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], Randy Aryee[1,2], 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

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

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

Summary

Predicting enzyme kinetics directly from sequence remains a central challenge in computational biology, particularly in resolving the effects of mutations at catalytically essential residues. Existing models frequently overlook the functional consequences of such perturbations, often defaulting to wild-type predictions even in cases of substantial activity loss, thereby limiting their reliability for enzyme design and mechanistic inference. Here, we introduce RealKcat, a machine learning framework trained on KinHub-27k, a rigorously curated dataset of 27,176 experimentally reported enzyme–substrate entries consolidated from BRENDA, SABIO-RK, and UniProt and verified across 2,158 primary sources. To ensure biochemical realism, kinetic parameters were collapsed into order-of-magnitude bins, enabling predictions that are tolerant to experimental noise yet sensitive to functional shifts. RealKcat integrates ESM embeddings for enzyme sequences with ChemBERTa embeddings of affiliated substrate, producing a unified feature space of the chemical conversion that supports robust multi-class classification of both catalytic turnover (k_{cat}) and substrate affinity (K_M). Across cross-validation, hold-out, out-of-distribution, and few-shot evaluations—including a dense mutational landscape of alkaline phosphatase (PafA)—RealKcat consistently capturead the direction and magnitude of mutation-induced changes, while preserving discrimination in both wild-type and mutant contexts. Importantly, structural descriptors were deliberately excluded, as naive integration of structural features has been shown to impair model generalization, underscoring the primacy of rigorous dataset curation, biologically informed task formulation, and balanced evaluation metrics. RealKcat establishes a scalable and mutation-sensitive framework for enzyme kinetics prediction, offering a biologically grounded platform for enzyme engineering, metabolic modeling, and therapeutic design.

Significance Statement

Enzymes catalyze biochemical reactions that sustain life, and accurate measurement of their efficiency—expressed through turnover number (k_{cat}) and substrate affinity (K_M)—is fundamental to biotechnology, synthetic biology, and even pharmaceutical innovation. Yet experimental assays remain prohibitive, time-intensive, and sensitive to conditions such as pH, temperature, and ionic strength of the assay buffer, while existing computational approaches often lack sensitivity to catalytic-site mutations and are constrained by inconsistencies in public databases. RealKcat addresses these gaps by introducing a rigorously curated dataset (KinHub-27k) derived from manual review of 2,158 articles and augmented with 5,278 synthetic catalytic variants generated through alanine substitution at annotated catalytic residues. Leveraging protein and substrate embeddings and a classification scheme based on order-of-magnitude kinetic bins, RealKcat achieves state-of-the-art functional e-accuracy and, critically, demonstrates sensitivity to catalytic perturbations. By adopting e-accuracy—a performance metric that evaluates predictions within ±1 order of magnitude, aligning with the practical utility of enzyme kinetics—RealKcat provides biologically meaningful assessments that conventional metrics often obscure. This work establishes a robust, mutation-aware predictive platform that advances computational enzyme design and extends applicability to biomanufacturing, metabolic engineering, and precision medicine.

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 mapping of metabolic landscapes in non-model species, engineering of enzymes for specific substrates, and optimization of biochemical pathways, thereby reducing reliance on costly experimental screening. Such predictive capability is particularly valuable for pathway optimization—whether for sustainable metabolite production or for dissecting disease-linked metabolic shifts—where trial-and-error experimentation is prohibitive and does not reveal mechanistic underpinnings. Because functionally diverse enzymes are frequently linked to sequence variation, including engineered mutations and naturally occurring polymorphisms, robust and mutation-sensitive models are essential both for industrial biocatalysis and for identifying disease prognosis markers (*3, 4*).

Current computational approaches, however, often lack the sensitivity to detect functionally relevant alterations at catalytically crucial residues (*5*). Compounding this challenge, most reported kinetic measurements are generated under *in vitro* conditions that do not capture the dynamic biochemical environments of cells, leading to predictions that can diverge from *in vivo* behavior. This gap limits the accurate modeling of metabolic shifts, drug responses, and synthetic pathway performance, particularly in non-model organisms and complex human systems. With growing demand for precision enzyme design in biotechnology, metabolic engineering, and biomedicine, there is a pressing need for machine learning (ML) frameworks that offer mutation-aware, catalytically sensitive predictions built on rigorously curated datasets.

Recent advances in ML have produced promising enzyme kinetics predictors, including DLKcat, TurNuP, UniKP, CatPred, and EITLEM-Kinetics, each demonstrating the utility of data-driven approaches in modeling enzyme–substrate kinetics (*6–10*). DLKcat applies convolutional and graph neural networks to predict turnover numbers (k_{cat}) across diverse pairs, though its performance depends heavily on dataset diversity. TurNuP extends this by employing ESM-1b sequence encodings (*7, 11*) with RDKit-derived reaction fingerprints in a gradient-boosted algorithm, improving generalizability for poorly represented enzymes. UniKP incorporates environmental descriptors such as pH and temperature into a two-layer model that encodes sequences and substrates (*8*), though its accuracy is constrained by training data quality. CatPred

leverages neural networks trained on SABIO-RK and BRENDA (*12, 13*), achieving state-of-the-art accuracy with approximately 76% of k_{cat} predictions and 84% of K_M predictions within one order of magnitude (*10*). EITLEM-Kinetics advances further by employing an ensemble iterative transfer learning approach to predict kinetic parameters for mutants with low sequence similarity (*14*). These models employ random splitting schemes to generate training, validation, and test sets; TurNuP additionally excluded overlapping enzyme sequences across splits, while CatPred enforced the stricter condition that no enzyme–substrate pair was duplicated across partitions.

While these tools mark significant progress, they exhibit critical limitations in mutation sensitivity. In particular, alanine substitutions at catalytic residues yield near-identical predictions across models, reflecting reliance on global sequence similarity rather than explicit catalytic awareness. Moreover, CatPred's concatenation of substrate and cofactor SMILES into a composite reaction fingerprint overlooks the independent effects of distinct ligands, constraining interpretability. For enzyme engineering, mutation sensitivity and dataset quality are paramount: accurate prediction of catalytic consequences requires not only sophisticated ML frameworks but also carefully curated training resources with consistent treatment of both wild-type and mutant variants. To address these challenges, we present RealKcat, a framework that integrates rigorous curation with optimized gradient-boosted classifiers to deliver robust, mutation-sensitive predictions of enzyme kinetics. RealKcat is trained on the KinHub-27k dataset, comprising 27,176 experimentally validated entries from BRENDA and SABIO-RK and augmented with 5,278 synthetic inactive sequences generated by alanine substitution at UniProt-annotated catalytic residues (*15, 16*). Dataset inconsistencies (n = 1,804) were resolved through direct review of 2,158 primary sources, ensuring fidelity of kinetic parameters, sequences, and substrate assignments. By combining ESM sequence embeddings, which capture evolutionary and biochemical context (*17*), with ChemBERTa substrate embeddings derived from isomeric SMILES (*18*), RealKcat constructs a rich feature representation that enhances functionally relevant predictive accuracy and establishes a robust foundation for catalytically sensitive enzyme modeling.

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 (*19*)) allowing for differences *in vitro* to *in vivo* kinetic rates, to capture experimental phenotypes with fidelity (*20–22*).

RealKcat. is a mutation-aware, classification-based framework for enzyme kinetics prediction that integrates rigorous data curation, order-of-magnitude clustering, and mechanistically guided negatives with advanced protein and substrate embeddings. We demonstrate that RealKcat achieves robust generalization across diverse evaluation regimes, including wild-type and mutant test sets, out-of-distribution catalytic variants, and a high-throughput mutational landscape of alkaline phosphatase (PafA). By emphasizing e-accuracy as a biologically grounded metric alongside balanced statistical measures such as Matthews Correlation Coefficient (MCC), RealKcat delivers predictions that are both functionally interpretable and statistically conservative. Beyond benchmarking, we demonstrate that RealKcat is capable of capturing mutation-induced kinetic shifts at catalytically relevant residues, highlighting its potential utility for enzyme design, metabolic engineering, and biomedical applications where mutation sensitivity is essential.

Results

Building a Curated Benchmark for Enzyme Kinetics

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). For enzymatic reactions involving multiple substrates, we included separate entries for each substrate, effectively allowing the same enzyme to appear more than once in the dataset. This design acknowledges the standard assumption underlying many reported *in vitro* characterizations—that assays are performed under saturating concentrations of co-substrates, such that the measured kinetic parameter reflects dependence on a single varied substrate (*23*). By separating substrates explicitly, the model can learn substrate-specific contributions to catalytic parameters without conflating effects from experimental co-substrate conditions. The curated KinHub-27k dataset

and accompanying code are available for download and inference (https://chowdhurylab.github.io/downloads.html).

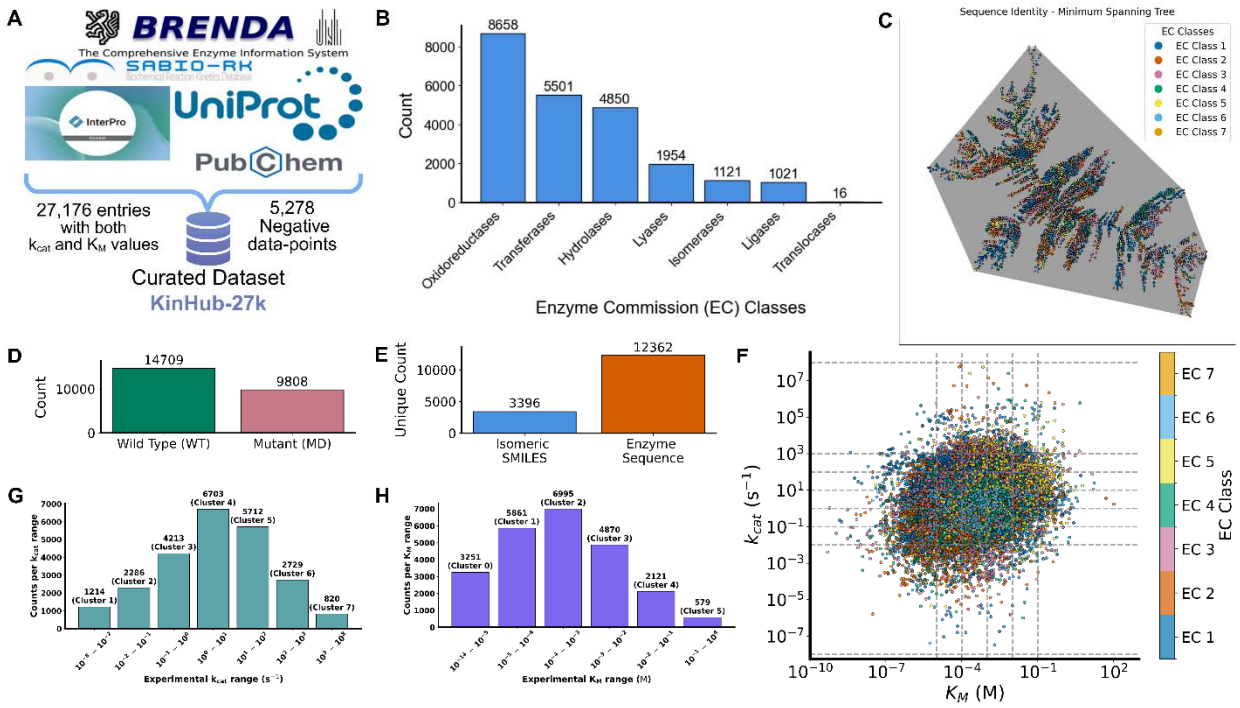


Fig. 1. Comprehensive dataset curation, classification, and distribution for RealKcat model training. (A) Raw data for KinHub-27k was collated from multiple databases, including BRENDA, SABIO-RK, UniProt, and PubChem, resulting in a curated dataset with 27,176 entries containing both and K_M values, and an additional 5278 synthetic inactive sequences **(B)** Distribution of enzyme entries across Enzyme Commission (EC) classes, highlighting a broad representation of enzymatic functions. **(C)** Network graph based on pairwise sequence identity, illustrating clustering by EC class and the diversity of sequence space represented in the dataset. **(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 values, depicting the core dynamic range of kinetic parameters. **(G & H)** Distribution of kinetic parameter clusters for and K_M, respectively, categorized by order of magnitude to facilitate robust multiclass classification.

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 3,396 unique SMILES, 12,362 sequences, and spans 1,499 unique organisms (Fig. 1E). Sequences longer than 1024 characters were removed, and duplicate entries were consolidated by grouping identical sequence and isomeric SMILES pairs. Rather than retaining the extreme (max–min) kinetic values—a common practice in prior models that can over-emphasize atypical

or poorly replicated measurements—we observed in our own preliminary processing that this approach could produce spurious variability, with several unique sequence–substrate pairs differing by as much as 6–8 orders of magnitude (Supplemental Fig. S1). To mitigate this, duplicates were collapsed by taking the median k_{cat} and median K_M, yielding robust central estimates that preserve the underlying biological signal while reducing the influence of outliers. This refinement produced a dataset of 24,517 unique entries with 14,709 wild-type entries and about 9808 mutant entries (Fig. 1D). Figure 1B shows the distribution of enzyme commission (EC) numbers, while Fig. 1C visualizes sequence identity–based relationships across enzymes. The network was constructed from pairwise global alignments, with distances defined as 1−identity, and simplified using a minimum spanning tree. Clustering by EC class highlights both conserved families and the broader diversity of sequence space encompassed in KinHub-27k. In order to strengthen the exposure of RealKcat functionally impaired sequences, we added 5278 mechanistically guided synthetic negative set by substituting UniProt-annotated "active-site" residues with alanine. In addition, to ensure robust target representations, duplicate sequence–substrate pairs were collapsed by median pooling prior to partitioning, a refinement that set the stage for reformulating kinetic prediction as a classification problem. Having established a curated dataset, we next reformulated kinetic prediction as a classification problem to better reflect biological interpretation.

Log-Scaled Binning Towards Functional Catalytic Regimes

To address the intrinsic exponential variability of enzyme kinetic parameters, k_{cat} and K_M were reformulated from continuous regression targets into biologically meaningful, $\log_{10}$-spaced bins, transforming the task into one of discriminating parameter ranges rather than predicting exact point estimates (Fig. 1F–H), with additional clusters for extreme values (see Materials and Methods for details). This design reflects how kinetic data are interpreted in practice—where order-of-magnitude resolution is often sufficient for functional inference—and imposes boundaries that separate distinct catalytic and binding regimes while reserving additional clusters for extreme values. By aligning prediction labels with magnitude intervals, the model is trained to focus on functionally relevant scale differences, thereby dampening the influence of experimental noise and assay-specific variability that persist even after median pooling of duplicate sequence–substrate pairs.

Moreover, the resulting formulation also enables the use of e-accuracy, a biologically grounded metric that rewards predictions within one bin of the measured value, acknowledging that slight deviations in magnitude rarely alter biochemical interpretation (*14, 24*). This is particularly critical given that, even after rigorous curation, many enzyme–substrate pairs exhibit spreads of 6–8 orders of magnitude (Supplemental Fig. S1), driven by differences in organism source, assay protocols, and measurement conditions. In this framework, binning and e-accuracy operate synergistically to produce a stable, interpretable, and functionally relevant classification setting, strengthening RealKcat's robustness for enzyme design and metabolic modeling applications.

Synthetic Negatives Anchor Catalytic Loss Detection

To strengthen RealKcat's ability to distinguish catalytically relevant changes, we expanded the dataset with a mechanistically guided synthetic negative set. This was generated by systematically substituting UniProt-annotated catalytic "active-sites" residues with alanine (see *Materials and Methods*), a perturbation most often expected to impair catalytic activity. This procedure yielded 5,278 high-confidence synthetic negatives, anchored by 22 experimentally curated WT–MD_{Ala} pairs that consistently exhibited substantial catalytic impairment, (Supplemental Table 1)—typically ≥2 orders-of-magnitude reductions in k_{cat} and ≥1 order increases in K_M for active site substitutions—without recourse to arbitrary zero-assignment policies. These curated examples provided benchmarks for calibrating synthetic negatives against realistic biochemical effects.

To further refine the set, we trained a binary meta-label classifier to discriminate between alanine substitutions that caused substantial catalytic loss and those that retained measurable activity. These curated WT–MD_{Ala} pairs served as the foundation for training on a mechanistically descriptor set, which included mutational severity (aggregate BLOSUM62 costs for alanine replacement), substrate physicochemical properties (molecular weight, logP, ring count, hydrogen-bond donor/acceptor counts), interaction-scale metrics (number of known active-site and substrate-to-enzyme molecular weight ratio), and the wild-type kinetic bin as a contextual reference. Under leave-one-out cross-validation, the classifier achieved accuracies of 77.27% for k_{cat} drop prediction (confusion: [5, 3; 2, 12]) and 81.82% for K_M increase prediction (confusion: [4, 2; 2, 14]), providing sufficient discrimination policy to guide relabeling of synthetic negatives. Applied across the alanine-scanned dataset, this strategy produced 5278

high-confidence negatives. Importantly, these negatives remained sequence-close to their wild-type counterparts but were reassigned to kinetically divergent bins, creating hard counter examples that increased the discriminative challenge during model training. By integrating these negatives, we impose on RealKcat to learn sensitivity to catalytic relevance, with expected robustness in separating functional from non-functional variants during downstream evaluation.

Minimizing Data Leakage Through Structured Partitioning

To provide an independent evaluation set beyond the cross-validation training and evaluation, we first established a randomized catalytic inference hold-out comprising 100 positive entries proportionally sampled from the wild-type (WT) and mutant-derived (MD) pools under a fixed random seed, thereby preserving the endogenous WT:MD ratio while ensuring independence from the training partitions. In parallel, 10 synthetic negatives were withheld from the negative test pool; these were generated by alanine substitution at annotated catalytic residues. Although their cognate wild-type sequences were allowed in the positive training partitions downstream, the corresponding negative sequences were never exposed during training, positioning them as in-distribution with respect to enzyme identity yet unseen as explicit negatives. This two-tier design enabled evaluation of catalytic inference across both biological variants and synthetic chemotypes, complementing the cross-validation partitions.

Beyond this hold-out, a careful manifold-guided clustering of the negative set yielded Affinity Propagation clusters, including a small “miscellaneous” bin, ensuring broad structural coverage while preventing over-representation of densely sampled regions (see Materials and Methods). This resulted in curated positives comprised 14,161 wild-type and 9,372 mutant entries after the randomized hold-out (based on median-pooled targets), forming the basis for subsequent modeling. Synthetic negatives were consistently paired with their cognate wild-type enzymes, and a 22-sample “anchor” synthetic set was withheld for drop-policy labeling of unlabeled synthetic negatives and for benchmarking. To preserve out-of-distribution integrity, sequence similarity between the anchor set and all other subsets (positives and negatives) was limited to ≤60% global identity. Across five folds, composite-label stratification preserved balanced kinetic class and WT:MD proportions, ensuring consistent distributions. Final train/validation/test proportions were tightly controlled (~80/10/10 for CV folds and ~95/2.5/2.5 for the negative set), supporting robust, leakage-minimized, and reproducible model evaluation.

Sequence–Substrate Embeddings Capture Biochemical Context

With the inputs and target labels defined, we next compared two enzyme representation schemes. Specifically, we evaluated enzyme representations derived from two recent protein language models: ESM-2 and ESM-C. Both models, trained on large-scale sequence corpora, generate residue-level contextualized vectors that capture evolutionary and biochemical constraints across protein families (*11, 17, 25*). In this study, residue embeddings were aggregated into fixed-length vectors to serve as enzyme descriptors, a step that improves computational tractability for large-scale. For substrates, we used ChemBERTa embeddings (768 dimensions), derived from isomeric SMILES strings, which capture information related to stereochemistry, functional groups, and other physicochemical properties encoded in molecular structure (*18*). Enzyme and substrate embeddings were concatenated to form a 1,920-dimensional joint representation (1,920 dimensions for ESM-C as in Fig. 2A, 2,048 for ESM-2), integrating sequence context and substrate chemistry within a unified feature space. Although such concatenation does not explicitly model enzyme–substrate interactions, it provides a joint input space from which the classifier can infer associations between catalytic variation and molecular context. In addition, across random test and hold-out splits, ESM-2 and ESM-C performed comparably (e-accuracy, AUC-PR), whereas the main separation appeared in MCC on the out-of-distribution “anchor” set (n=22) and on a high-throughput alkaline phosphatase dataset (PafA) few-shot test, where ESM-C was higher (Fig. 3A). Given our emphasis on generalization under distribution shift and few-shot mutation testing, we report ESM-C as the default representation, while noting that ESM-2 yielded comparable performance across most splits.

