Training and performance evaluation of RealKcat across accuracy metrics and cluster distributions for k_{cat} and K_M. (A) Various model performance across different data splits. **(B)** Accuracy (%) for training, validation, and test sets across k_{cat} and K_M parameters, showing strong performance consistency. **(C)** e-Accuracy within ±1 log unit, indicating RealKcat's precision within one order of magnitude for kinetic predictions. **(D)** Recall (%) across data splits, highlighting RealKcat's sensitivity to true positive rates. **(E)** F1-score (%) comparisons further validate the model's balanced precision and recall. **(F&I)** t-SNE visualizations of true vs. predicted clusters for k_{cat} and K_M, showcasing effective high-dimensional separation and clustering, underscoring RealKcat's robustness in capturing complex kinetic spaces. **(G&H)** Confusion matrices for k_{cat} and K_M cluster ranges reveal high accuracy within central clusters and low misclassification at extremes.

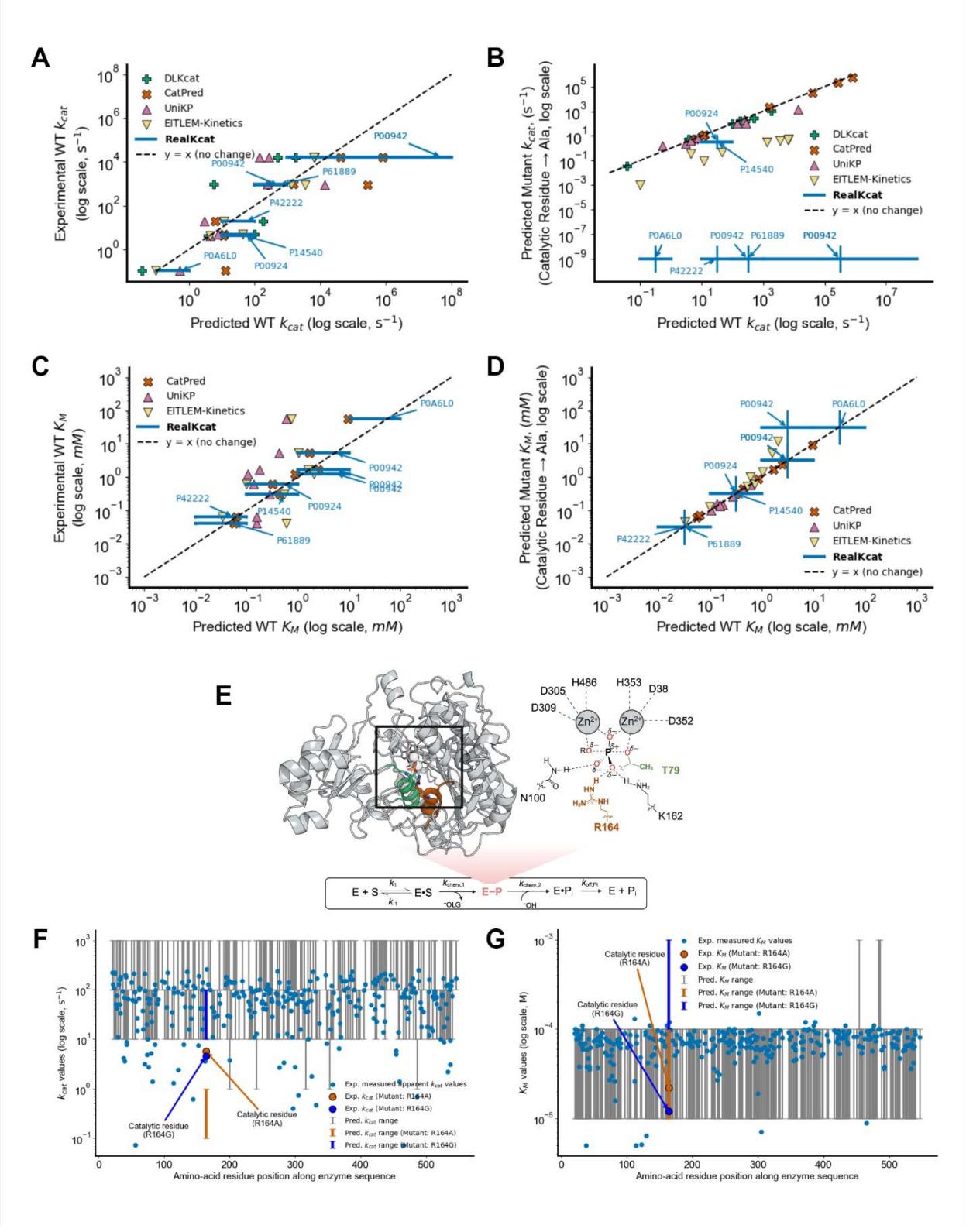


Fig. 4. Validation of RealKcat's catalytic relevance prediction and sensitivity to mutations across enzyme-substrate complexes. (A-D) Comparison of predicted versus experimental kinetic parameters k_{cat} and K_M for wild-type and mutant enzymes. **(A)** Predicted versus

experimental k_{cat} values for wild-type enzymes, demonstrating high alignment with experimental data, benchmarked against other models. **(B)** Predicted k_{cat} values for catalytically relevant alanine mutations, highlighting RealKcat's sensitivity to reduced activity upon mutation of key residues. **(C)** Wild-type K_M predictions, showing strong correlation with measured values across diverse substrates. **(D)** Predicted K_M values for catalytic residue mutants. **(E)** Structure of alkaline phosphatase PafA with highlighted catalytic residues in transition state, showing positions T79 and R164 used in mutant testing. **(F-G)** Sequence-wide distribution of k_{cat} **(F)** and K_M **(G)** values across the PafA test dataset, with RealKcat predictions compared to experimental data, specifically for key R164A and R164G mutations.