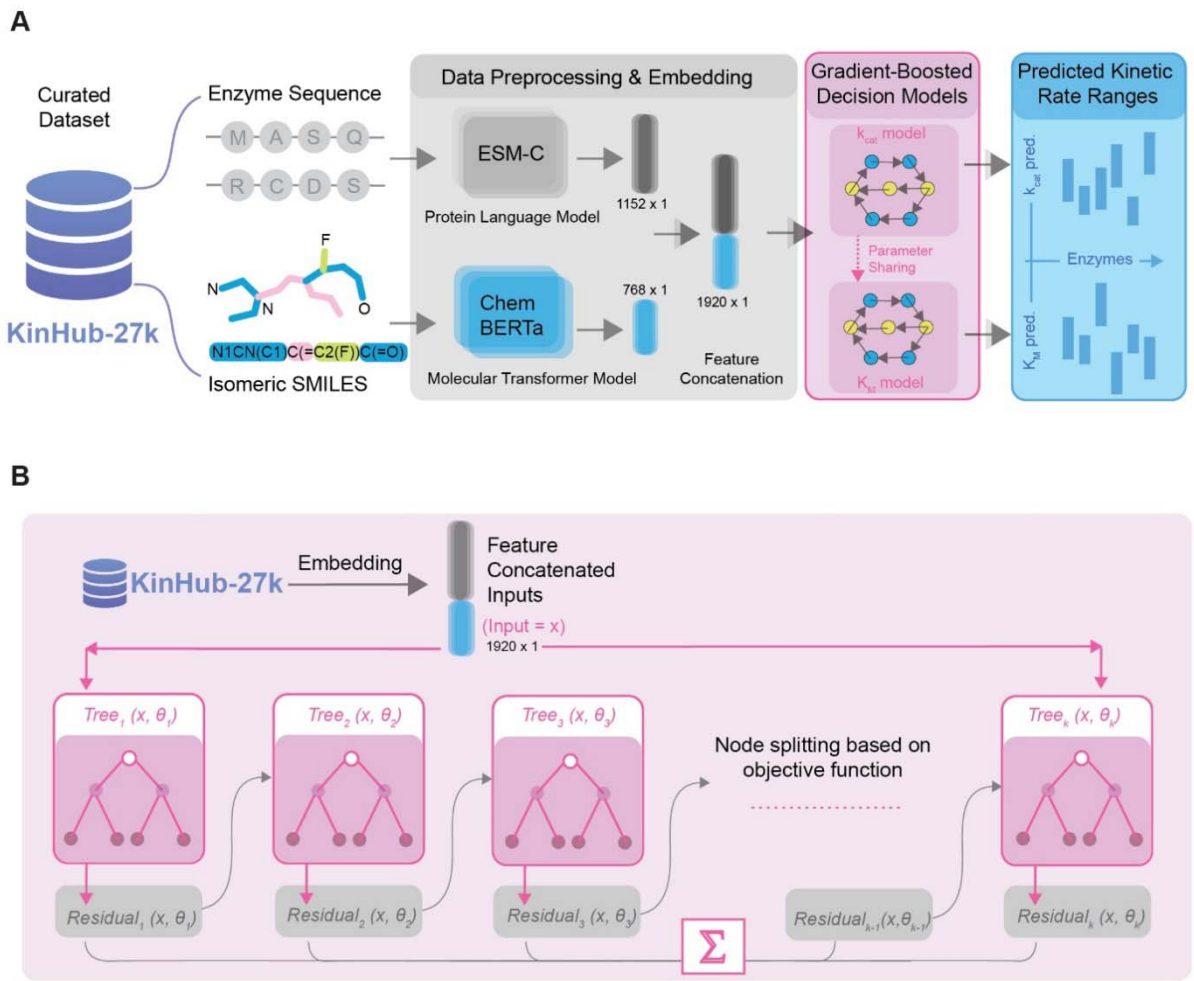


Fig. 2. RealKcat model workflow and architecture for enzyme kinetics prediction. (A) The KinHub-27k dataset provides enzyme sequence and substrate information, embedded with ESM-C (1152-dimensional) and ChemBERTa (768-dimensional) models, respectively. The combined 1920-dimensional feature vectors are 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

A Classification Framework for Enzyme Kinetics Prediction

RealKcat framework was trained using gradient-boosted additive modeling (XGBoost) to classify enzyme kinetic parameters into biologically meaningful $\log_{10}$-spaced bins (**Fig. 2B**). We emphasize from the outset that our objective was not to exhaustively search for an "optimal" algorithm but to demonstrate that, given a rigorously curated dataset and biologically grounded formulation, robust predictive performance can be achieved with a stable, interpretable, and

reproducible classifier. Gradient boosting was chosen for its ability to model complex non-linear dependencies, its resilience to overfitting across heterogeneous feature spaces, and its transparent decision structure (*26, 27*). This choice reflects a deliberate focus on robustness and reproducibility rather than marginal improvements in score maximization.

Model performance is summarized in **Fig. 3A–C**, which reports e-accuracy (within ±1 log unit) and macro-averaged MCC, a balanced classification metric that incorporates all true/false positives and negatives, across test, hold-out, out-of-distribution (OOD), and few-shot evaluations. On the test set, RealKcat achieved 87.10% e-accuracy for k_{cat} and 91.72% e-accuracy for K_M (Fig. 3B), with corresponding MCC values of 0.53 for k_{cat} and 0.50 for K_M (Fig. 3C). These metrics remained consistent across validation and hold-out partitions, underscoring that the model moderately generalizes beyond its training distribution. Meanwhile, to address the strong class imbalance across kinetic bins, we applied SMOTE oversampling (*28*) exclusively to the training partition of each cross-validation fold, while leaving validation and test sets untouched to ensure fair generalization. This balancing step prevented majority-class dominance, stabilized performance in minority clusters, and directly contributed to the consistently high macro-MCC observed in training (Fig. 3C). Moreover, K_M predictions exhibit consistently higher e-accuracy than k_{cat} (91.72% vs 87.10% on test set), reflecting the fact that substrate affinities cluster within narrower log10-ranges (Fig. 1F), whereas turnover numbers span broader, assay-sensitive dynamic ranges (Supplemental Fig. S1). This highlights the inherent difficulty of predicting catalytic rates and underscores that RealKcat conforms to the established hierarchy of predictability, with K_M values being inherently more tractable than $k_{cat.}$

While e-accuracy captures the biological and environmental variational reality that order-of-magnitude predictions suffice for most functional inferences, MCC provides a balanced statistical measure that accounts for all entries in the confusion matrix and prevents dominance by majority classes. This is especially critical given the natural skew of enzyme kinetics data toward central bins (Fig. 1G-H). Exact-bin accuracy, F1-scores, recall, and area under the precision–recall curve (AUC-PR, which is more informative than ROC curves under class imbalance), are reported in **Supplemental Fig. S2**, where lower values illustrate the limitations of conventional metrics under class imbalance. In contrast, the stability of e-accuracy and MCC across splits (Fig. 3A–C) highlights why these measures are more conservative and robust indicators of predictive quality.

Prediction under distribution shift was further assessed on an OOD anchor set of 22 negatives and a few-shot alkaline phosphatase (PafA) dataset. In these cases, MCC slightly distinguished ESM-C from ESM-2, with modest but consistent improvements for ESM-C (Fig. 3A). Although performance across other metrics was comparable, these results emphasize the relevance of embedding choice for prediction in sparse or low-data settings, where robustness rather than average accuracy is most critical. Generalization across homology boundaries also supports this interpretation: stratification of predictions by sequence identity (Fig. 3D–E) shows that e-accuracy remain fairly stable down to 60% identity relative to training, with only modest declines below 40%. This indicates that the embedding framework captures catalytic information beyond sequence similarity, enabling reliable extrapolation to enzymes not represented in the training set.

Resolution at the class level is evident from the confusion matrices for k_{cat} and K_M (**Fig. 3F–G**). Predictions are concentrated along the diagonal, with strongest agreement in the central bins but crucially, even in sparsely populated extremes, most predictions fall within ±1 log unit of the experimental cluster. This validates the use of e-accuracy as a biologically relevant measure, as near-miss predictions at the order-of-magnitude level still preserve functional interpretation. The t-SNE projections (Fig. 3H–I) provide further confirmation, with predicted classes reproducing the global structure of experimental clusters in reduced space, demonstrating that RealKcat has internalized meaningful distinctions across kinetic regimes rather than collapsing toward majority classes. For deployment, we report results from fold-5, pre-specified as the inference model, while population-level reproducibility is summarized as cross-fold mean ± SD.

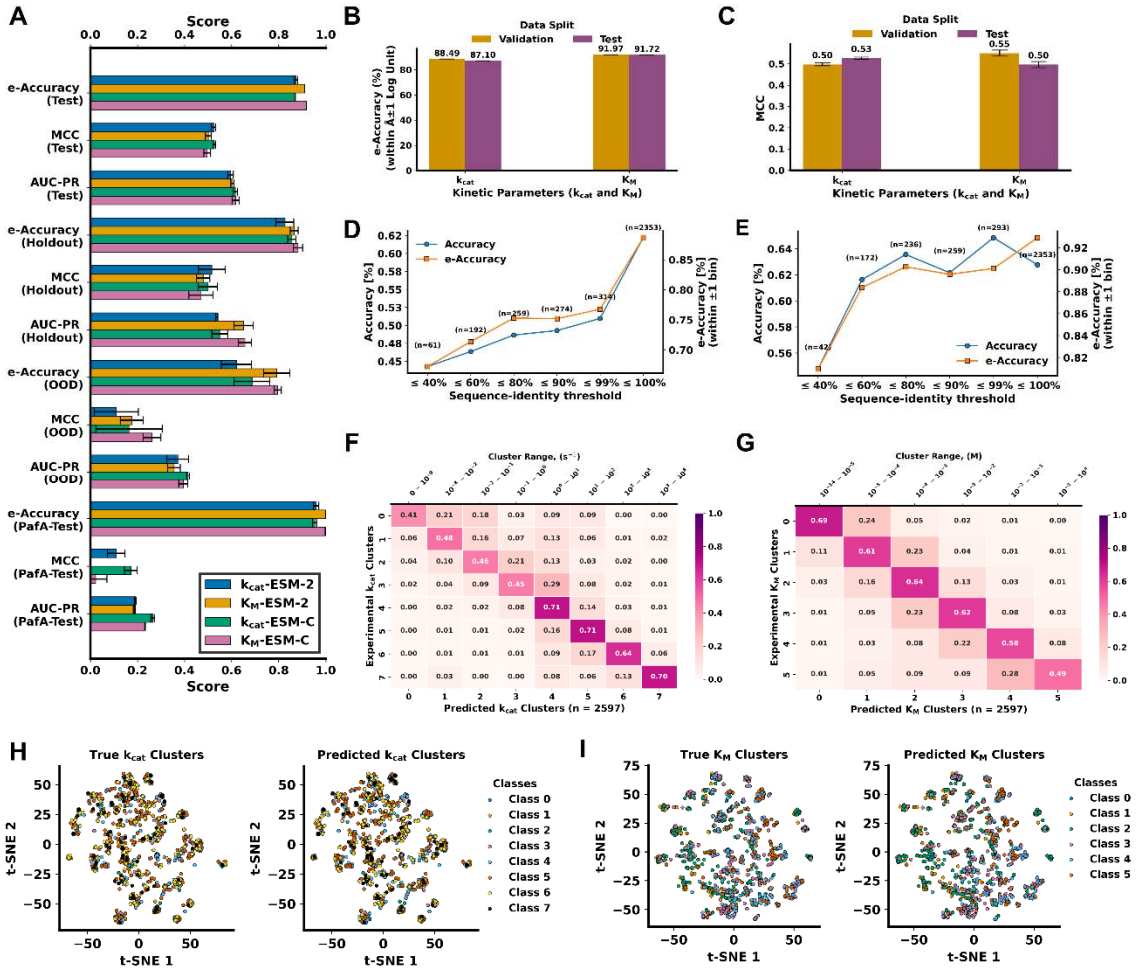


Fig. 3. Evaluation of RealKcat performance across kinetic parameters and data splits. (A) Comparison of model performance using ESM-2 versus ESM-C embeddings for and K_M across test, hold-out, out-of-distribution (OOD), and PafA few-shot splits, reported as e-accuracy, Matthews correlation coefficient (MCC), and area under the precision–recall curve (AUC-PR). **(B–C)** Performance consistency with ESM-C embeddings across validation and test sets: **(B)** e-accuracy (within ±1 log unit) and **(C)** MCC, demonstrating stable predictive capacity for both kinetic parameters. Panels A–C report the mean across all five folds ± standard deviation. **(D–E)** Generalization with ESM-C across sequence-identity thresholds for **(D)** and K_M **(E)**. Accuracy and e-accuracy are plotted against sequence similarity to the training set, showing reliable extrapolation to low-identity regions. **(F–G)** Confusion matrices for **(F)** and **(G)** K_M predictions, indicating strong recovery of experimental clusters with highest accuracy in central kinetic ranges and lower misclassification at extremes. **(H&I)** t-SNE visualizations of true vs. predicted clusters for and K_M, showcasing high-dimensional separation and clustering.

Mutation Sensitivity Across Catalytic Residues

RealKcat demonstrated sensitivity to point mutations at catalytically relevant residues, capturing the expected shifts in kinetic parameters when active-site positions are perturbed. To

evaluate mutation awareness under controlled but distinct conditions, we adopted a two-tier scheme. First, we assessed performance on the established out-of-distribution (OOD) anchor set comprising 22 experimentally characterized wild-type (WT) and alanine variants at UniProt-annotated catalytic residues that were withheld from training. Second, we evaluated a negative hold-out consisting of 10 synthetic alanine variants drawn at random from the partitioned negative test pool; their cognate WT enzymes appeared in the positive training partitions, but the exact negative sequences were never used for training. This design probes two distinct generalization regimes: (i) an OOD setting in which both enzyme–substrate pairs and mutation outcomes are unseen, and (ii) an in-distribution-by-identity setting in which the enzyme family is familiar but the negative examples are novel, enabling an appraisal of mutation sensitivity independent of trivial sequence memorization.

Across the 22 OOD WT–MD_{Ala} pairs (Supplemental Table 1), RealKcat reproduced both the directionality and order-of-magnitude shifts in catalytic impairment for k_{cat} and K_M, with the majority of mutant predictions falling within ±1 bin of the experimental values (Fig. 4A–B). These pairwise comparisons show that predicted WT→MD trajectories align with experimentally observed changes, particularly for direct-contact substitutions at annotated catalytic residues. Longitudinal visualizations (Fig. S3) further illustrate this behavior, with predicted trajectories closely tracking experimental curves and remaining largely within the ±1-bin tolerance bands for both parameters. A critical observation, however, is that misclassified mutants tend to be predicted near the corresponding wild-type bin, indicating a fallback toward baseline activity rather than spurious overestimation of catalytic loss. This pattern highlights both the conservative nature of the model and the difficulty of extrapolating large mutational effects in an out-of-distribution regime. Together, these analyses indicate that RealKcat captures mutation-induced movement across kinetic bins for catalytic residues in a setting that is intentionally distribution-shifted from training.

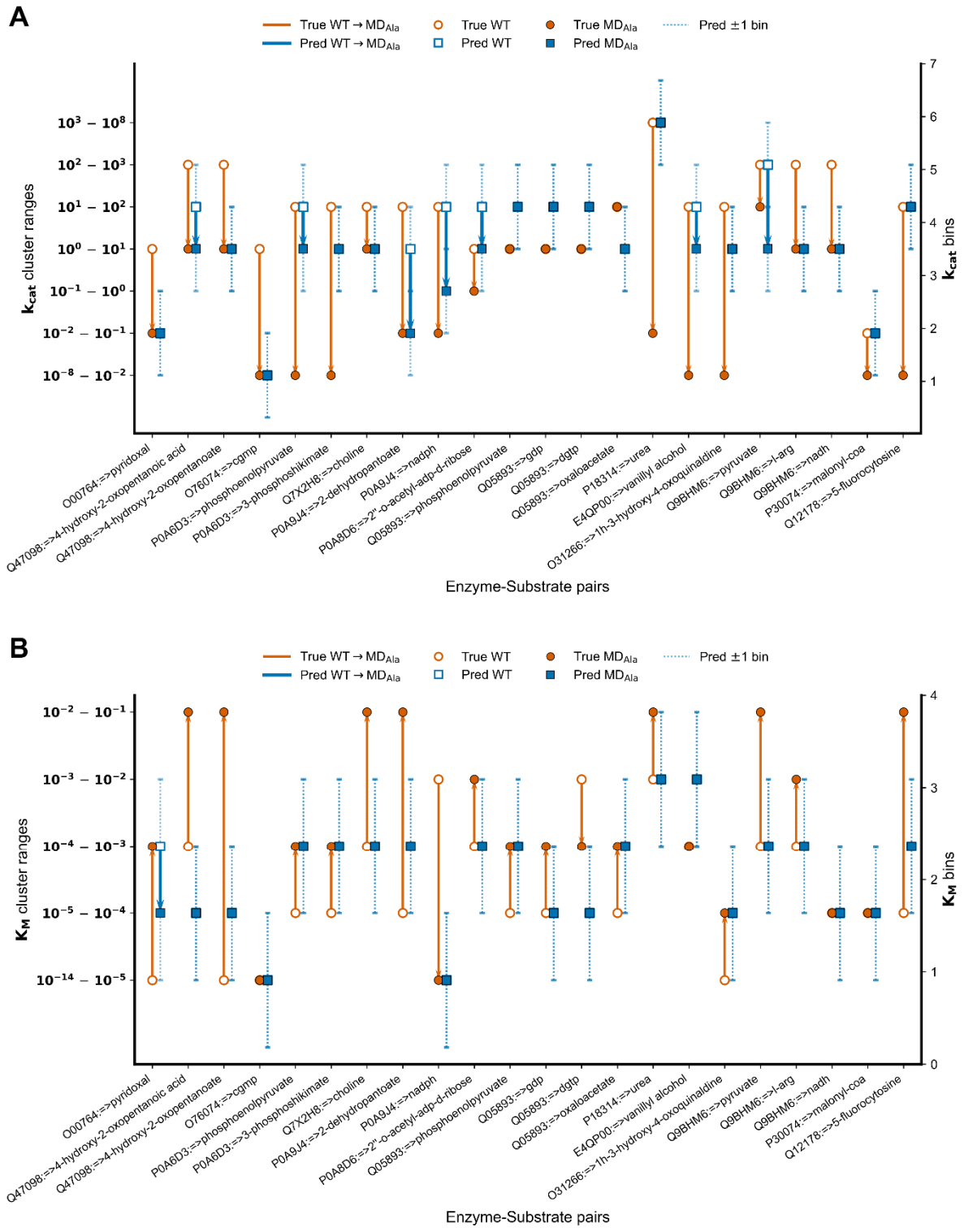


Fig. 4. Out-of-distribution anchor evaluation on experimentally validated WT–MD$_{Ala}$ pairs at catalytic residues. (A) Predictions of clusters for 22 WT–MD$_{Ala}$ pairs curated from UniProt-annotated catalytic residues. Orange markers indicate experimentally measured wild-

type (WT, open circle) and mutant (MD_{Ala}, filled circle) values; blue markers denote RealKcat predictions for WT (open square) and MDAla (filled square). Lines connect WT–MD_{Ala} pairs to illustrate mutation-induced shifts. Dotted vertical bars indicate the ±1 bin tolerance window used in e-accuracy assessment. Predicted shifts largely align with experimental measurements, with most mutant values falling within ±1 order of magnitude of the experimental cluster. **(B)** Equivalent analysis for K_M clusters across the same 22 WT–MD_{Ala} pairs. RealKcat predictions capture the direction and magnitude of affinity changes following alanine substitution at catalytic residues, with the majority of mutant values within ±1 bin of experimental measurements.

For the negative hold-out of 10 synthetic alanine variants (Table 1), which included not only single-point substitutions but also several dual-mutant cases (samples 2, 6, and 10), mutation effects were benchmarked against a binary meta-label (drop vs. no-drop for k_{cat}; increase vs. no-increase for K_M) derived from mechanistic descriptors (*Materials and Methods*). On this set, RealKcat achieved precision = 1.00 and recall = 0.80 for k_{cat} drops (F1 = 0.87; MCC = 0.51) and precision = 1.00 and recall = 0.40 for K_M increases (F1 = 0.57; MCC = 0.50), with corresponding confusion matrices shown in Fig. 5A–B. We further examined how well RealKcat could predict categorical k_{cat} and K_M bins for alanine substitutions at all known catalytic residues, including those involving simultaneous perturbation at more than one active-site position. In scatter plots comparing predicted and experimental values, RealKcat's wild-type predictions (Fig. 5C, E) aligned closely with experimental values and were largely within ±1 log bin of the identity line, comparable to other available zero-shot inference models (DLKcat, CatPred, UniKP, and EITLEM-Kinetics). However, we emphasize that RealKcat is not intended as an algorithmic competitor to prior deep-learning frameworks such as DLKcat or CatPred, but rather as a dataset- and evaluation-driven advance. While these published models are included as benchmarks to reflect mechanistic expectations, all primary performance claims for RealKcat are confined to standardized evaluation settings within KinHub-27k.

When extended to catalytic-site mutants, RealKcat captured both the magnitude and direction of expected (meta-labeled) mutation-induced changes. For k_{cat}, MD_{Ala} mutant predictions (Fig. 5D) consistently deviated from their wild-type cognates in accordance with the drop versus no-drop meta-labels, reproducing the separation pattern observed in the categorical evaluation (Fig. 5A). For K_M most MD_{Ala} mutant predictions (Fig. 5F) tracked the expected increase versus no-increase trends, again maintaining separation between perturbed and unaltered variants. In general, the scatter plots (Fig. 5C–F) extend the categorical observations of Fig. 5A–B into the continuous prediction domain, illustrating that the binary drop/increase labels

are reflected in RealKcat's predicted kinetic clusters. Although the evaluation set (n=10) is intentionally small to enforce strict hold-out conditions, these results demonstrate that RealKcat can distinguish major activity losses from retention even in an in-distribution-by-identity regime, where enzyme families are familiar but the explicit negative variants themselves remain unseen.

Taken together, the OOD anchor and in-distribution negative evaluations provide complementary perspectives: in both regimes RealKcat preserves the direction of WT→mutant effects and maintains bin-level agreement with experimental outcomes. Moreover, this behavior extends to non-alanine and broader OOD mutants (Fig. S4), further underscoring that model generalization is not limited to alanine substitutions. We stress that this section evaluates mutation sensitivity—that is, whether predicted movements across kinetic bins are consistent with perturbations at catalytic residues—rather than fine-grained ranking within bins. As expected for an order-of-magnitude classification framework, sub-bin resolution is not achieved; nevertheless, the observed separations provide convergent evidence that RealKcat internalizes biochemical constraints at catalytic sites and generalizes these constraints across distinct evaluation conditions.

Table 1. Independent in-distribution synthetic negative hold-out set of 10 catalytic-site alanine variants. Summary of wild-type (WT) enzymes whose corresponding alanine-substituted mutants used for in-distribution hold-out evaluation. These MD variants are excluded from training to provide a controlled evaluation of RealKcat's mutation sensitivity in an in-distribution-by-identity regime.

Case #	UniProt ID	Substrate Name	Catalytic Residues	k_{cat} [s^{-1}]	K_M [mM]	Organism	Pathway
1	P17516	3alpha-hydroxy-5beta-pregnane-20-one	Y55	$3.3x10^{-2}$	$1.6x10^{-3}$	*Homo sapiens*	steroid metabolism
2	Q9P2T1	gmp	C186, T188	$2.7x10^{-1}$	$1.7x10^{-2}$	*Homo sapiens*	purine metabolism
3	Q6FFQ0	succinate semialdehyde	E264, C301	$5.5x10^{-2}$	$4.5x10^{-1}$	*Acinetobacter baylyi*	carbohydrate acid metabolism
4	P9WNX5	nadp+	C130, H256	$8.5x10^{0}$	$6.5x10^{-2}$	*Mycobacterium tuberculosis*	amino-acid biosynthesis
5	P37213	diphosphate	H454, C839	$8.7x10^{1}$	$6.8x10^{-3}$	*Entamoeba histolytica*	
6	P00924	mn2+	E212, K346	$7.2x10^{1}$	$1.9x10^{-2}$	*Saccharomyces cerevisiae*	carbohydrate degradation/ glycolysis
7	O81413	lipoamide	H479	$3.9x10^{2}$	$3.4x10^{0}$	*Glycine max*	
8	Q91WR5	3beta-hydroxyandrost-5-en-17-one	Y55	$6.8x10^{-2}$	$4.0x10^{-4}$	*Mus musculus*	steroid metabolism
9	Q2MHD9	5alpha-androstane-3alpha,17beta-diol	Y55	$7.0x10^{-2}$	$1.7x10^{-2}$	*Rattus norvegicus*	

10	P0A6L2	l-aspartate 4-semialdehyde	Y133, K161	$1.2x10^{2}$	$1.5x10^{-1}$	*Escherichia coli*	amino-acid biosynthesis

A

	Main: DROP	Main: NO-DROP	Row total
Meta: DROP	TP 7 (70%)	FN 2 (20%)	9 (90%)
Meta: NO-DROP	FP 0 (0%)	TN 1 (10%)	1 (10%)

Precision=1.00, Recall=0.78, F1=0.87, MCC=0.51, tau=2 bins

Case	WT True	WT Pred	MD Pred	Δ bin	Shift	Meta Label	Predicted (WT-MD)	Match
1	2	2	0	2	↓ 2	DROP	DROP	YES
2	3	3	3	0	0	NO-DROP	NO-DROP	YES
3	2	5	1	4	↓ 4	DROP	DROP	YES
4	4	4	1	3	↓ 3	DROP	DROP	YES
5	5	5	5	0	0	DROP	NO-DROP	NO
6	5	5	2	3	↓ 3	DROP	DROP	YES
7	6	6	6	0	0	DROP	NO-DROP	NO
8	2	2	0	2	↓ 2	DROP	DROP	YES
9	2	3	0	3	↓ 3	DROP	DROP	YES
10	6	6	4	2	↓ 2	DROP	DROP	YES

B

	Main: INCREASE	Main: NO-INCREASE	Row total
Meta: INCREASE	TP 2 (20%)	FN 3 (30%)	5 (50%)
Meta: NO-INCREASE	FP 0 (0%)	TN 5 (50%)	5 (50%)

Precision=1.00, Recall=0.40, F1=0.57, MCC=0.50, tau=-1 bins

Case	WT True	WT Pred	MD Pred	Δ bin	Shift	Meta Label	Predicted (WT-MD)	Match
1	0	0	0	0	0	NO-INCREASE	NO-INCREASE	YES
2	1	1	1	0	0	INCREASE	NO-INCREASE	NO
3	2	2	5	-3	↑ 3	INCREASE	INCREASE	YES
4	1	1	1	0	0	NO-INCREASE	NO-INCREASE	YES
5	1	1	1	0	0	INCREASE	NO-INCREASE	NO
6	1	1	1	0	0	INCREASE	NO-INCREASE	NO
7	3	3	3	0	0	NO-INCREASE	NO-INCREASE	YES
8	0	0	0	0	0	NO-INCREASE	NO-INCREASE	YES
9	1	1	1	0	0	NO-INCREASE	NO-INCREASE	YES
10	2	2	5	-3	↑ 3	INCREASE	INCREASE	YES

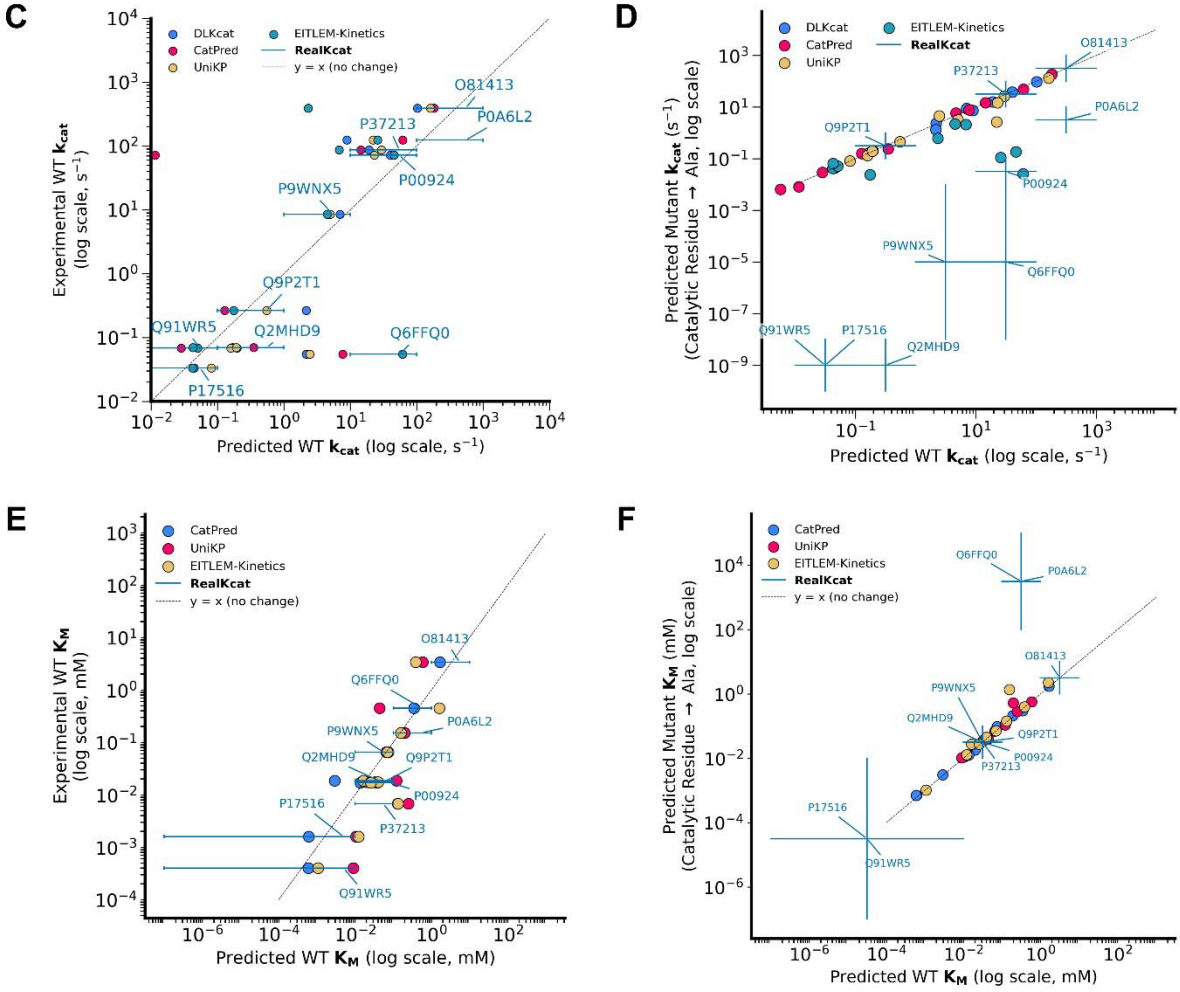


Fig. 5. Evaluation of mutation-aware predictions on a synthetic negative hold-out of 10 catalytic-site alanine variants. (A–B) Confusion matrices and summarizing binary

classification of catalytic loss versus retention across 10 held-out synthetic alanine variants. **(A)** "DROP" for k_{cat}, **(B)** "INCREASE" for K_M. Predictions were benchmarked against meta-labels derived from mechanistic descriptors. **(C–F)** Scatter plots comparing predicted versus experimental kinetic values for wild-type (WT) and mutant (MD_{Ala}) enzymes. **(C–D)** k_{cat} values for WT **(C)** and MD_{Ala} mutants **(D)**. **(E–F)** K_M values for WT **(E)** and MD_{Ala} mutants **(F)**. The diagonal line denotes the identity relationship (x = y). In panels C–F, RealKcat predictions are shown as blue line segments connecting the corresponding experimental and predicted values. The length of each line segment denotes the span of the predicted log-bin cluster, while the direction of shift away from the identity line in panels D and F indicates whether the model predicted a decrease (leftward/downward) or increase (rightward/upward) relative to experimental values. These line segments are unique to RealKcat and should not be conflated with either the dotted identity line (x = y), which indicates perfect agreement, or the markers of other models (DLKcat, CatPred, UniKP, EITLEM-Kinetics), which are displayed as colored points without connecting segments.

Few-Shot Generalization in a High-Throughput Mutational Landscape

To extend mutation-aware testing beyond curated catalytic residues and synthetic negatives, we employed the high-throughput dataset of the alkaline phosphatase (PafA) from *Elizabethkingia meningoseptica*, originally reported by Markin et al. (*29*). This dataset systematically introduces single-point mutations (glycine, valine and alanine) across the enzyme sequence to probe catalytic and structural roles, yielding kinetic measurements (k_{cat} and K_M) for 1,016 variants with the substrate carboxy 4-methylumbelliferyl phosphate ester (cMUP), a binding-limited substrate. For evaluation with RealKcat, we partitioned the dataset into 398 training, 310 validation, and 310 test variants, ensuring that the training distribution spanned the high, medium, and low ranges of observed turnover values. Crucially, while the wild type (WT) and a catalytic-site variant (T79S) were present in the training set, the key catalytic mutations R164A and R164G—located at a residue central to substrate coordination (Fig. 6A)—were held out for the test split to specifically assess inferencing on unobserved but catalytically significant mutations.

Across the PafA test set, RealKcat achieved 51% exact-bin accuracy for k_{cat} and 94.2% for K_M, with markedly higher e-accuracy values of 95.2% (k_{cat}) and 100% (K_M) within ±1 order of magnitude (Fig. 6C–D). As shown in Fig. 6B, the PafA test subset spans multiple kinetic bins

with no single class dominating the distribution, confirming that these results are not attributable to class imbalance or majority-class effects. At the residue level, RealKcat correctly reproduced the bin-level placement and directional shifts associated with catalytic substitutions: both R164A and R164G were predicted within the same experimental bins as their *in vitro* measurements (Fig. 6C–D), with R164A and R164G showing marked reductions in turnover. This agreement underscores the model's sensitivity to side-chain chemistry and conformational flexibility at catalytic sites. Nevertheless, we also observed occasional off-target predictions for other mutations, in which predicted values deviated from experimental bins without systematic bias and not within ±1 order of magnitude of experimental bins. Structural analysis suggests these discrepancies may arise from the dense network of contacts at these off-target residues, for instance H486 is highly interconnected with at least 17 neighboring residues (see Fig. S5). Perturbations at this residue are therefore likely to propagate non-locally, complicating predictive accuracy. At the same time, a recurring pattern was observed in cases of misclassification: when RealKcat failed to capture the mutation-induced impairment, predictions often fell back to a conservative baseline near the wild-type bin, typically within ±1 log unit of the WT value. This behavior reflects a conservative bias that prioritizes stability over speculative large deviations, but it also reveals where mutation effects may be underestimated.

To probe these failures systematically, we scanned for mutants experimentally reported to exhibit ≥2-bin drops from the WT reference bin (e.g., WT at bin 6, mutant ≤ bin 3) and then examined whether RealKcat could recover these severe impairments within ±1 log-bin tolerance. Strikingly, only D305 was predicted correctly within ±1 bin of the experimental cluster, while the majority of other severe-drop mutants were misclassified back toward the wild-type reference, reflecting a conservative fallback bias. These failures clustered primarily at loop-like or disordered regions, or at structurally dense positions where ≥10 residues fall within a 6 Å radius. Examples include G56, G465, L297, D144, F332, and H486—sites embedded in flexible or highly interconnected regions where mutation-induced perturbations likely propagate in ways not captured by sequence embeddings alone. In such contexts, RealKcat underestimates the magnitude of catalytic loss, highlighting its current limitation in modeling destabilizing mutations at disordered or densely packed sites without explicit structural context.

Importantly, this analysis underscores the value of ±1 log-bin e-accuracy as a physiologically relevant measure. By maintaining predictive fidelity in such few-shot settings,

RealKcat demonstrates generalizability to densely sampled mutational landscapes, complementing its performance on curated OOD anchors and synthetic negatives. Moreover, experimental enzymatic turnover at catalytic residues can vary by up to an order of magnitude across assay conditions due to effects of pH, ionic strength, temperature, and protonation states (Fig. S1). Within this biological variability, RealKcat consistently reproduced the qualitative directionality and quantitative bin-level placement of PafA mutations.

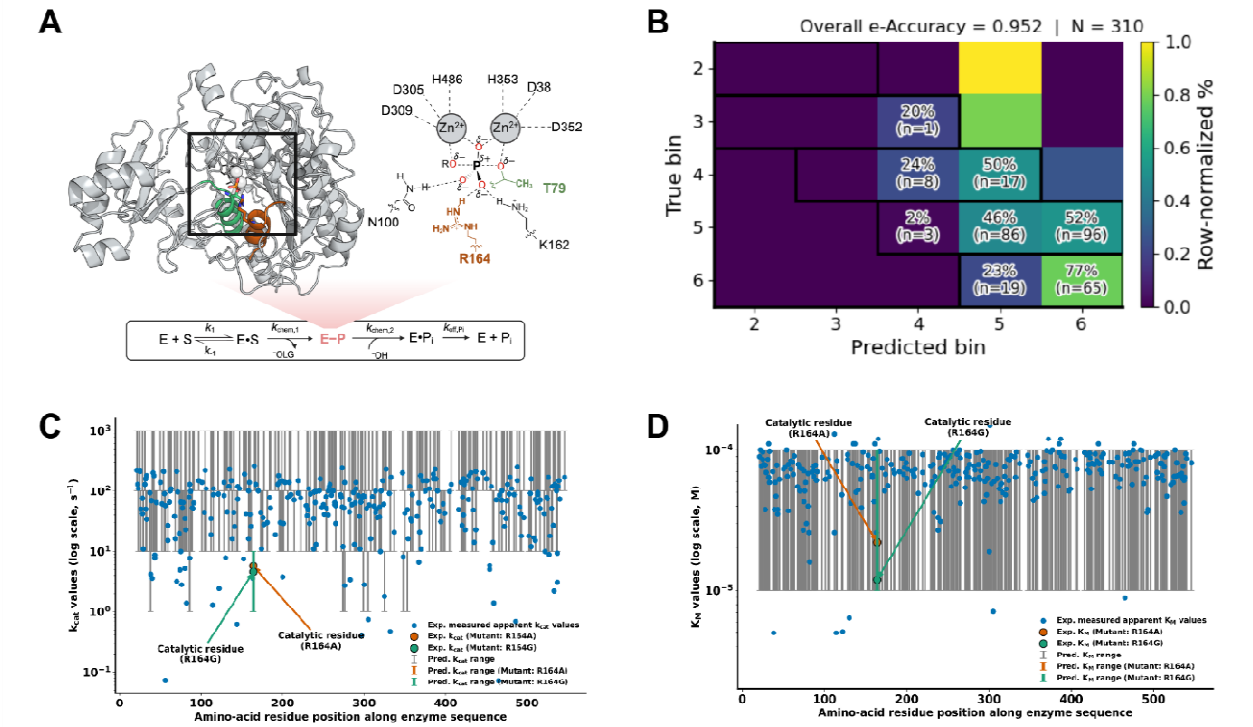


Fig. 6. Few-Shot Learning Evaluation of RealKcat's catalytic relevance prediction and sensitivity to mutations of alkaline phosphatase. (A-D) (A) Structure of alkaline phosphatase PafA with highlighted catalytic residues in transition state, showing positions T79 and R164 used in mutant testing. **(B)** Confusion-matrix-style visualization of predicted versus true bins (N = 310). Each cell shows the row-normalized percentage and the number of samples in parentheses; the overall ± 1 bin e-accuracy = 0.952. No single class dominates, confirming balanced coverage across bins 3–6. **(C-D)** Sequence-wide distribution of **(C)** and K_M **(D)** values across the PafA test dataset, with RealKcat predictions compared to experimental data, specifically for key R164A and R164G mutations.

Discussion

RealKcat provides a mutation-sensitive, catalytically aware framework for predicting enzyme kinetics by integrating rigorous dataset curation, pretrained biochemical language models, and a robust classification-based modeling scheme. Rather than relying on methodological novelty at the algorithmic level, our approach demonstrates how careful attention to data quality, representation, and biologically grounded evaluation can substantially improve reliability and generalizability. The KinHub-27k dataset, built through meticulous reconciliation of thousands of source articles and manual resolution of inconsistencies, underpins this framework and distinguishes it from prior enzyme kinetics resources. Median pooling of duplicates, mechanistically guided generation of negatives, and stratified data partitioning together establish a dataset and modeling strategy that is resistant to spurious variability and leakage—issues that often obscure performance in regression-based formulations. By recasting k_{cat} and K_M as order-of-magnitude classification tasks, RealKcat addresses the intrinsic variability of kinetic measurements and provides predictions in physiologically interpretable ranges, bridging the gap between *in vitro* assay heterogeneity and *in vivo* functional relevance (*24*).

Moreover, RealKcat's dual-input formulation is not an architectural embellishment but a biological necessity. Kinetic parameters such as k_{cat} and K_M arise from the coupled enzyme–substrate context; enzyme-only or substrate-only models cannot resolve this interdependence and therefore underperform when tasked with catalysis-specific predictions. By grounding our modeling unit in enzyme–substrate pairs, RealKcat aligns computational inference with the biochemical basis of catalysis, avoiding the pitfalls of single-input formulations more appropriate for binding screens than for turnover and affinity estimation. Importantly, most experimental kinetic assays are performed under saturating concentrations of co-substrates and cofactors, such that reported k_{cat} and K_M values are conditional on the specific enzyme–substrate pair being assayed. This experimental design makes it biologically inconsistent to decouple enzyme from substrate in predictive modeling: a single-input model cannot capture the conditional dependency imposed by assay context. Hence, single-input ablations were considered out-of-scope for RealKcat, as they would conflict directly with our rationale and reduce biological interpretability.

Furthermore, the classification-based formulation adopted in RealKcat is not merely a technical simplification but a deliberate reflection of how kinetic data are interpreted in practice.

Turnover numbers and affinities are rarely known with precision beyond orders of magnitude, and even for the same enzyme can vary by several bins, as much as three orders of magnitude, depending on experimental context (*1, 24, 30*). For instance, according to criteria from Bar-Even et al. (*24*), 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, reflecting known variation in catalytic efficiency. Moreover, *in vivo* data themselves can be noisy and context-dependent due to the temporal variability and regulatory dynamics inherent to living systems (*31, 32*), complicating their direct interpretation.