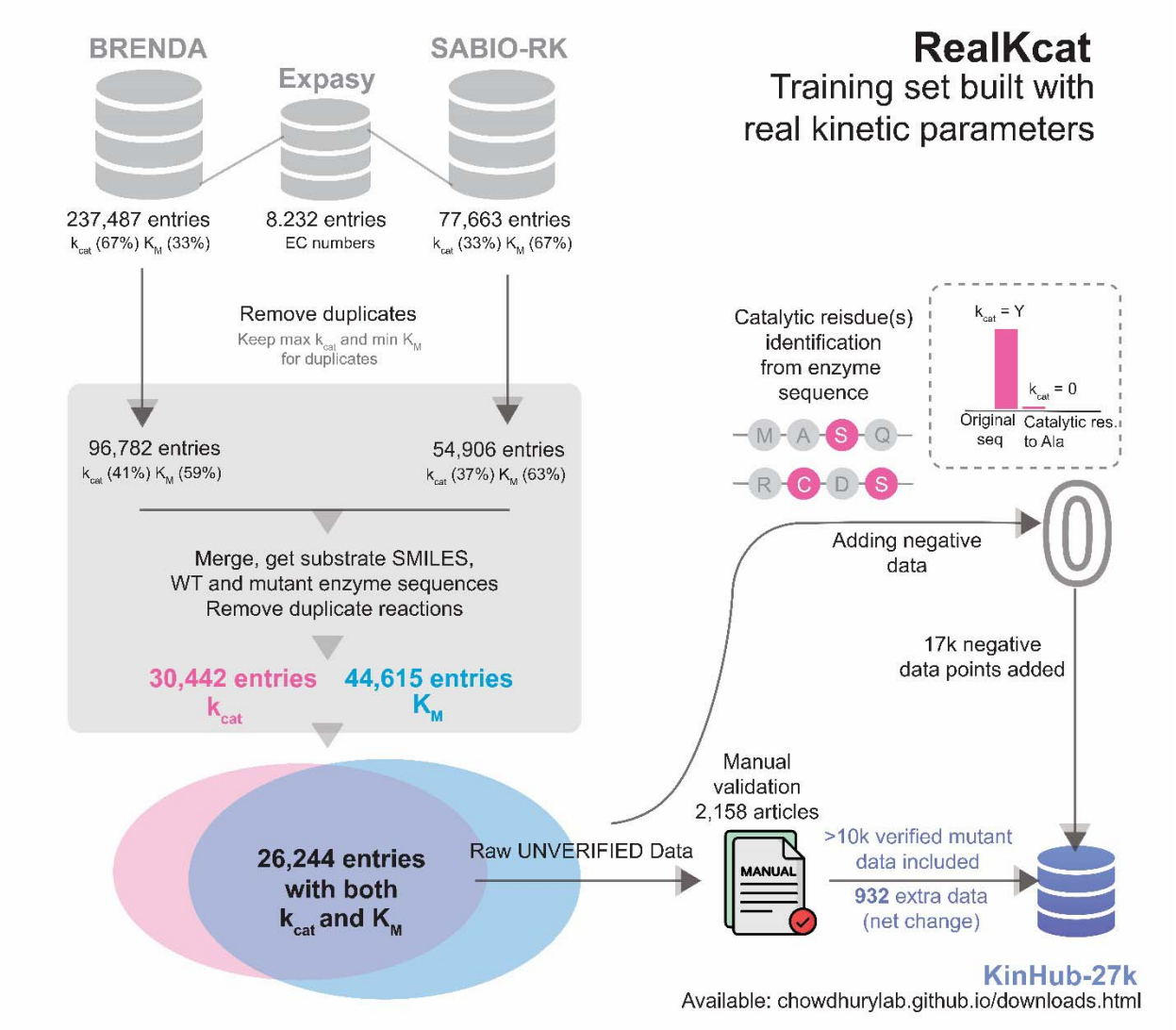


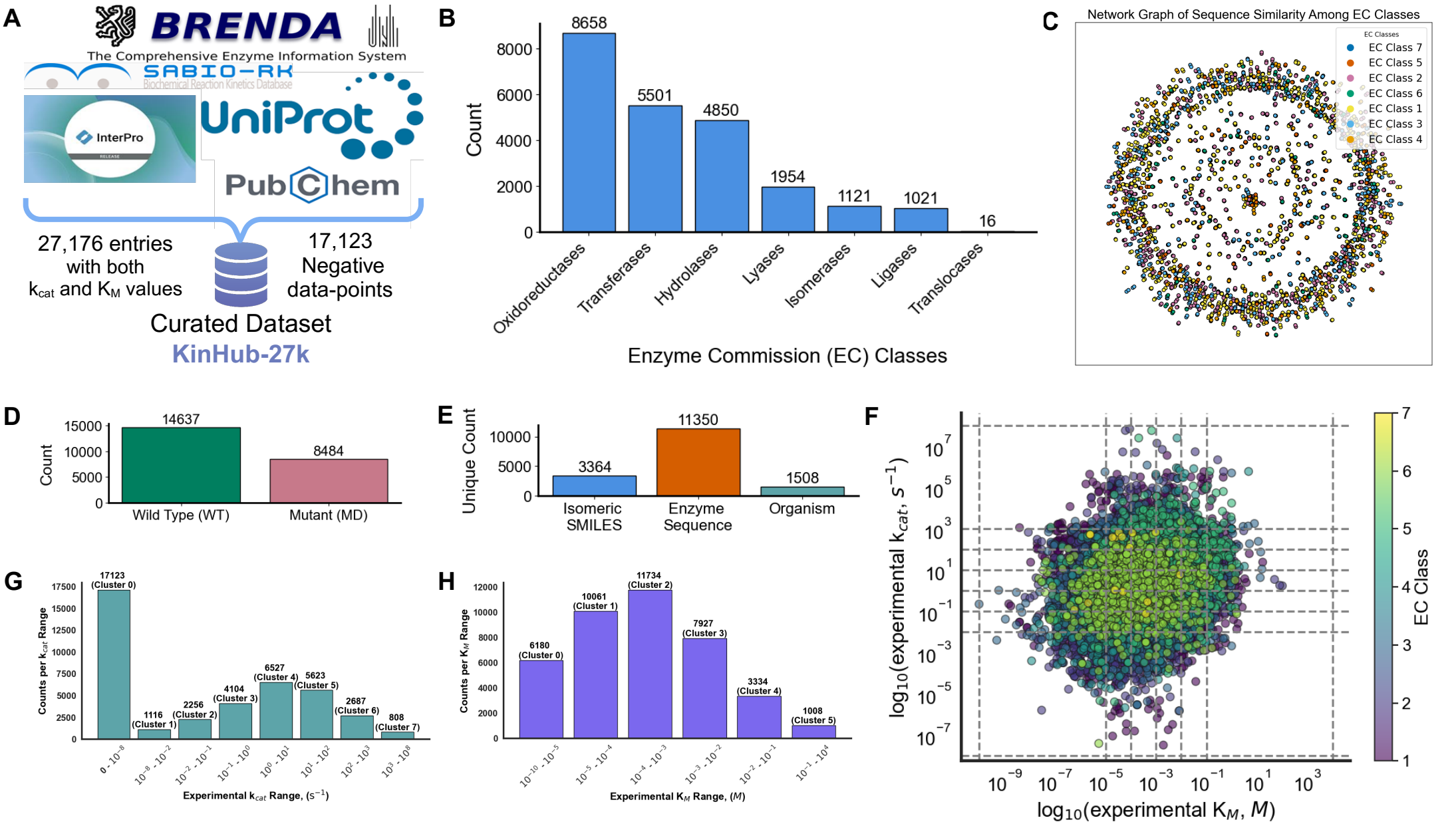
Fig. 5. Data collection steps for constructing KinHub-27k from BRENDA and SABIO-RK databases. Raw enzyme kinetic data were sourced from BRENDA (237,487 entries) and SABIO-RK (77,663 entries), covering 8,232 unique EC numbers. (A) Initial processing involved removing duplicate entries while retaining maximum k_{cat} and minimum K_M values, resulting in 96,782 entries from BRENDA and 54,906 from SABIO-RK. (B) After merging the datasets, obtaining SMILES for substrates, and distinguishing between wild-type and mutant sequences, further duplicates were removed, yielding 30,442 unique entries for k_{cat} and 44,615 for K_M. The resulting dataset, KinHub-27k, contains 26,244 entries with paired k_{cat} and K_M values, prepared for manual validation through article review of the original data sources.

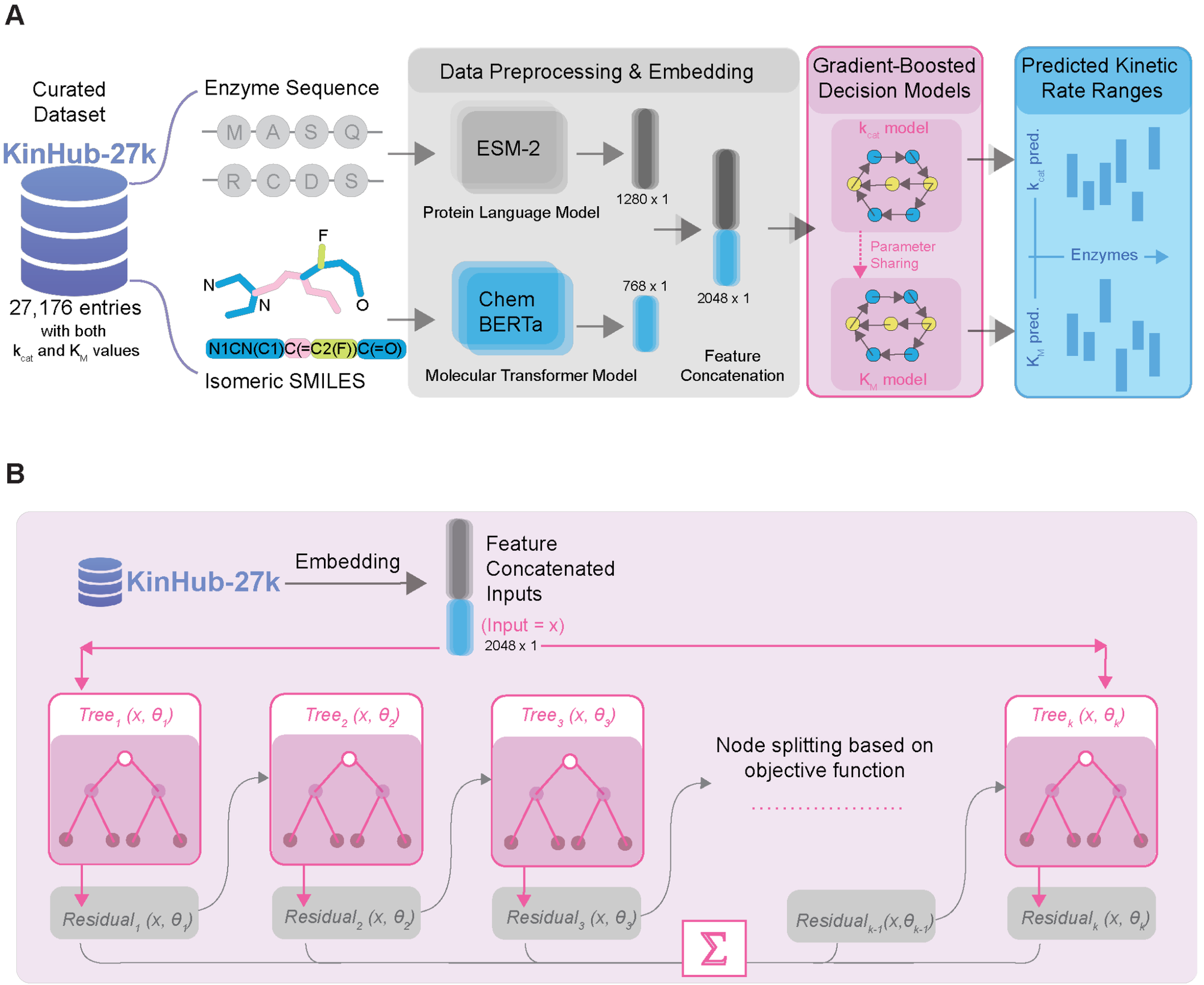
Table 1. Independent out-of-distribution validation set used for RealKcat model evaluation. This dataset was specifically chosen to test RealKcat's predictive and catalytic residue awareness performance on unseen enzyme-substrate pairs distribution. (PPP – Pentose Phosphate Pathway, TCA- Tricarboxylic Acid Cycle)