By training the model to operate at this resolution, we emphasize robustness and functional interpretability over pointwise accuracy. The use of e-accuracy as a central evaluation metric acknowledges this reality, rewarding predictions within ±1 order of magnitude of experimental values, while macro-averaged MCC ensures balanced evaluation across abundant and rare kinetic classes. Together, these measures prioritize stability and biological meaning over superficial improvements in regression metrics. This evaluation strategy also clarifies why apparent "failures" often reflect a conservative bias: when RealKcat misclassifies, predictions frequently fall back toward wild-type activity rather than producing implausible deviations, underscoring both its strengths and its current limitations. Technically, from the inputs, residue-level embeddings were aggregated via mean pooling to yield fixed-length descriptors that capture conserved evolutionary signals, sequence-context dependencies, and biochemical constraints across enzyme families. While such pooling reduces positional resolution and may obscure localized catalytic effects, it provides a stable and tractable representation for large-scale training. Importantly, RealKcat's sensitivity to catalytic mutations does not arise from explicit residue-level encoding, but from the curated inclusion of mutant entries and mechanistically guided negatives within KinHub-27k. This supervision compels even pooled embeddings to learn mutation-aware distinctions, enabling generalization across out-of-distribution anchors, synthetic negatives, and dense mutational landscapes.

Our analyses further demonstrate that mutation sensitivity is a defining feature of RealKcat. Across three complementary tiers—out-of-distribution anchors, in-distribution synthetic negatives, and the dense PafA mutational landscape—the model consistently

reproduced mutation-induced movements in kinetic bins and maintained bin-level alignment with experimental outcomes. These results highlight the framework's ability to capture sequence-residue-function relationships beyond trivial memorization, with particular robustness at annotated catalytic residues. At the same time, the PafA evaluation revealed critical limitations which are common to all tools at the moment. RealKcat occasionally underestimates large functional losses, particularly at disordered or densely connected sites such as G56, L297, or H486, where mutational effects propagate through non-local networks that are not fully encoded in sequence embeddings. These observations suggest that while the current model captures catalytic awareness fairly at residue level, it remains limited in resolving mutation effects mediated by structural disorder, allosteric coupling, or long-range interactions. Future incorporation of structure-derived features—such as residue contact density, dynamic flexibility, or cavity context—may address these shortcomings and extend predictive fidelity in challenging regions. We also believe that release of experimental datasets where multiple substrates are tested for activity with the same enzyme (similar to our PafA dataset) will elevate the accuracy and catalytic awareness of RealKcat (and like tools) to the next echelon. One can imagine using updated versions of RealKcat with additional structural priors for identifying mutations to improve activity of target enzymatic reactions.

The implications of RealKcat extend across multiple domains. In enzyme engineering and synthetic biology, the ability to resolve mutation-induced shifts in catalytic parameters provides a rational guide for pathway optimization, reducing reliance on costly experimental screening. In biomedical research, the mutation-sensitive framework offers a predictive platform for prioritizing variants of uncertain significance, enabling early identification of substitutions likely to impair catalytic activity or alter substrate affinity (*33*). Such capabilities are particularly relevant for metabolic disorders and pharmacogenomics, where subtle shifts in kinetic parameters can have outsized physiological consequences (*34*). In systems biology, RealKcat could support enzyme-constrained and proteome-constrained metabolic models by supplying reliable turnover and affinity estimates in magnitude ranges directly compatible with constraint-based formulations (*35–37*). By narrowing parameter uncertainty, it could reduce heavy reliance on stochastic surrogates and improves the interpretability of flux predictions under varying environmental conditions (*38–40*). Across these applications, the emphasis on biologically

meaningful classification ranges ensures that RealKcat provides not only accurate predictions but also outputs that can be directly integrated into experimental and computational workflows.

In conclusion, RealKcat represents a step toward biologically interpretable, mutation-aware prediction of enzyme kinetics. Its robustness arises not from algorithmic novelty but from principled dataset curation, mechanistically informed negatives, and carefully chosen evaluation strategies that reflect biochemical practice. While limitations remain in capturing large mutational effects at structurally complex or disordered regions, the framework already establishes a reliable baseline for both predictive and exploratory applications. By embedding enzyme kinetics into an order-of-magnitude classification paradigm, RealKcat aligns computational inference with the practical resolution of experimental biology, offering a versatile tool for enzyme engineering, disease modeling, and systems-level simulations. Future work integrating structural and dynamic descriptors promises to extend these capabilities further, enabling predictive models that not only track catalytic variation but also anticipate the broader consequences of mutations within the complexity of cellular environments.

Limitations of the Study

Despite its robustness, RealKcat operates under several constraints that define its current scope. The curated KinHub-27k dataset, while extensive, may be biased toward well-characterized classes such as hydrolases and transferases, with translocases and certain ligases underrepresented (**Fig. 1B**), reflecting broader reporting biases in enzymology rather than modeling artifacts. Substrate embeddings derived from isomeric SMILES capture stereochemistry and functional groups but cannot fully represent conformational flexibility or three-dimensional electrostatics, limiting predictions for stereochemistry-rich or structurally complex molecules. Furthermore, framing kinetics as an order-of-magnitude classification task introduces interpretive strengths and weaknesses: e-accuracy acknowledges the physiological variability of enzyme kinetics, which can fluctuate by ±1 log unit across assays due to differences in pH, ionic strength, and temperature (25, 29–31), yet it cannot resolve subtle intra-bin effects, as seen in the PafA analysis where R164A, R164G, and R164V collapsed into the same kinetic cluster despite differing severities of impairment. In addition, RealKcat shows a conservative fallback toward wild-type predictions when challenged with severe mutations, particularly in out-of-distribution or structurally dense contexts. This bias masks the true

magnitude of loss for mutations at loop-like or highly interconnected regions such as G56, L297, and H486. The construction of synthetic negatives through alanine substitution at catalytic residues similarly enforces a binary drop/no-drop labeling scheme that strengthens catalytic sensitivity but risks overstating inactivation and neglecting nuanced partial-function variants, an issue compounded by inconsistent metadata on assay conditions within KinHub-27k. Finally, while the dataset includes about 10,000 mutant entries, it is skewed toward loss-of-function variants, offering limited exposure to gain-of-function or context-dependent activity enhancements. Addressing these limitations will require principled integration of structural priors (e.g., residue depth, cavity accessibility, contact density), incorporation of condition-aware modeling to capture assay-dependent variability, and expansion toward balanced mutational landscapes. Notably, naive structural integration has been shown in related work (e.g., CatPred (*10*)) to degrade performance when context is not carefully modeled, underscoring that future refinements must be guided by mechanistic insights and usage of a catalytically relevant sub-structure for each enzyme that stabilizes the active site through direct hydrogen bonding or allostery. We have put forward a recipe to collate such sub-structures for >19,000 experimentally crystallized hydrolases in a reaction core database (RC-Hydrolase) which is now housed within the Protein Data Bank (*41*). These are purported to be stronger regulators of model performance than data volume.

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] (*12*) and SABIO-RK [as of May 2024] (*13*), using custom Python scripts [steps in **Fig. 7**]. 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 non-missing fields including those with unique PubMed IDs (PMIDs) to aid verification downstream. This resulted in 96,782 entries from BRENDA and 54,906 from SABIO-RK. Substrate isomeric SMILES keys were obtained from PubChem database (*42*) using IUPAC name of substrates, ensuring consistency across databases to eliminate redundancies. We specifically used the InChI intermediate rather than directly downloading SMILES strings from

PubChem, since multiple non-canonical SMILES notations are often stored for the same compound. Conversion through InChI ensured a unique, standardized representation of each substrate before canonicalization with RDKit, thereby minimizing structural inconsistencies and improving reproducibility across downstream embeddings. After extensive data cleaning, entries from both databases were merged, incorporating sequence data from UniProt (*15*) 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 33,442 k_{cat} and 48,103 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).

Following merging, duplicate entries sharing the same isomeric SMILES–sequence pair (i.e., identical enzyme–substrate combinations, regardless of publication source) were collapsed by taking the median value of each kinetic parameter. This refinement replaced the popular maximum k_{cat} and minimum K_M filtering approaches, which—while useful for highlighting theoretical performance limits—proved sensitive to atypical or poorly replicated measurements (Fig. S1). Reported values for the same enzyme–substrate pair often varied substantially due to differences in experimental conditions, assay protocols, and the range of organisms or expression systems from which the enzymes were sourced. Median pooling therefore provided a more robust representative value for each unique pair, minimizing bias from outlier measurements while preserving the underlying biological signal across diverse physiological and environmental constraints. The resulting dataset—comprising 14,709 wild-type and 9,808 mutant (MD) entries—captures a broad spectrum of catalytic and affinity ranges while avoiding redundancy.

Prior to median pooling, we do recognize that a significant proportion of these entries might have discrepancies compared to the original references due to data entry errors (*24*). To this end, 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 1804 entries and adding 1,072 new entries (Table 2). For instance, the revised k_{cat} and K_M values for the human m-NAD-malate dehydrogenase (D102E) mutant in the conversion of fumarate to pyruvate are 60.61 s^{-1} and 0.00091 M, respectively—correcting the previously reported

erroneous values of 44.14 s□¹ and 0.00255 M. Similarly, for the *Lactobacillus pentosus* D-lactate dehydrogenase (Y52A) mutant catalyzing 2-ketocaproate, the corrected k_{cat} and K_M values are 102 s□¹ and 0.0007 M, replacing the inaccurate earlier 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 or ESM-C does not accept sequences longer than this length. Pairwise sequence identity was computed across all unique enzyme sequences using all-vs-all global alignments (Biopython pairwise2.align.globalxx, implementing the Needleman–Wunsch algorithm with unit match scores and zero mismatch/gap penalties). The resulting percent identity scores were used to construct a graph representation, with distances defined as 1−identity. A minimum spanning tree (MST) was extracted to provide a parsimonious backbone of the sequence space, and each node was annotated by its primary EC class. The network was visualized using a distance-aware layout (**Fig. 1C**), thereby highlighting intra-class homology while also capturing the global diversity of sequence relationships represented in KinHub-27k.

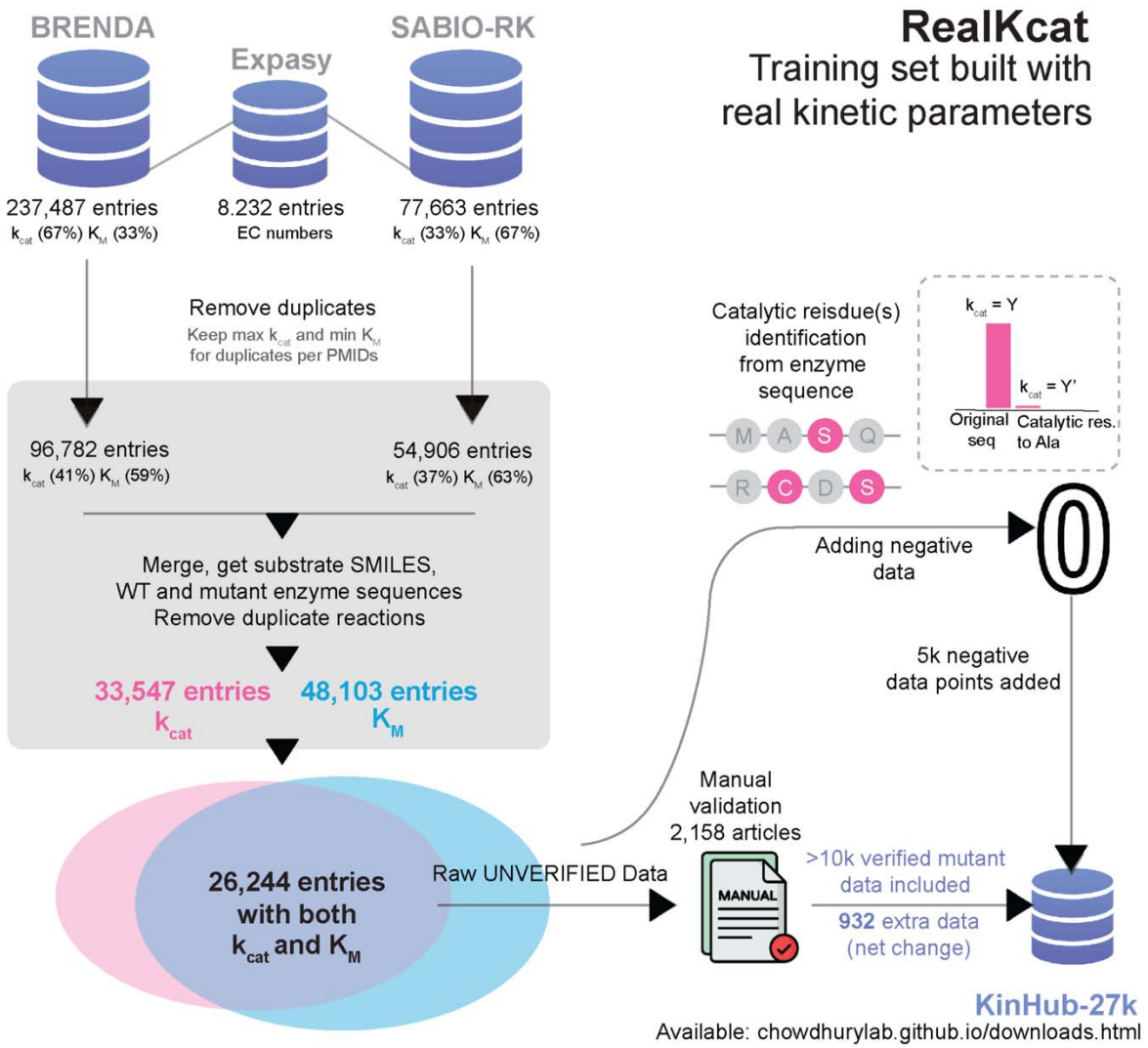


Fig. 7. Data collection steps for constructing KinHub-27k from BRENDA and SABIO-RK databases. Raw enzyme kinetic data were sourced from BRENDA (237,487 entries) and SABIO-RK (77,663 entries), covering 8,232 unique EC numbers. Initial processing involved removing duplicate entries while retaining maximum and minimum K_M values, resulting in 96,782 entries from BRENDA and 54,906 from SABIO-RK. After merging the datasets, obtaining SMILES for substrates, and distinguishing between wild-type and mutant sequences, further duplicates were removed, yielding 33,547 unique entries for and 48,103 for K_M. The resulting dataset, KinHub-27k, contains 26,244 entries with paired and K_M values, prepared for manual validation through article review of the original data sources. Note that negative datapoints were added for training but not assigned = 0 (see Materials and Methods for labeling policy).

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.

Parameter	corrected	removed
K_M	788	10
k_{cat}	618	10
substrates	240	8
mutations	18	21
duplicates	-	91
additional data	1072	-

Binning Strategy for Enzyme Kinetic Parameters

In RealKcat, continuous enzyme kinetic parameters k_{cat} and K_M values are reformulated as multi-class classification tasks to address the substantial order-of-magnitude variability inherent in enzyme kinetics and the practical limitations of direct regression. Kinetic parameters were then discretized into biologically meaningful, $\log_{10}$-spaced bins, enabling robust range prediction while reducing the impact of experimental noise. Specifically, 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, 1G, and 1H). 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 $[10^{-14},10^{-5},10^{-4},10^{-3},10^{-2},10^{-1},10^{4}]$ 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.

Mechanistically Guided Synthetic Negatives and Hybrid Relabeling

To model catalytic loss events and expand training coverage, we constructed a mechanistically guided synthetic negative dataset from UniProt-annotated catalytic, substrate-binding, and cofactor-binding residues. Each wild-type sequence underwent alanine-scanning mutagenesis, a strategy that preserves backbone conformation while removing side-chain chemistry to selectively attenuate catalytic function with minimal structural disruption. For multi-residue active sites, all annotated active site residues were substituted simultaneously to emulate severe functional impairment. This process yielded 5,278 high-confidence synthetic negatives, including 22 experimentally curated wild-type–mutant ($WT–MD_{Ala}$) pairs spanning

diverse enzyme–substrate contexts (see Supplemental Table 1). A ≤60% global sequence identity threshold was enforced only between all data subsets and the 22-sample synthetic negative anchor set to maintain OOD separation. The similarity distributions shown in Fig. 3D–E were computed between the test and training partitions for performance stratification and may therefore include identities exceeding 60%. Kinetic effects in these curated pairs were quantified as the bin difference $\Delta b = b_{WT} - b_{MD}$ on the clustered kinetic scale (i.e. the same $\log_{10}$-spaced bins employed in the data preprocessing pipeline). For k_{cat}, a significant drop was defined as $\Delta k_{cat} \geq 2$ bins (i.e. ≥ two orders-of-magnitude reduction), ensuring that only major catalytic losses were labeled positive. For K_M, a drop was defined as $\Delta K_M \leq -1$ bin (≥ one order-of-magnitude increase), reflecting weaker substrate affinity. This asymmetric thresholding is consistent with biochemical expectations that active-site disruptions predominantly affect turnover rather than binding, unless direct contact residues are altered.

Because most synthetic negatives lacked experimental kinetic measurements, we implemented a two-stage relabeling framework. The curated WT–MD_{Ala} pairs served as the training set for a meta-learned binary classifier, which was then used to assign **drop/no-drop** labels to the broader unlabeled synthetic set. For each enzyme–substrate pair, we computed a suite of mechanistically relevant descriptors spanning mutational severity, substrate chemistry, and interaction scale. Mutational severity was quantified as the sum of BLOSUM62 substitution scores for alanine replacement of annotated residues, providing an evolutionary substitution cost metric. Substrate chemistry was characterized by molecular weight, Crippen's logP, total ring count, hydrogen bond donor count, and hydrogen bond acceptor count from RDKit-parsed isomeric SMILES. Interaction-scale metrics included active-site density (annotated residues per sequence length) and substrate-to-enzyme molecular weight ratio, with the wild-type kinetic bin for the relevant parameter included as a baseline activity indicator.

Binary classification for both k_{cat} and K_M drop prediction tasks was performed using the *AutoTabPFNClassifier*, a Transformer-based meta-learned prior optimized for tabular data (*43*). Given the small curated set, Leave-One-Out Cross-Validation (LOOCV) was employed to maximize data usage while maintaining strict independence between training and test samples. In each fold, features were standardized using a *StandardScaler* fit exclusively on the training set to prevent information leakage. *AutoTabPFN* models were trained with a per-fold time limit of 1000 seconds on GPU and evaluated via classification accuracy and aggregated confusion

matrices. Final models for k_{cat} and K_M were retrained on their complete datasets and serialized alongside their scalers using *joblib*, enabling reproducibility and deployment within the downstream RealKcat training pipeline. Hence, the resulting binary classifiers were used to annotate unlabeled synthetic negatives as drop or no-drop. These labels informed *a hybrid relabeling policy*: in k_{cat} tasks, drops were randomly reassigned to bins well below (≥ two orders-of-magnitude reduction) the WT value and non-drops constrained within ±1 bin; in K_M tasks, drops shifted to higher bins (≥ one order-of-magnitude increase) while non-drops were restricted to WT or lower bins. Thus, ensuring kinetic plausibility, generating hard negatives that are sequence-close to actives but mechanistically divergent in catalytic space.

Partitioning Strategy for Leakage-Minimized Data Splits

To prevent over-representation of densely sampled regions and maintain structural diversity, the negative dataset was partitioned using a manifold-guided strategy. The enzyme subspace (the first portion of the concatenated features corresponds to enzyme embeddings) was projected into two dimensions using PaCMAP, a dimensionality reduction method that preserves both global and local relationships while mitigating distortions common in high-dimensional Euclidean space. Affinity Propagation (AP) clustering was then applied to the PaCMAP projection to identify exemplar-based sequence clusters. Clusters containing fewer than five members were reassigned to a "miscellaneous" bin (label −1) to prevent small groups from disproportionately influencing stratification. Partitioning proceeded in two stages. First, a stratified hold-out split with respect to AP cluster labels and their associated drop/no-drop classifications for both k_{cat} and K_M set aside 5% of the negative set as a temporary pool, preserving the empirical distribution of these labels (including the −1 "miscellaneous" bin). Second, the temporary pool was bisected evenly into validation and test subsets, yielding approximate final proportions of ~95%/2.5%/2.5% for train/val/test. All random seeds were fixed (random_state=42) to allow exact regeneration of splits across runs.

To preserve biological representativeness and statistical rigor, we implemented a two-tier, source-aware positive set partitioning framework. The aim was to maintain the endogenous balance between wild-type (WT) and mutant-derived (MD) entries while controlling for potential covariates between biological source and kinetic class. First, a catalytic inference hold-out of 100 positive samples was constructed by proportionally sampling from WT and MD pools without replacement, preserving the natural WT:MD (60:40) ratio and ensuring statistical independence

from the training data. This was performed under a fixed random seed of 42. In parallel, a negative hold-out of 10 synthetic variants was selected at random from the partitioned negative test pool. These negatives were generated by alanine substitution at annotated catalytic residues. Although their corresponding wild-type sequences are present in the positive training partitions, the exact negative sequences were excluded from training. Consequently, this configuration ensures that negatives are evaluated in-distribution with respect to enzyme identity but remain unseen as explicit negative examples. This two-tier scheme enables evaluation across both positive (WT and MD) and synthetic negative chemotypes, providing a controlled framework for assessing generalization and mutation sensitivity.