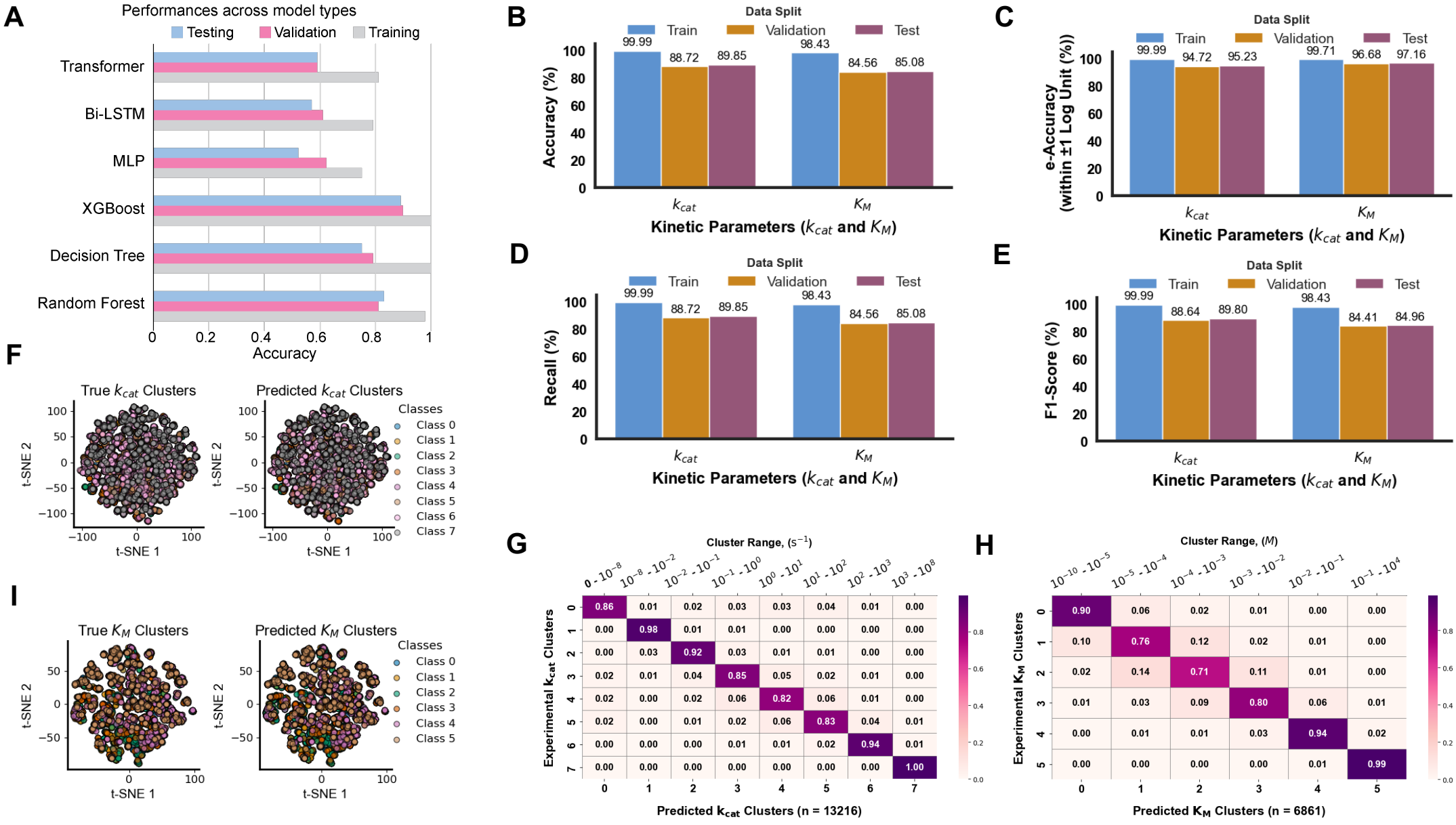
EC Number	UniProt ID	Substrate Name	Catalytic Residues	k_{cat} [s^{-1}]	K_M [mM]	Organism	Pathway
1.1.1.37	P61889	oxaloacetate	H177	$9.3x10^{2}$	$4.0x10^{-2}$	*E. coli*	TCA
5.3.1.1	P00942	dihydroxyacetone 3-phosphate	H95, E165	$8.6x10^{2}$	$1.7x10^{0}$	*S. cerevisiae*	Glycolysis
4.1.2.4	P0A6L0	2-deoxy-d-ribose	D102, K167, K201	$1.1x10^{-1}$	$5.7x10^{1}$	*E. coli*	PPP
4.2.1.11	P00924	3-phospho-d-glycerate	E212, K346	$5.1x10^{0}$	$6.2x10^{-1}$	*S. cerevisiae*	Glycolysis
4.2.1.11	P42222	2-phospho-d-glycerate	E212, K346	$1.9x10^{1}$	$6.5x10^{-2}$	*S. cerevisiae*	Glycolysis
5.3.1.1	P00942	d-glyceraldehyde 3-phosphate	H95, E165	$1.7x10^{4}$	$5.3x10^{0}$	*S. cerevisiae*	Glycolysis
4.1.2.13	P14540	d-glyceraldehyde 3-phosphate	D110	$4.1x10^{0}$	$3.0x10^{-1}$	*S. cerevisiae*	Glycolysis
5.3.1.1	P00942	phosphoenolpyruvate	H95, E165	$1.7x10^{4}$	$1.3x10^{0}$	*S. cerevisiae*	Glycolysis

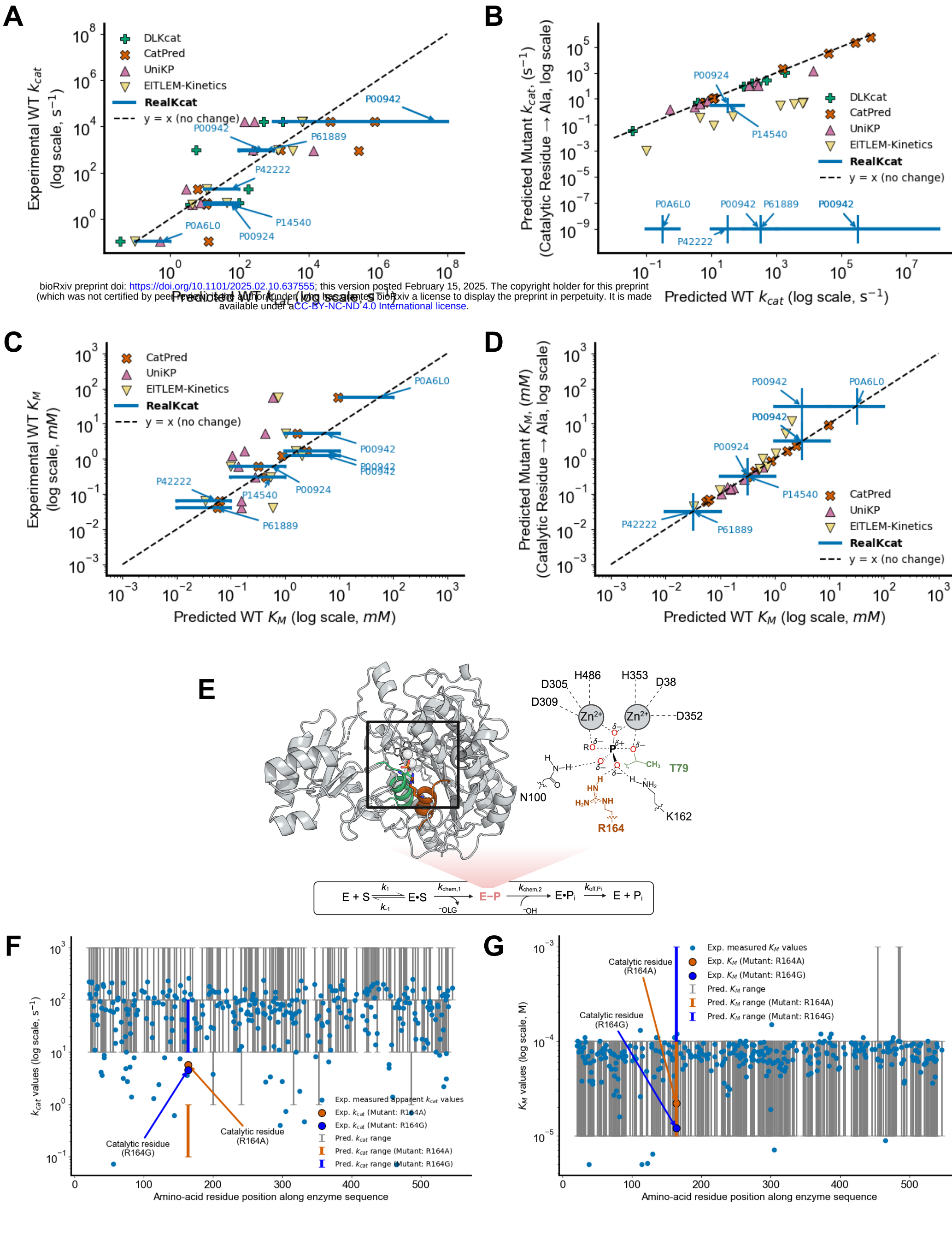
Table 2. Resolution and Additions in the KinHub-27k Mutant Dataset. This table details the curation process, highlighting resolutions and removals across parameters (K_M and k_{cat}) substrates, and mutations. 1804 discrepancies were resolved, with 1,072 new entries added, refining the dataset to 11,175 mutant entries for predictive modeling.

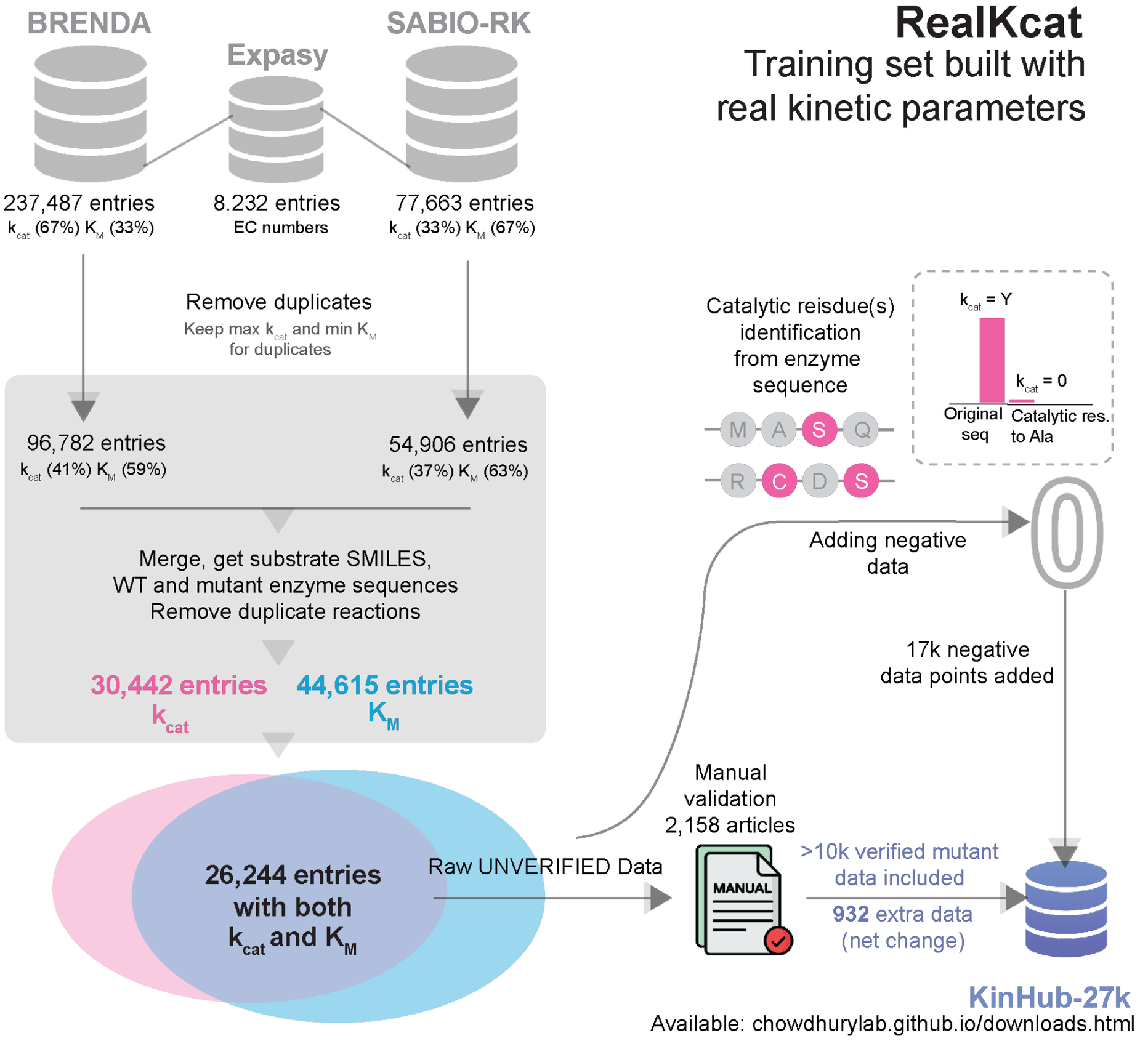
Parameters	Corrected	Removed
K_M	788	10
k_{cat}	618	10
Substrates	240	8
Mutations	18	21
Duplicates	-	91
Additional data	1072	-