The remaining positives underwent five-fold cross-validation stratified on a composite label that jointly encoded kinetic bin and biological source (WT/MD). This ensured balanced representation of both factors in every fold. Negative sets were handled in a paired, contrastive manner: synthetic negatives corresponding to a given WT enzyme were co-assigned to the same partition as their cognate positives. This pairing preserved the biological relevance of evaluation and ensured that synthetic negatives were consistently matched with their related WT samples across train/validation/test splits. Within each fold, the outer 20% of the data was evenly divided into validation and test subsets under the same composite-label stratification. This procedure yielded approximate 80/10/10 proportions for train/validation/test in each fold, while preserving both kinetic class distribution and WT:MD balance. In addition, all partitioning steps—including manifold-guided clustering for the negative set, proportional WT:MD sampling for the positive set, and positive–negative pairing—were performed with a fixed random seed (42). This ensured that every split could be exactly reproduced, maintaining consistency across experiments and enabling rigorous model comparison.

Feature Representation and Embeddings

In RealKcat's feature extraction pipeline, each sample from the curated KinHub-27k dataset comprises an enzyme sequence, an isomeric SMILES representation of the substrate, and assigned log-binned clusters for k_{cat} and K_M. Enzyme sequences were first evaluated using embeddings from both ESM-2 (esm2_t33_650M_UR50D) and ESM-C (600M; EvolutionaryScale/esmc-600m-2024-12), protein language models trained on large-scale sequence corpora to produce context-aware embeddings (*11, 25*). For both models, the CLS ([0]) and terminal/padding tokens were excluded, and residue-level embeddings across positions 1…L

(where L is sequence length) were mean-pooled to produce sequence descriptors, ensuring that representations reflected residue content rather than global classification tokens. Embeddings were drawn from the final layers (layer 33 for ESM-2 and layer 36 for ESM-C) to capture conserved evolutionary signals, sequence-context dependencies, and biochemical constraints across enzyme families.

Although mean pooling inevitably reduces positional resolution and may obscure localized effects at catalytic residues, it provides a stable and computationally tractable representation for large-scale training. As our evaluations showed, this approach retains sufficient sensitivity to mutation-induced perturbations when paired with rigorous curation and negative sampling. Between the two models, both ESM-2 (1,280-dimensional) and ESM-C (1,152-dimensional) supported robust performance, but ESM-C consistently offered modest advantages in mutation-sensitive and out-of-distribution tests (see Results, **Fig. 3**). Based on these comparative results, ESM-C was adopted as the default enzyme representation in RealKcat.

On the other hand, substrate features were encoded using **ChemBERTa** (seyonec/PubChem10M_SMILES_BPE_450k) embeddings, derived from isomeric SMILES strings and averaged across all tokens (including the CLS token) including the CLS token of final hidden states to yield 768-dimensional molecular descriptors of all subtrates/co-factor/ions. This whole-sequence pooling preserves information from both token-level chemical substructures and the global sequence summary embedded in the CLS representation, capturing stereochemistry, functional groups, and physicochemical properties directly relevant to substrate recognition and catalysis (*18*).

We note that CLS token handling differed between enzyme and substrate embeddings. For ChemBERTa, we retained the CLS token in the mean-pooling operation, since this embedding has been pretrained to act as a global sequence summary that captures molecular-level context beyond token-local features. In contrast, for ESM protein embeddings, the CLS token was excluded, as prior benchmarks indicate that mean pooling across residue tokens (excluding CLS and padding) yields more stable and biochemically relevant sequence descriptors. This design choice reflects the distinct pretraining objectives of the two models and was confirmed in our diagnostics to provide more consistent performance across tasks.

For each enzyme–substrate pair, the ESM-C (1,152-D) and ChemBERTa (768-D) embeddings were concatenated to form a unified 1,920-dimensional joint feature vector,

integrating both evolutionary context and substrate chemistry in a balanced representation. This formulation does not enforce explicit interaction modeling, but rather provides a feature space from which the classifier can learn associations between catalytic variation and molecular context. Prior to model training, all features were globally standardized, and partition-specific normalization parameters were estimated strictly on training folds to avoid leakage.

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 (*26*), 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: $\boldsymbol{L(\theta)} = \sum_{i=1}^{n} \boldsymbol{l(y_i, \hat{y}_i)} + \sum_{i=1}^{K} \boldsymbol{\Omega(f_{(k)})}$ where $\boldsymbol{l(y_i, \hat{y}_i)}$ denotes the log-loss for multi-class classification, specifically capturing the errors between true labels $\boldsymbol{y_i}$ and predictions $\boldsymbol{\hat{y}_i}$. The second term, $\boldsymbol{\Omega(f_{(k)})} = (\boldsymbol{\gamma} \times \boldsymbol{T}) + (\boldsymbol{0.5} \times \boldsymbol{\lambda} \times ||\boldsymbol{w}||^2)$, is a regularization penalty that controls model complexity by penalizing each tree's structural complexity, with $\boldsymbol{T}$ representing the number of leaves in a tree and $\boldsymbol{w}$ 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. $||\boldsymbol{w}||^2$: 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 $||\boldsymbol{w}||^2$ 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. $\mathbf{0.5} \times \lambda \times ||\boldsymbol{w}||^2$: 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 $\boldsymbol{\Omega(f_{(k)})}$ 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 **additiv**e update for each tree in the boosting process can be described by: $\boldsymbol{f_{(k+1)}(X) = f_{(k)}(X) + \eta \times PseudoResiduals,}$ where $\boldsymbol{\eta}$ is tuned to gradually refine predictions without over-adjusting for noise in the data.

Training and Hyperparameter Optimization*:* Model hyperparameters were optimized through a random-search strategy implemented in Optuna under the earlier described non-conventional stratified 5-fold cross-validation scheme. Stratification preserved the distribution of kinetic parameter bins across folds, ensuring that both abundant and minority classes were adequately represented in each split. The Optuna search optimized over validation accuracy as the objective function, with candidate models evaluated on held-out validation partitions to prevent information leakage. The final hyperparameters selected for downstream training were: number of estimators = 1920, maximum depth = 11, learning rate (η) = 0.0021, maximum delta step (δ) = 0, and L1 regularization (α) = 1.13, L2 regularization (λ) = 14.803. Early stopping was employed with a patience of 10 rounds on the validation negative log-loss (cross-entropy loss) as the

monitoring metric, providing an effective safeguard against overfitting while preserving generalization.

To address class imbalance, particularly in underrepresented kinetic bins, Synthetic Minority Over-sampling Technique (SMOTE) oversampling was applied exclusively to the training partition of each fold following the method of Chawla *et al.* (*28*), thereby enhancing representation without leakage into validation or test sets. SMOTE generates synthetic minority-class instances by interpolating between a given sample and its k-nearest neighbors, thereby enhancing representation without altering class labels. We used the default parameter $k=5$, which yielded an approximately even distribution across kinetic bins while preserving intra-class variability. For reproducibility, a fixed random seed (42) was applied to ensure identical synthetic instance generation across runs. Importantly, oversampling was performed **after binning**, allowing targeted correction within each order-of-magnitude kinetic bin—an essential step given the naturally skewed distribution of enzyme kinetics data.

Synthetic negatives, introduced via the hybrid relabeling policy, were further down-weighted by a constant factor (0.001) to reduce their disproportionate influence while retaining their calibration role; weights for positive entries were left unchanged. All weights were assigned in full-batch form, and datasets were shuffled with a fixed random seed (42) for reproducibility. Following the initial training pass, an optional hard negative mining (HNM) procedure was applied: misclassified training examples—especially false negatives not within ±1 order of magnitude —were identified, their sample weights boosted by a factor of 10^2, and the reweighted examples reintroduced into the training pool. The model was then fine-tuned for an additional ≥10 boosting rounds (approximately 20% of the original boosting iterations), starting from the previous booster state. This two-stage scheme improved sensitivity to difficult decision boundaries, particularly for mutation-sensitive bins and structurally challenging negatives, while maintaining stability and interpretability across cross-validation folds.

Moreover, in our 5-fold stratified cross-validation scheme, each fold produced an independently trained RealKcat model with its own training/validation/test split. For benchmarking, we reported mean ± SD performance across all five folds (to reflect population-level reproducibility), but for downstream inference we needed a single deployed model. To this end, we selected **fold-5** as the representative model for inference. Fold-5 was chosen a priori to

be held aside as the final inference model in our experimental design, avoiding any post-hoc cherry-picking.

***Feature Standardization**:* Given the heterogeneous nature of the input features, each embedding group—ESM 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: $\boldsymbol{x}' = \frac{(x - \boldsymbol{\mu})}{\boldsymbol{\sigma}}$, 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 capacity was evaluated using a suite of metrics specifically chosen for multi-class classification under class imbalance. Standard accuracy, $Accuracy = \frac{\sum_i^N (\hat{y}_i = y_i)}{N}$, defined as the proportion of exact bin matches between predicted ($\hat{y}_i$) and true (y_i) labels, provides a baseline measure of overall correctness but is limited in this context due to the skewed distribution of kinetic parameters across bins. To complement this, we reported class-specific precision, recall, and F1-scores, macro-averaged across classes to prevent dominance by majority bins.

A central metric in this study is the **error-tolerant accuracy (e-accuracy)**, which defines correctness within ±1 $\log_{10}$-spaced bin of the experimental label: $e_accuracy = \frac{\sum_i^N (|\hat{y}_i - y_i| \leq 1)}{N}$, reflecting the biological reality that order-of-magnitude deviations rarely alter functional interpretation. To further assess balanced classification, we used the **Matthews Correlation Coefficient (MCC)**, macro-averaged across bins, which incorporates all entries of the confusion matrix and is less sensitive to class imbalance than raw accuracy. Finally, the **area under the precision–recall curve (AUC-PR)** was included as it provides a more informative measure than ROC curves under imbalanced conditions.

Visual diagnostics complemented these quantitative measures. **Confusion matrices** captured per-bin prediction fidelity, highlighting both central and extreme kinetic ranges, while **t-SNE projections** of concatenated embeddings overlaid with predicted versus true classes illustrated preservation of the global kinetic parameter space. All metrics, including accuracy,

precision, recall, F1-score, e-accuracy, MCC, and AUC-PR, were implemented in Python using the *scikit-learn* (sklearn) library.

Mutation-Aware Validation Datasets

To assess mutation sensitivity, we curated three complementary datasets. The first comprised 22 experimentally characterized wild-type and alanine-substituted pairs at UniProt-annotated catalytic residues, withheld entirely from training and used as an out-of-distribution (OOD) anchor set. The second comprised a synthetic hold-out of 10 alanine-substituted variants randomly drawn from the partitioned negative pool; their corresponding wild-type sequences were included in the positive training set, but the negative variants themselves were withheld, yielding an in-distribution-by-identity benchmark. For both sets, mutation effects were additionally summarized using binary meta-labels (k_{cat} drop vs no-drop, K_M increase vs no-increase) derived from mechanistic descriptors. An additional set of 10 non-alanine OOD mutants (≤ 80 % sequence identity relative to the training partition) was further used to confirm that performance generalizes beyond alanine substitutions (see Fig. S4). Finally, we employed the high-throughput alkaline phosphatase PafA from *Elizabethkingia meningoseptica* mutational dataset, consisting of 1,016 single-point variants assayed with carboxy 4-methylumbelliferyl phosphate ester (cMUP) substrate (*29*). This dataset was partitioned into 398 training, 310 validation, and 310 test variants, with WT and one catalytic variant (T79S) included in training, while catalytic-site substitutions at residue R164 (R164A, R164G, R164V) were withheld in the test/validation sets. Performance was quantified using e-accuracy, MCC, and meta-label classification metrics as detailed under *Performance and Evaluation Metrics*.

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 curation, 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.

Author contributions:

Conceptualization: R.C., R.S., A.M., A.O., K.A.S.
Methodology: R.C., A.O., K.A.S., A.M., A.B., 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., R.A., 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.

Funding:

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

NSF CAREER Award 1943310 (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 is available for download and inference at the link: https://chowdhurylab.github.io/downloads.html. Moreover, 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.

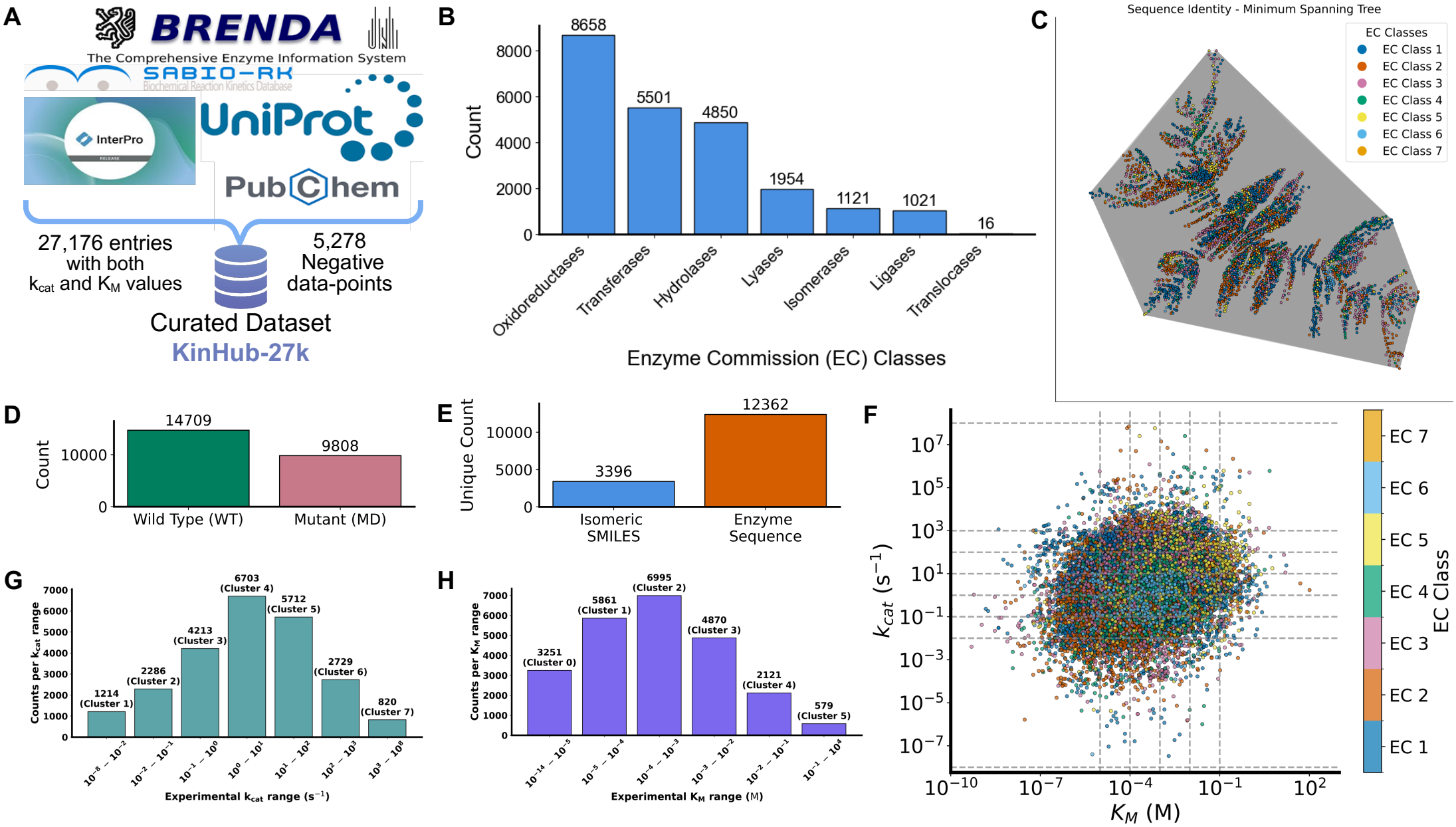
References

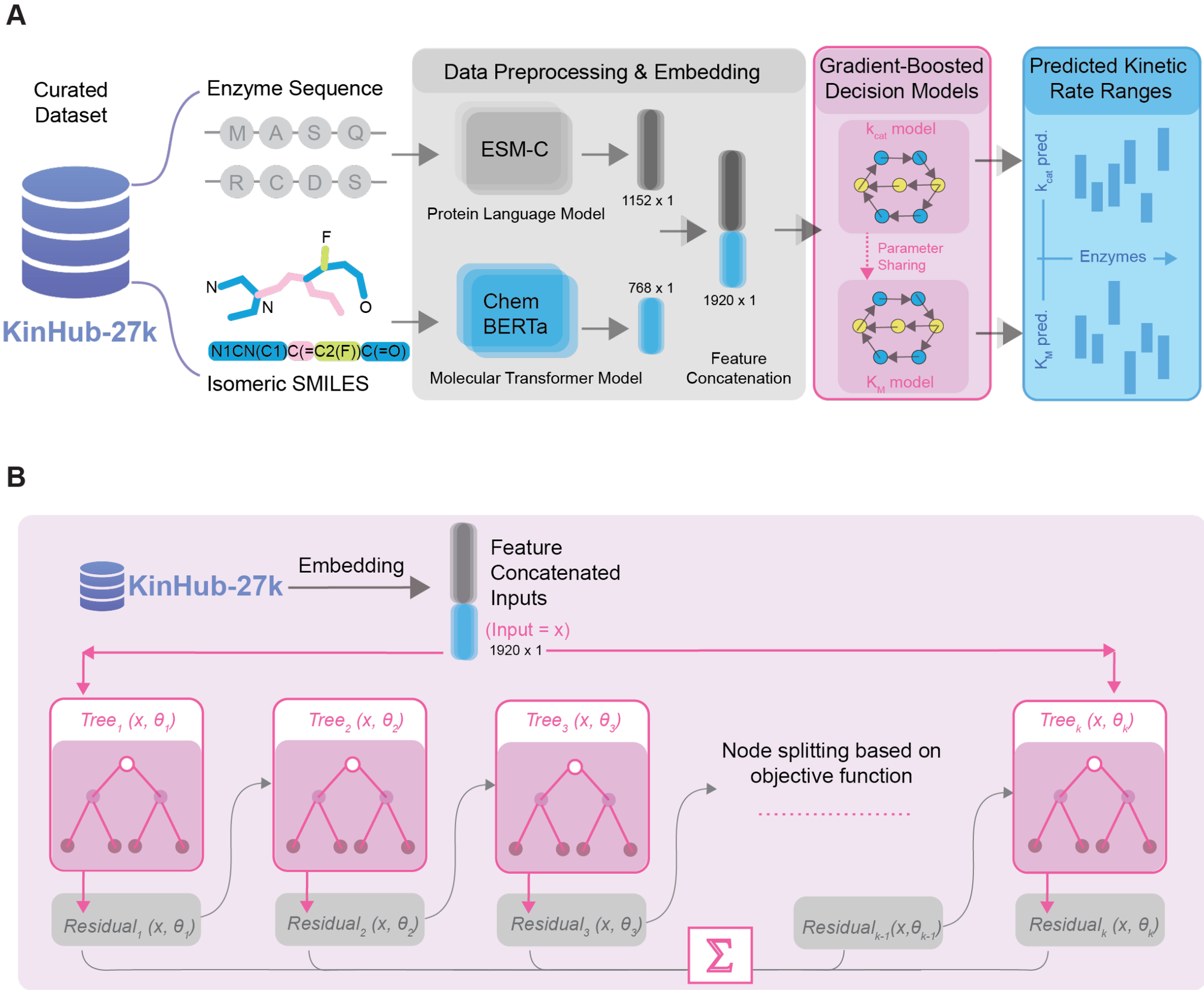
1. D. Davidi, R. Milo, Lessons on enzyme kinetics from quantitative proteomics. *Curr Opin Biotechnol* **46**, 81–89 (2017).
2. G. A. Holdgate, T. D. Meek, R. L. Grimley, Mechanistic enzymology in drug discovery: a fresh perspective. *Nature Reviews Drug Discovery 2017 17:2* **17**, 115–132 (2017).
3. 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. A. Kroll, M. J. Lercher, DLKcat cannot predict meaningful kcat values for mutants and unfamiliar enzymes. *Biol Methods Protoc* **9** (2024).
6. 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).
7. 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).
8. 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).
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

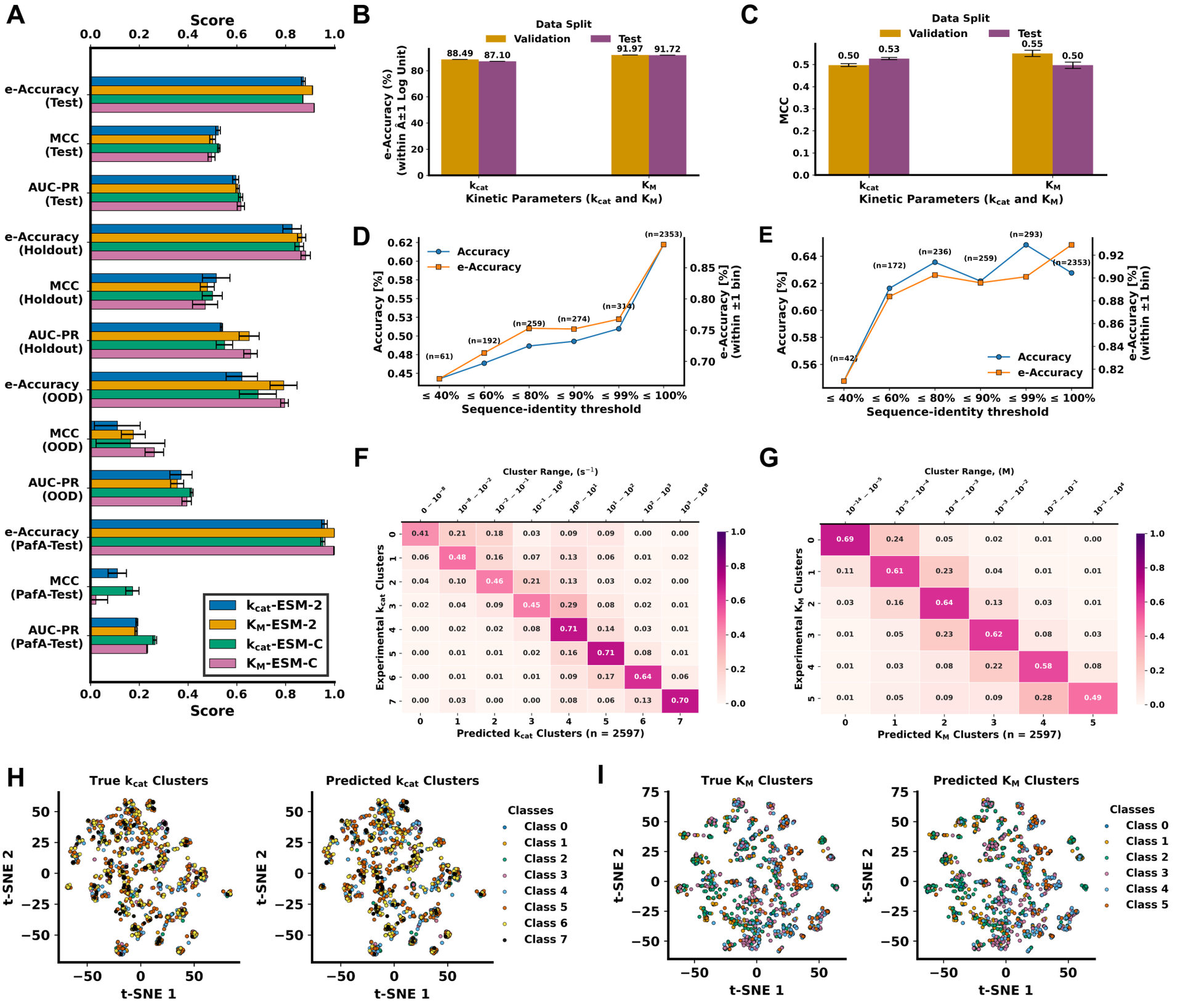
reveals protein structural correlates and improves metabolic models. *Nature Communications 2018 9:1* **9**, 1–10 (2018).
10. V. S. Boorla, C. D. Maranas, CatPred: a comprehensive framework for deep learning in vitro enzyme kinetic parameters. *Nature Communications 2025 16:1* **16**, 1–17 (2025).
11. 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).
12. 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).
13. D. Dudaš, U. Wittig, M. Rey, A. Weidemann, W. Müller, Improved insights into the SABIO-RK database via visualization. *Database* **2023**, 1–9 (2023).
14. 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).
15. 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).
16. 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).
17. 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).
18. S. Chithrananda, G. Grand, B. R. Deepchem, ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. (2020).
19. 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).

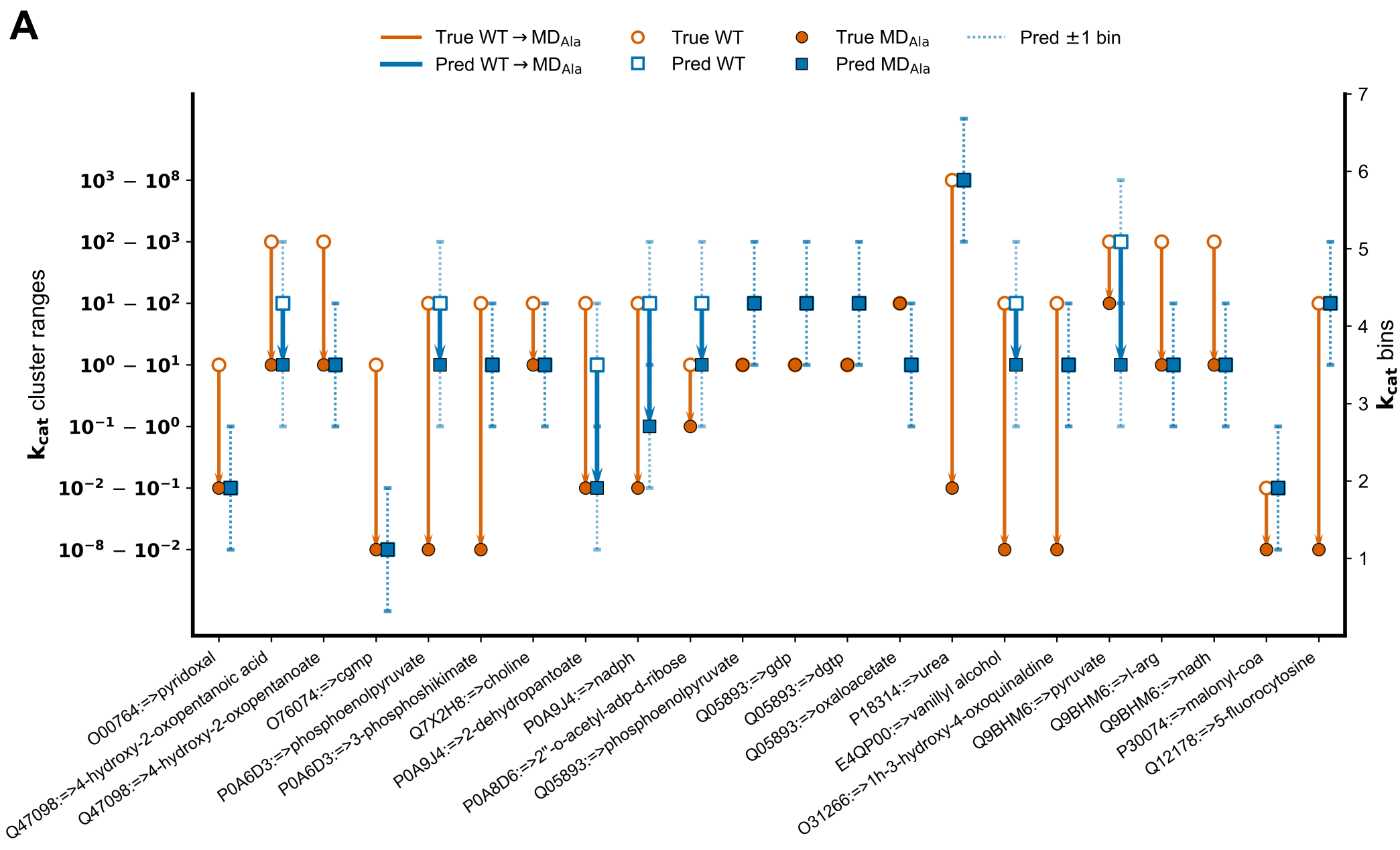
20. Y. Chen, F. Li, J. Nielsen, Genome-scale modeling of yeast metabolism: retrospectives and perspectives. *FEMS Yeast Res* **22** (2022).
21. H. Lu, F. Li, B. J. Sánchez, Z. Zhu, G. Li, I. Domenzain, S. Marcišauskas, P. M. Anton, D. Lappa, C. Lieven, M. E. Beber, N. Sonnenschein, E. J. Kerkhoven, J. 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).
22. A. Nilsson, J. Nielsen, B. O. Palsson, Metabolic Models of Protein Allocation Call for the Kinetome. *Cell Syst* **5**, 538–541 (2017).
23. X. S. Xie, Enzyme kinetics, past and present. *Science (1979)* **342**, 1457–1459 (2013).
24. 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).
25. ESM Team, Evolutionary Scale · ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning, *EvolutionaryScale Website* (2024). https://www.evolutionaryscale.ai/blog/esm-cambrian.
26. 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).
27. A. Abdel-Rehim, O. Orhobor, L. Hang, H. Ni, R. D. King, Protein-ligand binding affinity prediction exploiting sequence constituent homology. *Bioinformatics* **39** (2023).
28. 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).
29. 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).
30. D. Ringe, G. P.- Science, undefined 2008, How enzymes work. *science.orgD Ringe, GA PetskoScience, 2008•science.org* **320**, 1428–1429 (2008).
31. 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).
32. J. M. Woodley, Accelerating the implementation of biocatalysis in industry. *Appl Microbiol Biotechnol* **103**, 4733–4739 (2019).
33. 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).
34. M. A. A. Adam, C. D. Sohl, Probing altered enzyme activity in the biochemical characterization of cancer. *Biosci Rep* **42**, BSR20212002 (2022).
35. 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).
36. 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).

37. 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).
38. D. M. Blei, A. Kucukelbir, J. D. McAuliffe, Variational Inference: A Review for Statisticians. *J Am Stat Assoc* **112**, 859–877 (2017).
39. 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).
40. 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).
41. S. Ferdous, P. Gupta, Y. C. Teoh, K. A. Sajeevan, S. Vaishnavey, K. Hainlin, N. Reuel, A. Qamhiyah, N. R. Aluru, M. Kumar, R. Chowdhury, Improved Functional Classification of Hydrolases through Pairwise Structural Similarity of Reaction Cores. *bioRxiv*, 2025.01.09.632218 (2025).
42. 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).
43. N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. Bin Hoo, R. T. Schirrmeister, F. Hutter, Accurate predictions on small data with a tabular foundation model. *Nature 2025 637:8045* **637**, 319–326 (2025).


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
5,278 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
Sequence Identity - Minimum Spanning Tree
EC Classes
EC Class 1
EC Class 2
EC Class 3
EC Class 4
EC Class 5
EC Class 6
EC Class 7
D
Count
14709
9808
Wild Type (WT)
Mutant (MD)
E
Unique Count
3396
12362
Isomeric SMILES
Enzyme Sequence
F
k_{cat} (s^{-1})
K_M (M)
EC Class
EC 7
EC 6
EC 5
EC 4
EC 3
EC 2
EC 1
G
Counts per k_{cat} range
1214 (Cluster 1)
2286 (Cluster 2)
4213 (Cluster 3)
6703 (Cluster 4)
5712 (Cluster 5)
2729 (Cluster 6)
820 (Cluster 7)
Experimental k_{cat} range (s^{-1})
H
Counts per K_M range
3251 (Cluster 0)
5861 (Cluster 1)
6995 (Cluster 2)
4870 (Cluster 3)
2121 (Cluster 4)
579 (Cluster 5)
Experimental K_M range (M)


A
Curated Dataset
KinHub-27k
Enzyme Sequence
M A S Q
R C D S
F
N
N
O
N1CN(C1)C(=C2(F))C(=O)
Isomeric SMILES
Data Preprocessing & Embedding
ESM-C
Protein Language Model
1152 x 1
Chem BERTa
Molecular Transformer Model
768 x 1
1920 x 1
Feature Concatenation
Gradient-Boosted Decision Models
kcat model
Parameter Sharing
KM model
Predicted Kinetic Rate Ranges
kcat pred.
Enzymes
KM pred.
B
KinHub-27k
Embedding
Feature Concatenated Inputs
(Input = x)
1920 x 1
Tree1 (x, θ1)
Tree2 (x, θ2)
Tree3 (x, θ3)
Treek (x, θk)
Node splitting based on objective function
Residual1 (x, θ1)
Residual2 (x, θ2)
Residual3 (x, θ3)
Residualk-1(x,θk-1)
Residualk (x, θk)
Σ


A
Score
e-Accuracy (Test)
MCC (Test)
AUC-PR (Test)
e-Accuracy (Holdout)
MCC (Holdout)
AUC-PR (Holdout)
e-Accuracy (OOD)
MCC (OOD)
AUC-PR (OOD)
e-Accuracy (PafA-Test)
MCC (PafA-Test)
AUC-PR (PafA-Test)
kcat-ESM-2
KM-ESM-2
kcat-ESM-C
KM-ESM-C
B
Data Split
Validation
Test
e-Accuracy (%) (within Â±1 Log Unit)
88.49
87.10
91.97
91.72
Kinetic Parameters (kcat and KM)
C
MCC
0.50
0.53
0.55
0.50
D
Accuracy
e-Accuracy
Accuracy [%]
e-Accuracy [%] (within ±1 bin)
(n=61)
(n=192)
(n=259)
(n=274)
(n=314)
(n=2353)
≤ 40%
≤ 60%
≤ 80%
≤ 90%
≤ 99%
≤ 100%
Sequence-identity threshold
E
(n=42)
(n=172)
(n=236)
(n=259)
(n=293)
(n=2353)
F
Cluster Range, (s⁻¹)
Experimental kcat Clusters
Predicted kcat Clusters (n = 2597)
G
Cluster Range, (M)
Experimental KM Clusters
Predicted KM Clusters (n = 2597)
H
True kcat Clusters
Predicted kcat Clusters
t-SNE 1
t-SNE 2
Classes
Class 0
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
I
True KM Clusters
Predicted KM Clusters


A
True WT → MD$_{Ala}$
Pred WT → MD$_{Ala}$
True WT
Pred WT
True MD$_{Ala}$
Pred MD$_{Ala}$
Pred ±1 bin
k_{cat} cluster ranges
10^{3} – 10^{8}
10^{2} – 10^{3}
10^{1} – 10^{2}
10^{0} – 10^{1}
10^{-1} – 10^{0}
10^{-2} – 10^{-1}
10^{-8} – 10^{-2}
k_{cat} bins
O00764:=>pyridoxal
Q47098:=>4-hydroxy-2-oxopentanoic acid
Q47098:=>4-hydroxy-2-oxopentanoate
O76074:=>cgmp
P0A6D3:=>phosphoenolpyruvate
P0A6D3:=>3-phosphoshikimate
Q7X2H8:=>choline
P0A9J4:=>2-dehydropantoate
P0A9J4:=>nadph
P0A8D6:=>2'-o-acetyl-adp-d-ribose
Q05893:=>phosphoenolpyruvate
Q05893:=>gdp
Q05893:=>dgtp
Q05893:=>oxaloacetate
P18314:=>urea
E4QP00:=>vanillyl alcohol
O31266:=>1h-3-hydroxy-4-oxoquinaldine
Q9BHM6:=>pyruvate
Q9BHM6:=>l-arg
Q9BHM6:=>nadh
P30074:=>malonyl-coa
Q12178:=>5-fluorocytosine
Enzyme-Substrate pairs
B
True WT → MD$_{Ala}$
Pred WT → MD$_{Ala}$
True WT
Pred WT
True MD$_{Ala}$
Pred MD$_{Ala}$
Pred ±1 bin
K_M cluster ranges
10^{-2} – 10^{-1}
10^{-3} – 10^{-2}
10^{-4} – 10^{-3}
10^{-5} – 10^{-4}
10^{-14} – 10^{-5}
K_M bins
Enzyme-Substrate pairs

A

	Main: DROP	Main: NO-DROP	Row total
Meta: DROP	TP 7 (70%)	FN 2 (20%)	9 (90%)
Meta: NO-DROP	FP 0 (0%)	TN 1 (10%)	1 (10%)

Precision=1.00, Recall=0.78, F1=0.87, MCC=0.51, tau=2 bins

Case	WT True	WT Pred	MD Pred	Δ bin	Shift	Meta Label	Predicted (WT-MD)	Match
1	2	2	0	2	↓ 2	DROP	DROP	YES
2	3	3	3	0	0	NO-DROP	NO-DROP	YES
3	2	5	1	4	↓ 4	DROP	DROP	YES
4	4	4	1	3	↓ 3	DROP	DROP	YES
5	5	5	5	0	0	DROP	NO-DROP	NO
6	5	5	2	3	↓ 3	DROP	DROP	YES
7	6	6	6	0	0	DROP	NO-DROP	NO
8	2	2	0	2	↓ 2	DROP	DROP	YES
9	2	3	0	3	↓ 3	DROP	DROP	YES
10	6	6	4	2	↓ 2	DROP	DROP	YES

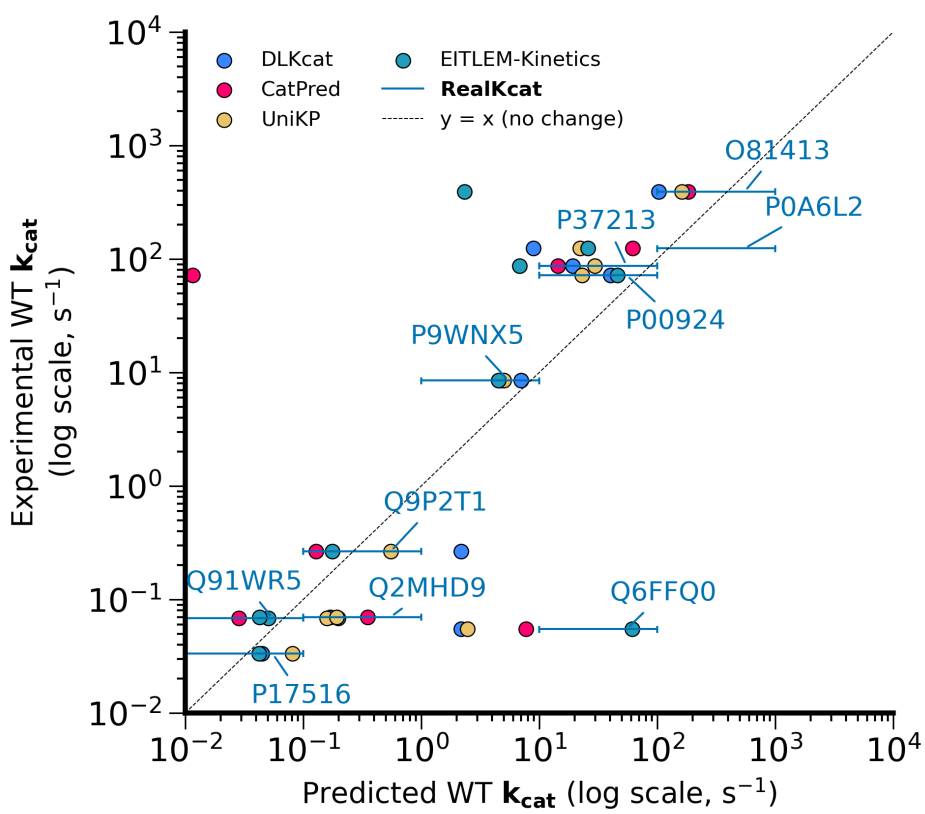
B

	Main: INCREASE	Main: NO-INCREASE	Row total
Meta: INCREASE	TP 2 (20%)	FN 3 (30%)	5 (50%)
Meta: NO-INCREASE	FP 0 (0%)	TN 5 (50%)	5 (50%)

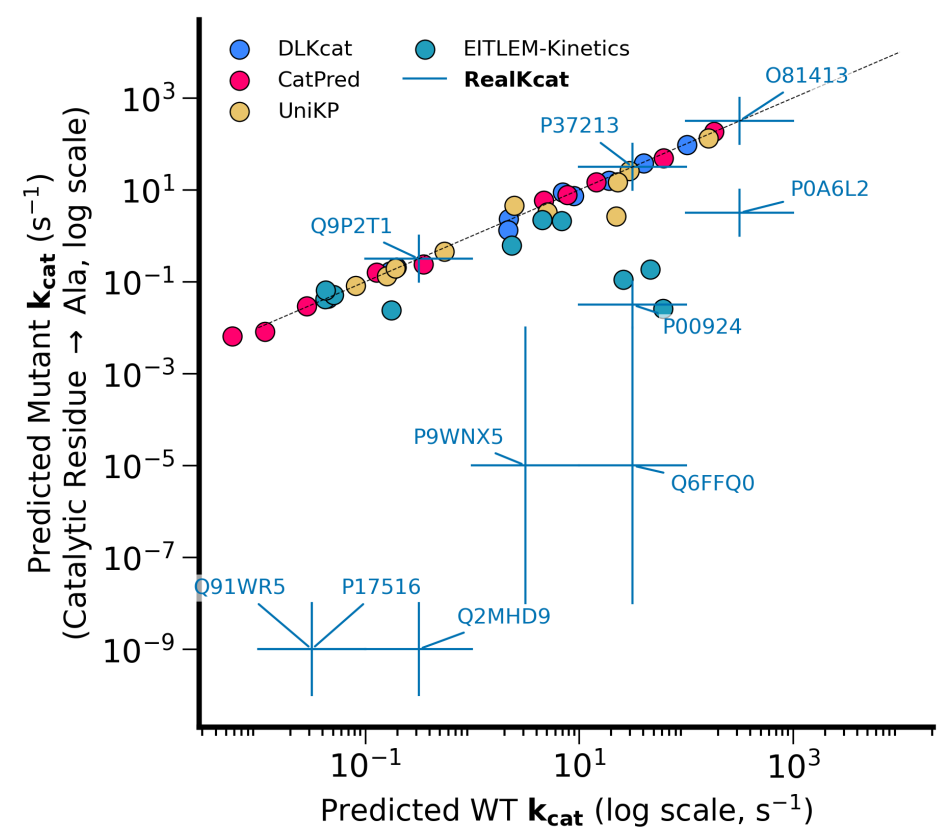
Precision=1.00, Recall=0.40, F1=0.57, MCC=0.50, tau=-1 bins