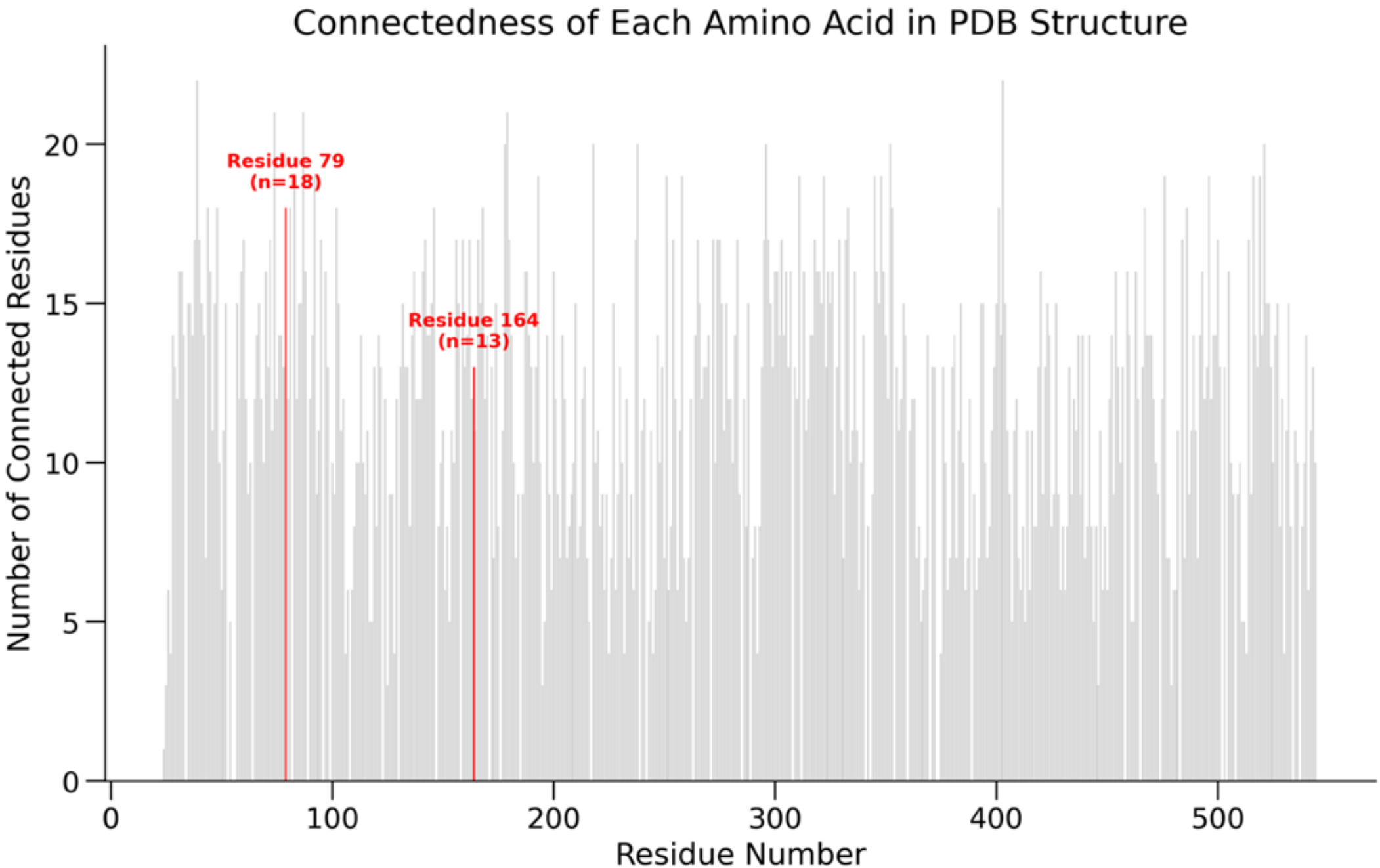

A
BRENDA
The Comprehensive Enzyme Information System
SABIO-RK
Biochemical Reaction Kinetics Database
InterPro
UniProt
PubChem
27,176 entries with both k_{cat} and K_M values
17,123 Negative data-points
Curated Dataset
KinHub-27k
B
Count
8658
5501
4850
1954
1121
1021
16
Oxidoreductases
Transferases
Hydrolases
Lyases
Isomerases
Ligases
Translocases
Enzyme Commission (EC) Classes
C
Network Graph of Sequence Similarity Among EC Classes
EC Classes
EC Class 7
EC Class 5
EC Class 2
EC Class 6
EC Class 1
EC Class 3
EC Class 4
D
Count
14637
8484
Wild Type (WT)
Mutant (MD)
E
Unique Count
3364
11350
1508
Isomeric SMILES
Enzyme Sequence
Organism
F
$\log_{10}$(experimental k_{cat}, s^{-1})
$\log_{10}$(experimental K_M, M)
EC Class
G
Counts per k_{cat} Range
17123 (Cluster 0)
1116 (Cluster 1)
2256 (Cluster 2)
4104 (Cluster 3)
6527 (Cluster 4)
5623 (Cluster 5)
2687 (Cluster 6)
808 (Cluster 7)
Experimental k_{cat} Range, (s^{-1})
H
Counts per K_M Range
6180 (Cluster 0)
10061 (Cluster 1)
11734 (Cluster 2)
7927 (Cluster 3)
3334 (Cluster 4)
1008 (Cluster 5)
Experimental K_M Range, (M)

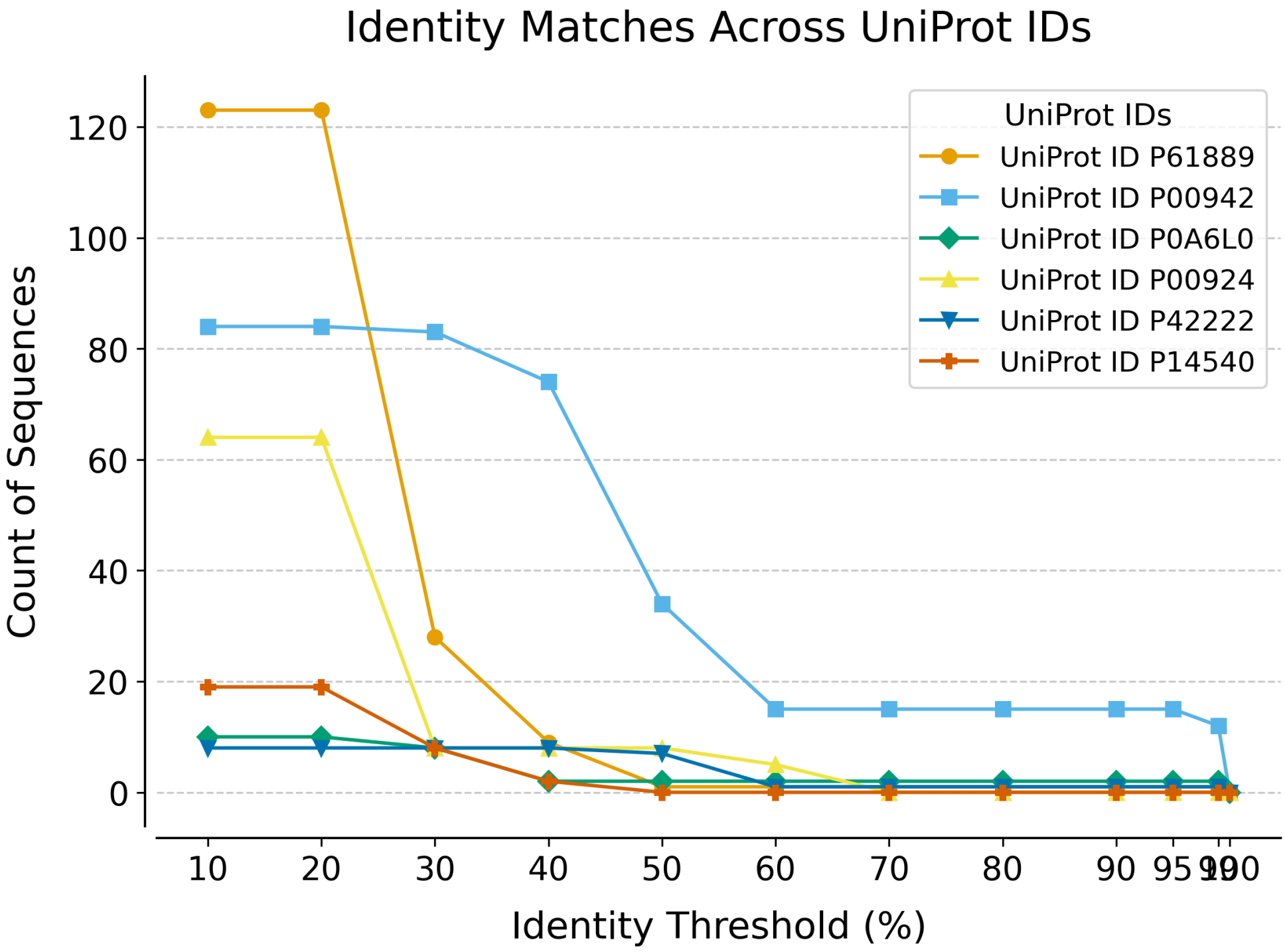

A
Curated Dataset
KinHub-27k
27,176 entries
with both
k_cat and K_M values
Enzyme Sequence
M A S Q
R C D S
N1CN(C1)C(=C2(F))C(=O)
Isomeric SMILES
Data Preprocessing & Embedding
ESM-2
Protein Language Model
1280 x 1
Chem BERTa
Molecular Transformer Model
768 x 1
2048 x 1
Feature Concatenation
Gradient-Boosted Decision Models
k_cat model
Parameter Sharing
K_M model
Predicted Kinetic Rate Ranges
k_cat pred.
Enzymes
K_M pred.
B
KinHub-27k
Embedding
Feature Concatenated Inputs
(Input = x)
2048 x 1
Tree_1 (x, θ_1)
Tree_2 (x, θ_2)
Tree_3 (x, θ_3)
Node splitting based on objective function
Tree_k (x, θ_k)
Residual_1 (x, θ_1)
Residual_2 (x, θ_2)
Residual_3 (x, θ_3)
Residual_k-1(x,θ_k-1)
Residual_k (x, θ_k)
Σ

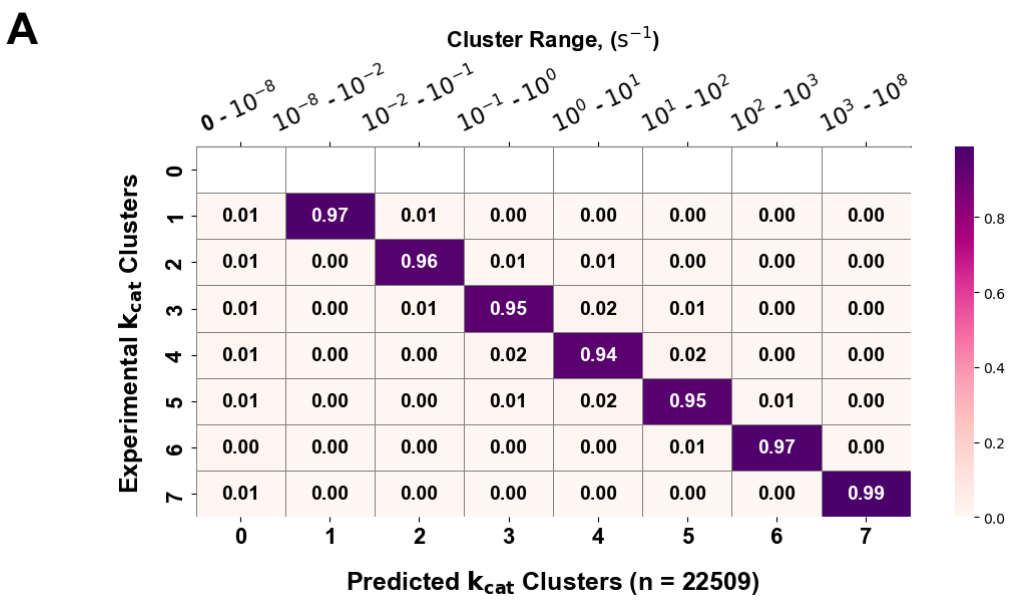

A
Performances across model types
Testing
Validation
Training
Transformer
Bi-LSTM
MLP
XGBoost
Decision Tree
Random Forest
0
0.2
0.4
0.6
0.8
1
Accuracy
B
Data Split
Train
Validation
Test
Accuracy (%)
99.99
88.72
89.85
98.43
84.56
85.08
k_{cat}
K_M
Kinetic Parameters (k_{cat} and K_M)
C
Data Split
Train
Validation
Test
e-Accuracy (within ±1 Log Unit (%))
99.99
94.72
95.23
99.71
96.68
97.16
k_{cat}
K_M
Kinetic Parameters (k_{cat} and K_M)
D
Data Split
Train
Validation
Test
Recall (%)
99.99
88.72
89.85
98.43
84.56
85.08
k_{cat}
K_M
Kinetic Parameters (k_{cat} and K_M)
E
Data Split
Train
Validation
Test
F1-Score (%)
99.99
88.64
89.80
98.43
84.41
84.96
k_{cat}
K_M
Kinetic Parameters (k_{cat} and K_M)
F
True k_{cat} Clusters
Predicted k_{cat} Clusters
Classes
Class 0
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
t-SNE 1
t-SNE 2
G
Cluster Range, (s^{-1})
0 - 10^{-8}
10^{-8} - 10^{-2}
10^{-2} - 10^{-1}
10^{-1} - 10^{0}
10^{0} - 10^{1}
10^{1} - 10^{2}
10^{2} - 10^{3}
10^{3} - 10^{8}
Experimental k_{cat} Clusters
0 0.86 0.01 0.02 0.03 0.03 0.04 0.01 0.00
1 0.00 0.98 0.01 0.01 0.00 0.00 0.00 0.00
2 0.00 0.03 0.92 0.03 0.01 0.01 0.00 0.00
3 0.02 0.01 0.04 0.85 0.05 0.02 0.01 0.00
4 0.02 0.00 0.02 0.06 0.82 0.06 0.01 0.00
5 0.02 0.00 0.01 0.02 0.06 0.83 0.04 0.01
6 0.00 0.00 0.00 0.01 0.01 0.02 0.94 0.01
7 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00
Predicted k_{cat} Clusters (n = 13216)
H
Cluster Range, (M)
10^{-10} - 10^{-5}
10^{-5} - 10^{-4}
10^{-4} - 10^{-3}
10^{-3} - 10^{-2}
10^{-2} - 10^{-1}
10^{-1} - 10^{4}
Experimental K_M Clusters
0 0.90 0.06 0.02 0.01 0.00 0.00
1 0.10 0.76 0.12 0.02 0.01 0.00
2 0.02 0.14 0.71 0.11 0.01 0.00
3 0.01 0.03 0.09 0.80 0.06 0.01
4 0.00 0.01 0.01 0.03 0.94 0.02
5 0.00 0.00 0.00 0.00 0.01 0.99
Predicted K_M Clusters (n = 6861)
I
True K_M Clusters
Predicted K_M Clusters
Classes
Class 0
Class 1
Class 2
Class 3
Class 4
Class 5
t-SNE 1
t-SNE 2