Case	WT True	WT Pred	MD Pred	Δ bin	Shift	Meta Label	Predicted (WT-MD)	Match
1	0	0	0	0	0	NO-INCREASE	NO-INCREASE	YES
2	1	1	1	0	0	INCREASE	NO-INCREASE	NO
3	2	2	5	-3	↑ 3	INCREASE	INCREASE	YES
4	1	1	1	0	0	NO-INCREASE	NO-INCREASE	YES
5	1	1	1	0	0	INCREASE	NO-INCREASE	NO
6	1	1	1	0	0	INCREASE	NO-INCREASE	NO
7	3	3	3	0	0	NO-INCREASE	NO-INCREASE	YES
8	0	0	0	0	0	NO-INCREASE	NO-INCREASE	YES
9	1	1	1	0	0	NO-INCREASE	NO-INCREASE	YES
10	2	2	5	-3	↑ 3	INCREASE	INCREASE	YES

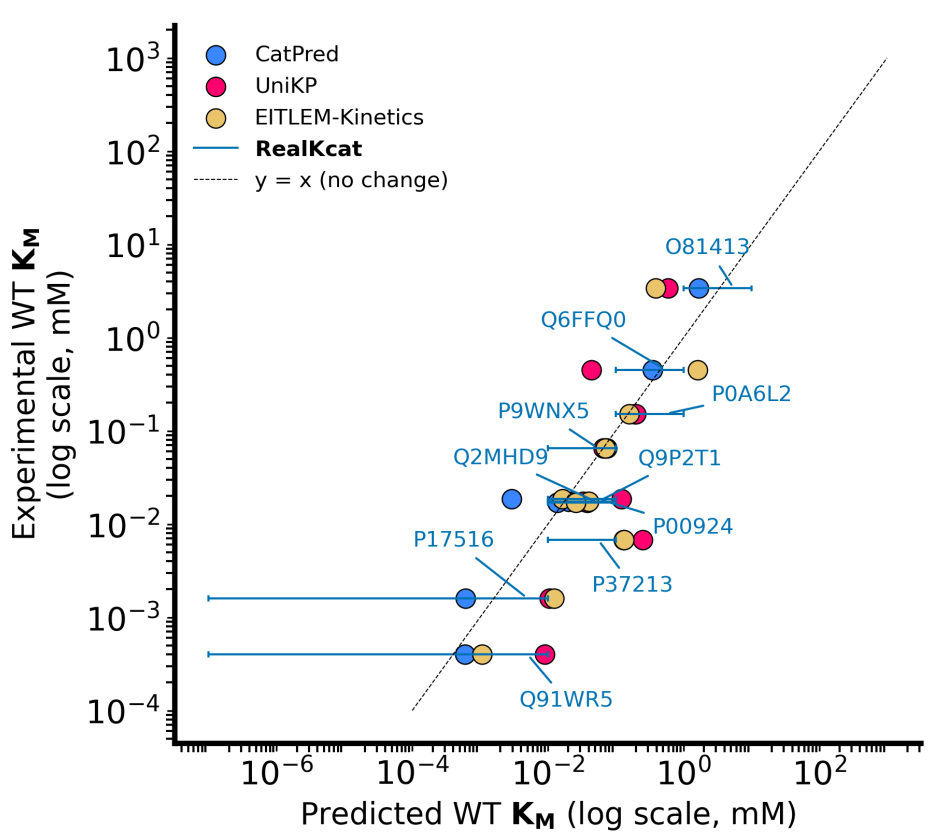
C



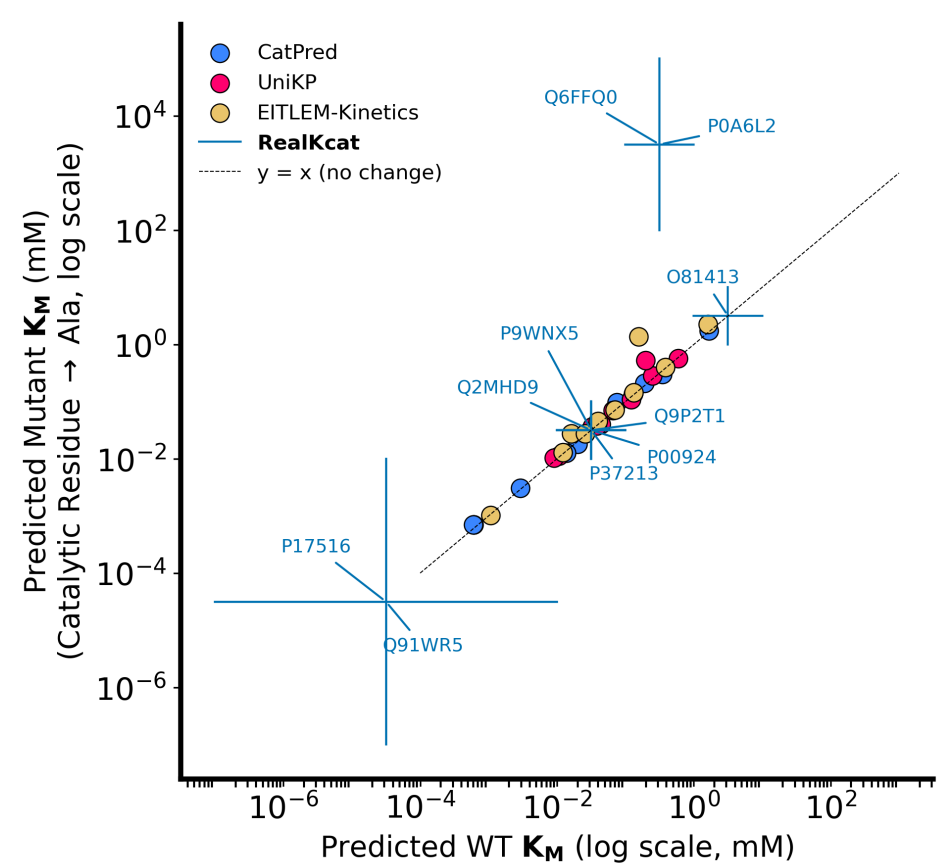
D

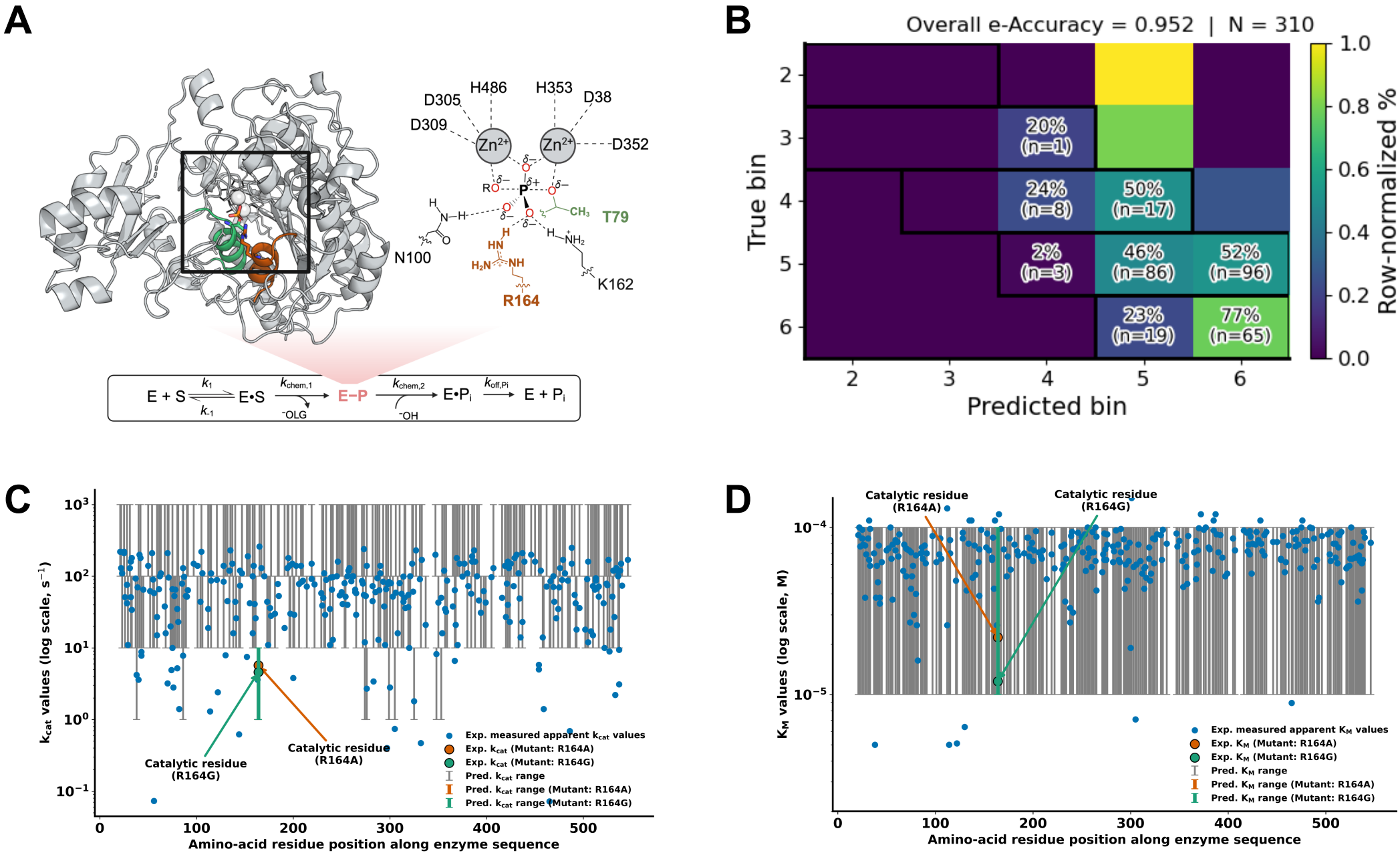


E



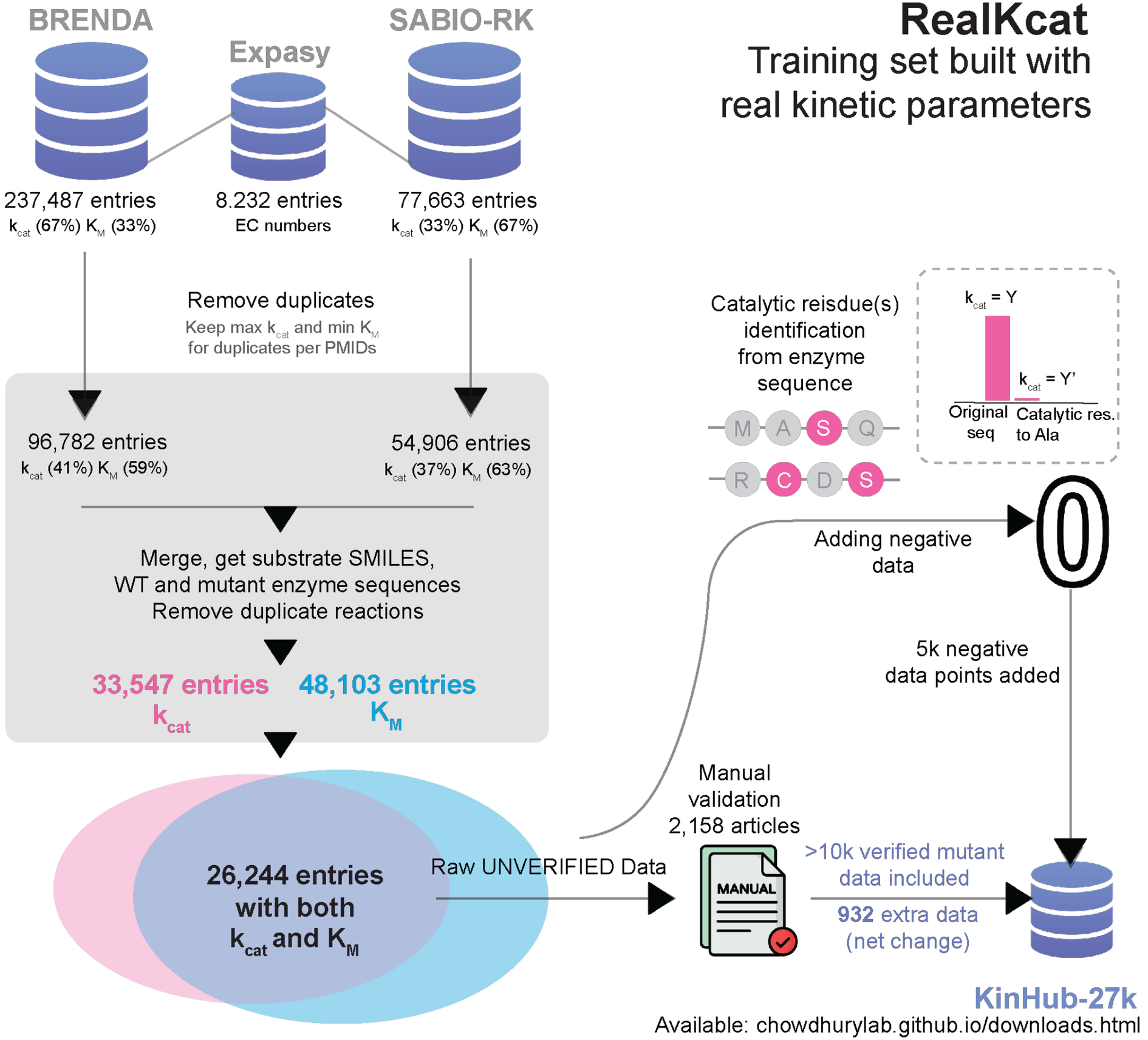
F



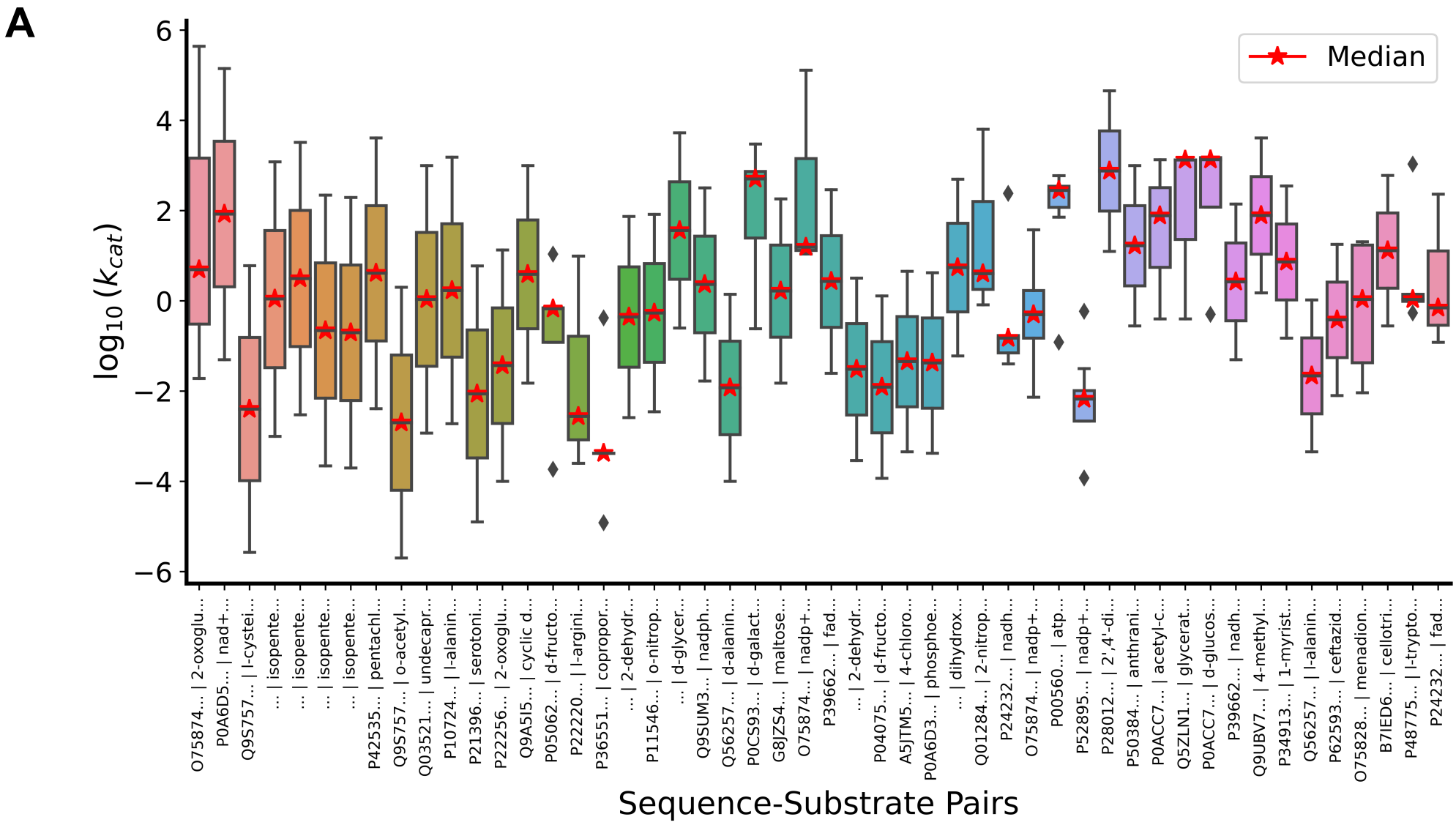

A
D305
H486
H353
D38
D309
D352
T79
N100
K162
R164
E + S
E•S
E–P
E•Pi
E + Pi
-OLG
-OH
B
Overall e-Accuracy = 0.952 | N = 310
True bin
Predicted bin
Row-normalized %
20% (n=1)
24% (n=8)
50% (n=17)
2% (n=3)
46% (n=86)
52% (n=96)
23% (n=19)
77% (n=65)
C
kcat values (log scale, s−1)
Catalytic residue (R164G)
Catalytic residue (R164A)
Exp. measured apparent kcat values
Exp. kcat (Mutant: R164A)
Exp. kcat (Mutant: R164G)
Pred. kcat range
Pred. kcat range (Mutant: R164A)
Pred. kcat range (Mutant: R164G)
Amino-acid residue position along enzyme sequence
D
KM values (log scale, M)
Catalytic residue (R164A)
Catalytic residue (R164G)
Exp. measured apparent KM values
Exp. KM (Mutant: R164A)
Exp. KM (Mutant: R164G)
Pred. KM range
Pred. KM range (Mutant: R164A)
Pred. KM range (Mutant: R164G)
Amino-acid residue position along enzyme sequence