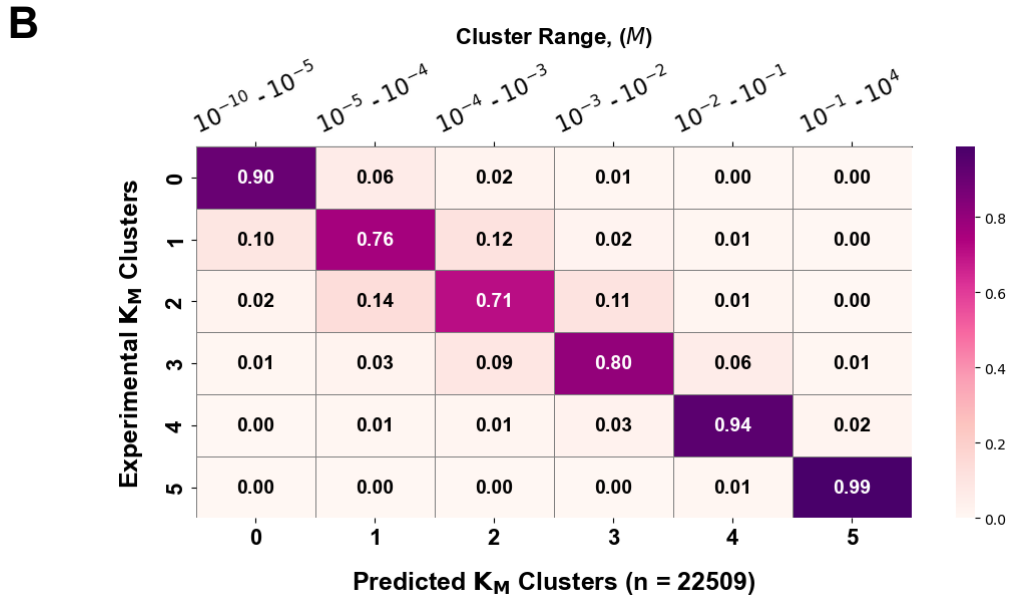



BRENDA
Expasy
SABIO-RK
237,487 entries
kcat (67%) KM (33%)
8.232 entries
EC numbers
77,663 entries
kcat (33%) KM (67%)
Remove duplicates
Keep max kcat and min KM for duplicates
96,782 entries
kcat (41%) KM (59%)
54,906 entries
kcat (37%) KM (63%)
Merge, get substrate SMILES,
WT and mutant enzyme sequences
Remove duplicate reactions
30,442 entries
kcat
44,615 entries
KM
26,244 entries with both kcat and KM
Raw UNVERIFIED Data
Manual validation 2,158 articles
MANUAL
>10k verified mutant data included
932 extra data (net change)
Catalytic reisdue(s) identification from enzyme sequence
M A S Q
R C D S
kcat = Y
kcat = 0
Original seq
Catalytic res. to Ala
Adding negative data
0
17k negative data points added
RealKcat
Training set built with real kinetic parameters
KinHub-27k
Available: chowdhurylab.github.io/downloads.html


Connectedness of Each Amino Acid in PDB Structure
Number of Connected Residues
Residue Number
Residue 79
(n=18)
Residue 164
(n=13)
0
5
10
15
20
0
100
200
300
400
500




A
Cluster Range, (s^{-1})
$0 - 10^{-8}$
$10^{-8} - 10^{-2}$
$10^{-2} - 10^{-1}$
$10^{-1} - 10^{0}$
$10^{0} - 10^{1}$
$10^{1} - 10^{2}$
$10^{2} - 10^{3}$
$10^{3} - 10^{8}$
Experimental k_{cat} Clusters
0 0.01 0.97 0.01 0.00 0.00 0.00 0.00 0.00
1 0.01 0.00 0.96 0.01 0.01 0.00 0.00 0.00
2 0.01 0.00 0.01 0.95 0.02 0.01 0.00 0.00
3 0.01 0.00 0.00 0.02 0.94 0.02 0.00 0.00
4 0.01 0.00 0.00 0.01 0.02 0.95 0.01 0.00
5 0.00 0.00 0.00 0.00 0.00 0.01 0.97 0.00
6 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.99
7
0 1 2 3 4 5 6 7
Predicted k_{cat} Clusters (n = 22509)
0.8
0.6
0.4
0.2
0.0


B
Cluster Range, (M)
$10^{-10} - 10^{-5}$
$10^{-5} - 10^{-4}$
$10^{-4} - 10^{-3}$
$10^{-3} - 10^{-2}$
$10^{-2} - 10^{-1}$
$10^{-1} - 10^{4}$
Experimental K_M Clusters
0 0.90 0.06 0.02 0.01 0.00 0.00
1 0.10 0.76 0.12 0.02 0.01 0.00
2 0.02 0.14 0.71 0.11 0.01 0.00
3 0.01 0.03 0.09 0.80 0.06 0.01
4 0.00 0.01 0.01 0.03 0.94 0.02
5 0.00 0.00 0.00 0.00 0.01 0.99
0 1 2 3 4 5
Predicted K_M Clusters (n = 22509)
0.8
0.6
0.4
0.2
0.0