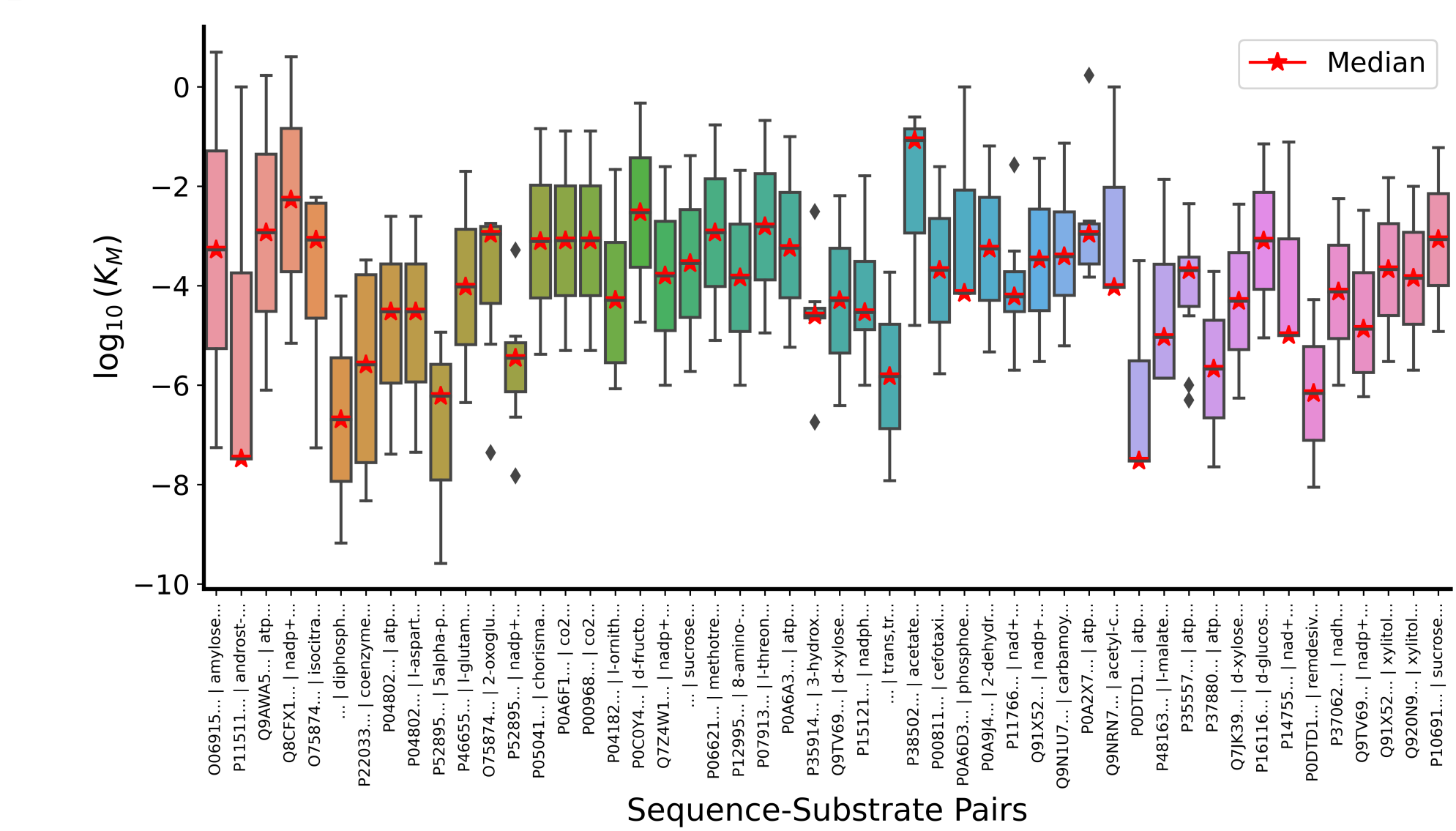
RealKcat

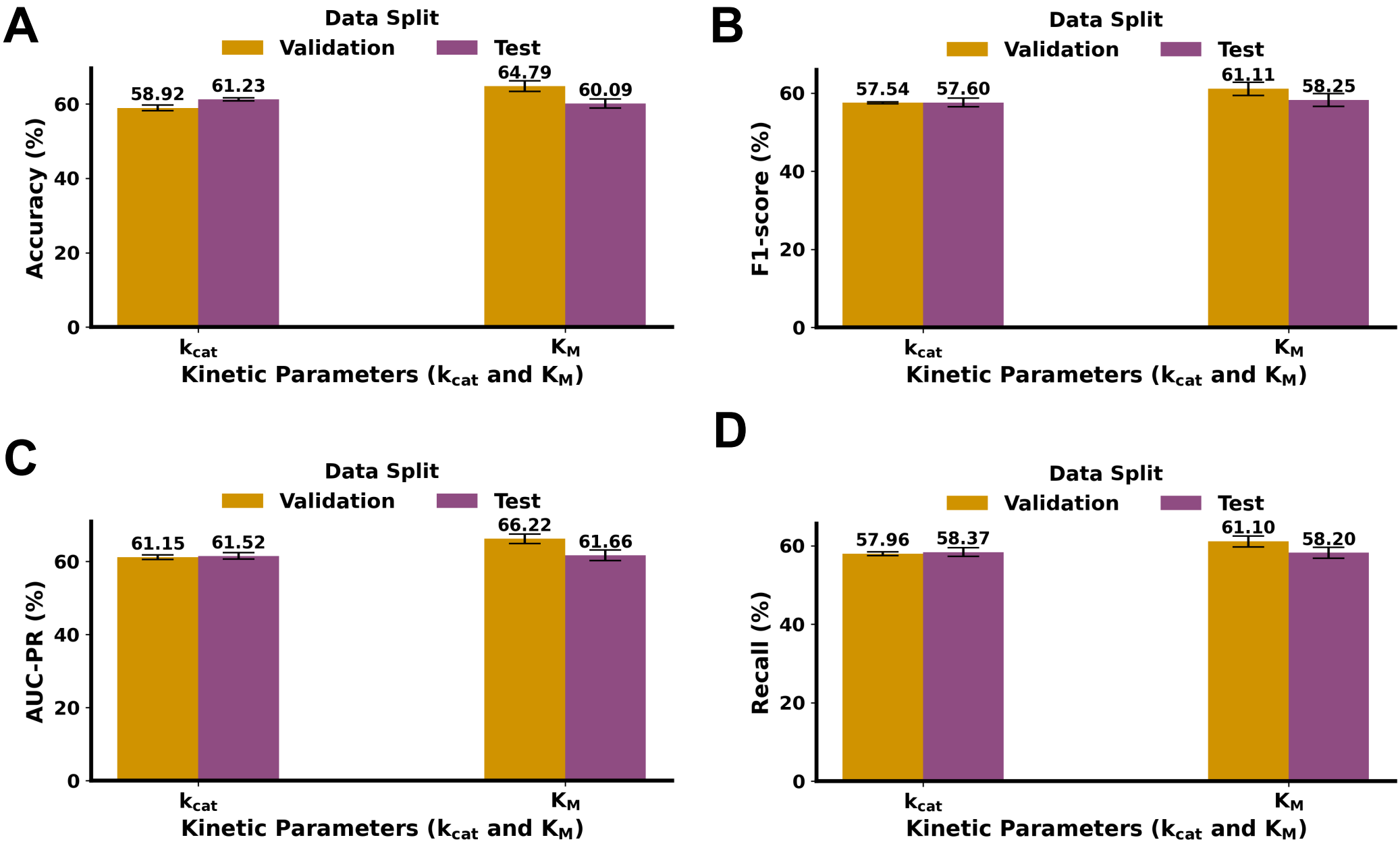
Training set built with real kinetic parameters

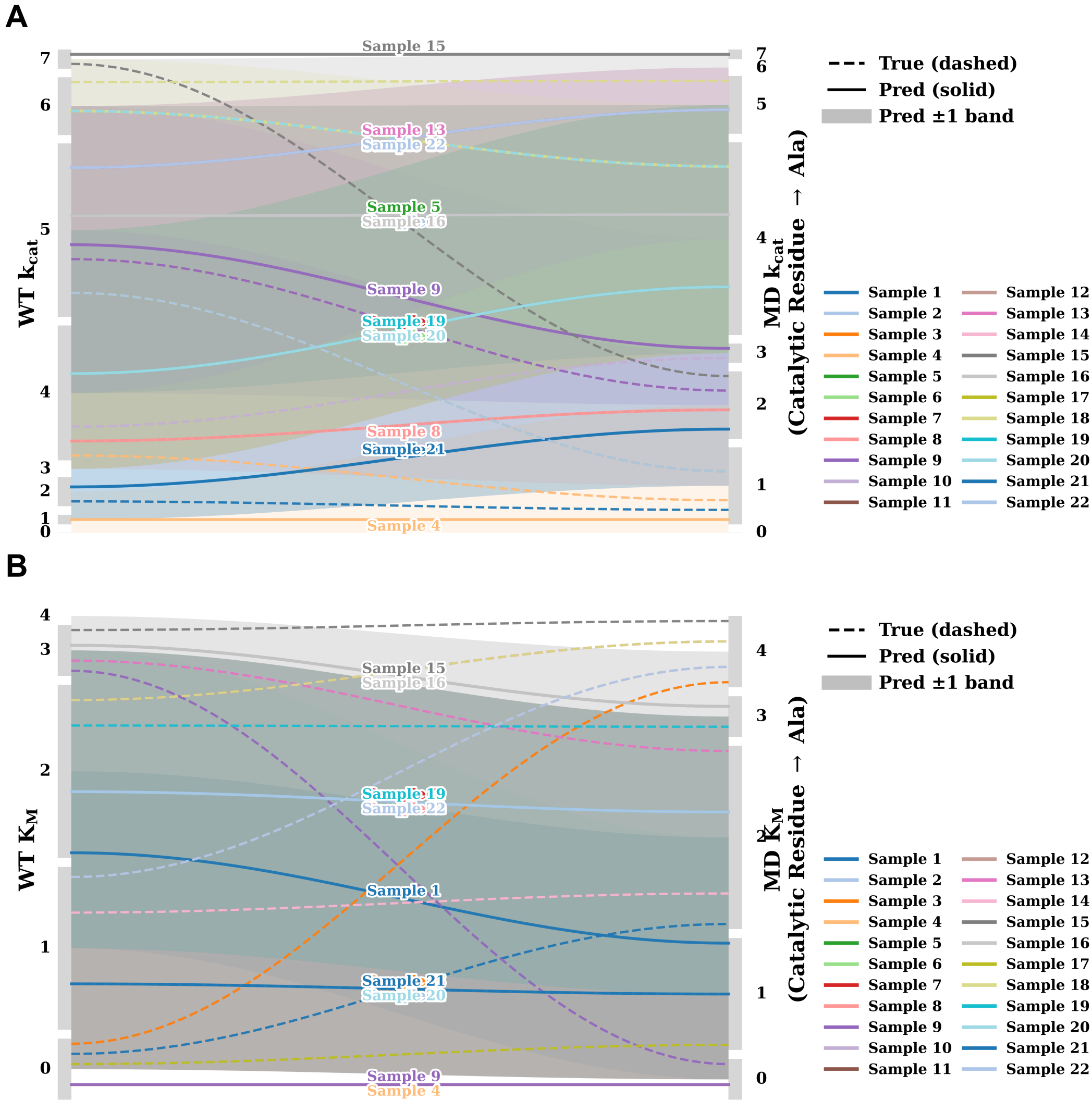


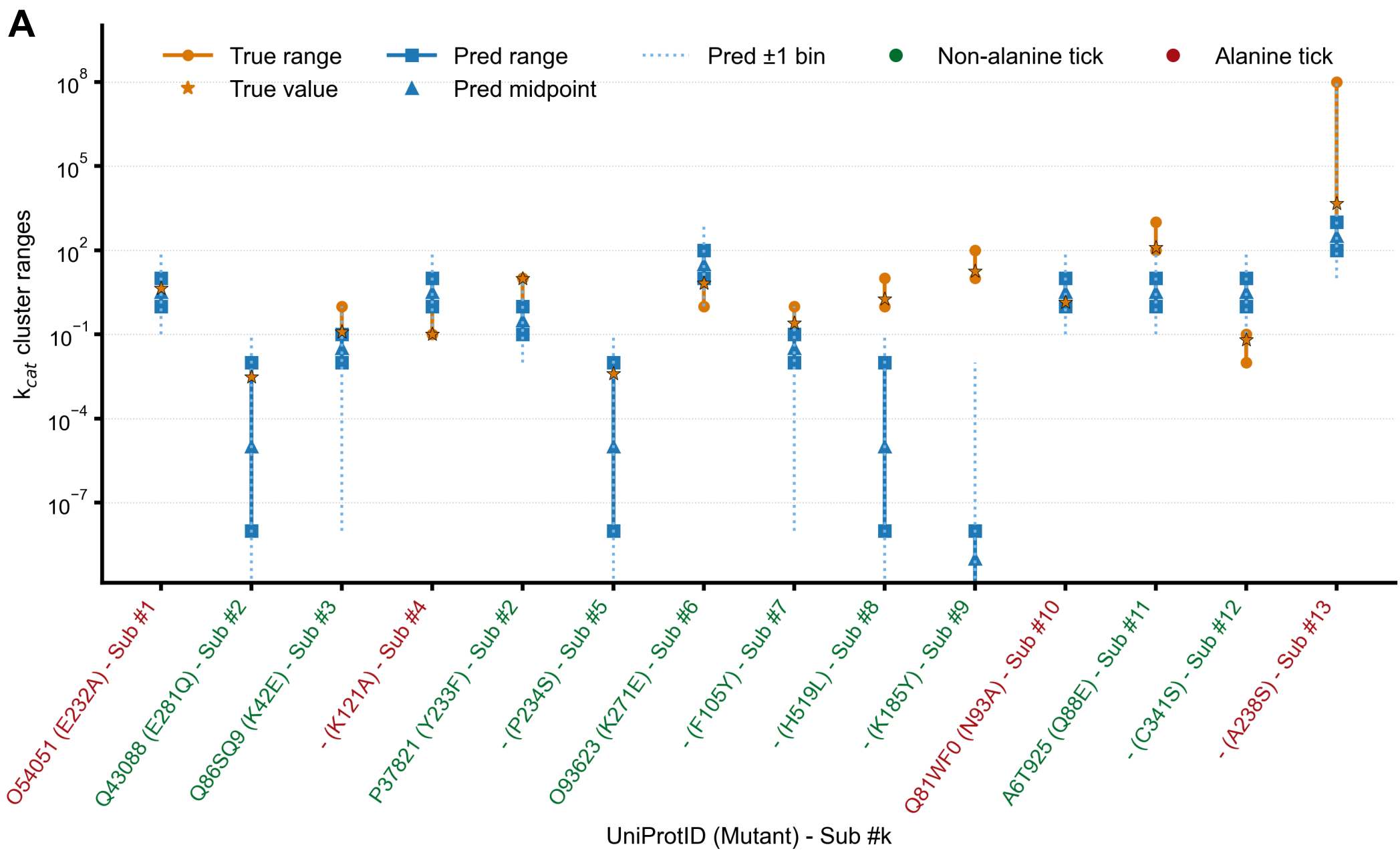
Available: chowdhurylab.github.io/downloads.html

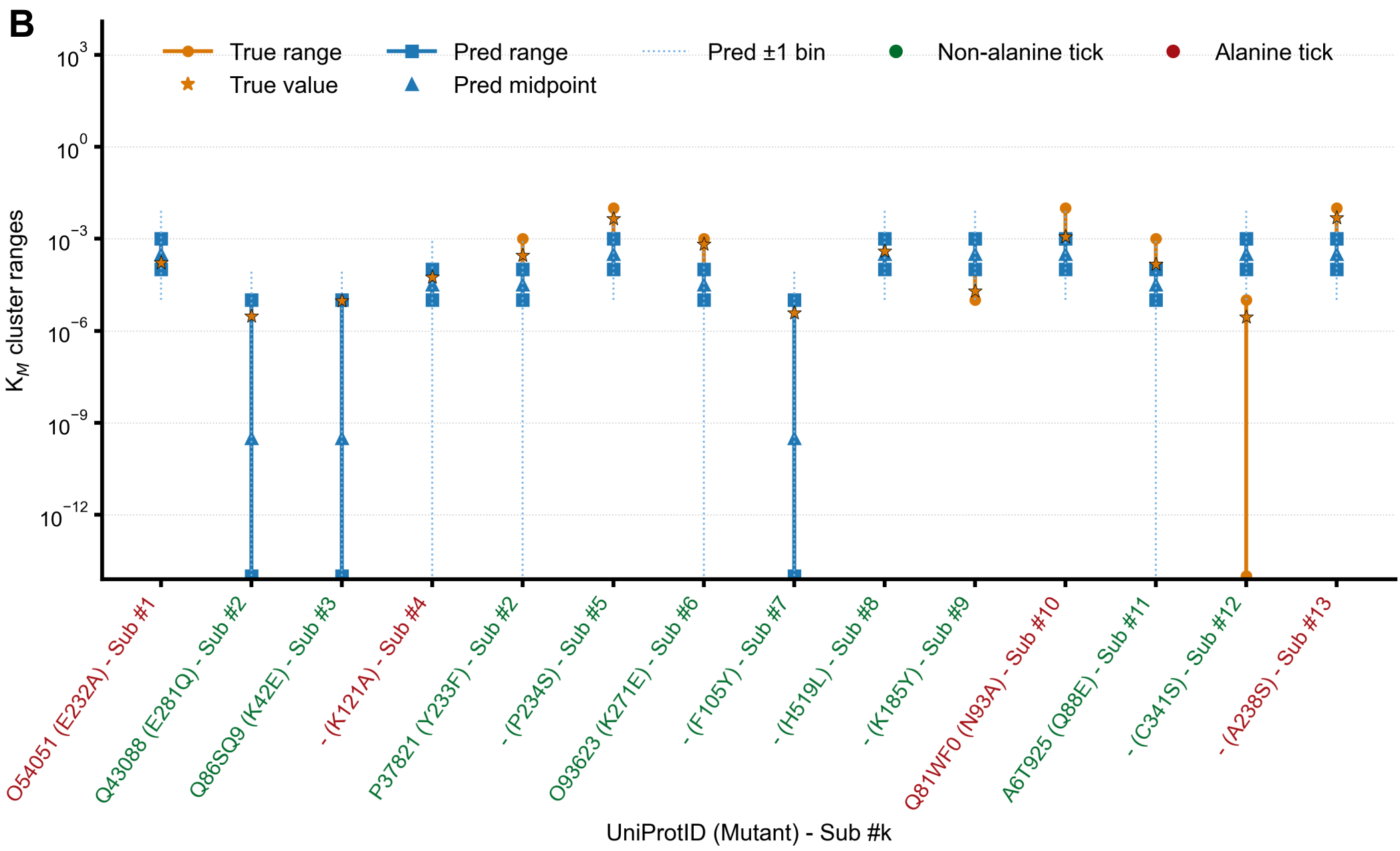

A
Median
$\log_{10}(k_{cat})$
6
4
2
0
−2
−4
−6
O75874... | 2-oxoglu...
P0A6D5... | nad+...
Q9S757... | l-cystei...
... | isopente...
... | isopente...
... | isopente...
... | isopente...
P42535... | pentachl...
Q9S757... | o-acetyl...
Q03521... | undecapr...
P10724... | l-alanin...
P21396... | serotoni...
P22256... | 2-oxoglu...
Q9A5I5... | cyclic d...
P05062... | d-fructo...
P22220... | l-argini...
P36551... | copropor...
... | 2-dehydr...
P11546... | o-nitrop...
... | d-glycer...
Q9SUM3... | nadph...
Q56257... | d-alanin...
P0CS93... | d-galact...
G8JZS4... | maltose...
O75874... | nadp+...
P39662... | fad...
... | 2-dehydr...
P04075... | d-fructo...
A5JTM5... | 4-chloro...
P0A6D3... | phosphoe...
... | dihydrox...
Q01284... | 2-nitrop...
P24232... | nadh...
O75874... | nadp+...
P00560... | atp...
P52895... | nadp+...
P28012... | 2',4'-di...
P50384... | anthrani...
P0ACC7... | acetyl-c...
Q5ZLN1... | glycerat...
P0ACC7... | d-glucos...
P39662... | nadh...
Q9UBV7... | 4-methyl...
P34913... | 1-myrist...
Q56257... | l-alanin...
P62593... | ceftazid...
O75828... | menadion...
B7IED6... | cellotri...
P48775... | l-trypto...
P24232... | fad...
Sequence-Substrate Pairs


B
Median
$\log_{10}(K_M)$
0
−2
−4
−6
−8
−10
O06915... | amylose...
P11511... | androst-...
Q9AWA5... | atp...
Q8CFX1... | nadp+...
O75874... | isocitra...
... | diphosph...
P22033... | coenzyme...
P04802... | atp...
P04802... | l-aspart...
P52895... | 5alpha-p...
P46655... | l-glutam...
O75874... | 2-oxoglu...
P52895... | nadp+...
P05041... | chorisma...
P0A6F1... | co2...
P00968... | co2...
P04182... | l-ornith...
P0C0Y4... | d-fructo...
Q7Z4W1... | nadp+...
... | sucrose...
P06621... | methotre...
P12995... | 8-amino-...
P07913... | l-threon...
P0A6A3... | atp...
P35914... | 3-hydrox...
Q9TV69... | d-xylose...
P15121... | nadph...
... | trans,tr...
P38502... | acetate...
P00811... | cefotaxi...
P0A6D3... | phosphoe...
P0A9J4... | 2-dehydr...
P11766... | nad+...
Q91X52... | nadp+...
Q9N1U7... | carbamoy...
P0A2X7... | atp...
Q9NRN7... | acetyl-c...
P0DTD1... | atp...
P48163... | l-malate...
P35557... | atp...
P37880... | atp...
Q7JK39... | d-xylose...
P16116... | d-glucos...
P14755... | nad+...
P0DTD1... | remdesiv...
P37062... | nadh...
Q9TV69... | nadp+...
Q91X52... | xylitol...
Q920N9... | xylitol...
P10691... | sucrose...
Sequence-Substrate Pairs


A
Data Split
Validation
Test
Accuracy (%)
58.92
61.23
64.79
60.09
0
20
40
60
k_{cat}
K_M
Kinetic Parameters (k_{cat} and K_M)
B
Data Split
Validation
Test
F1-score (%)
57.54
57.60
61.11
58.25
0
20
40
60
k_{cat}
K_M
Kinetic Parameters (k_{cat} and K_M)
C
Data Split
Validation
Test
AUC-PR (%)
61.15
61.52
66.22
61.66
0
20
40
60
k_{cat}
K_M
Kinetic Parameters (k_{cat} and K_M)
D
Data Split
Validation
Test
Recall (%)
57.96
58.37
61.10
58.20
0
20
40
60
k_{cat}
K_M
Kinetic Parameters (k_{cat} and K_M)

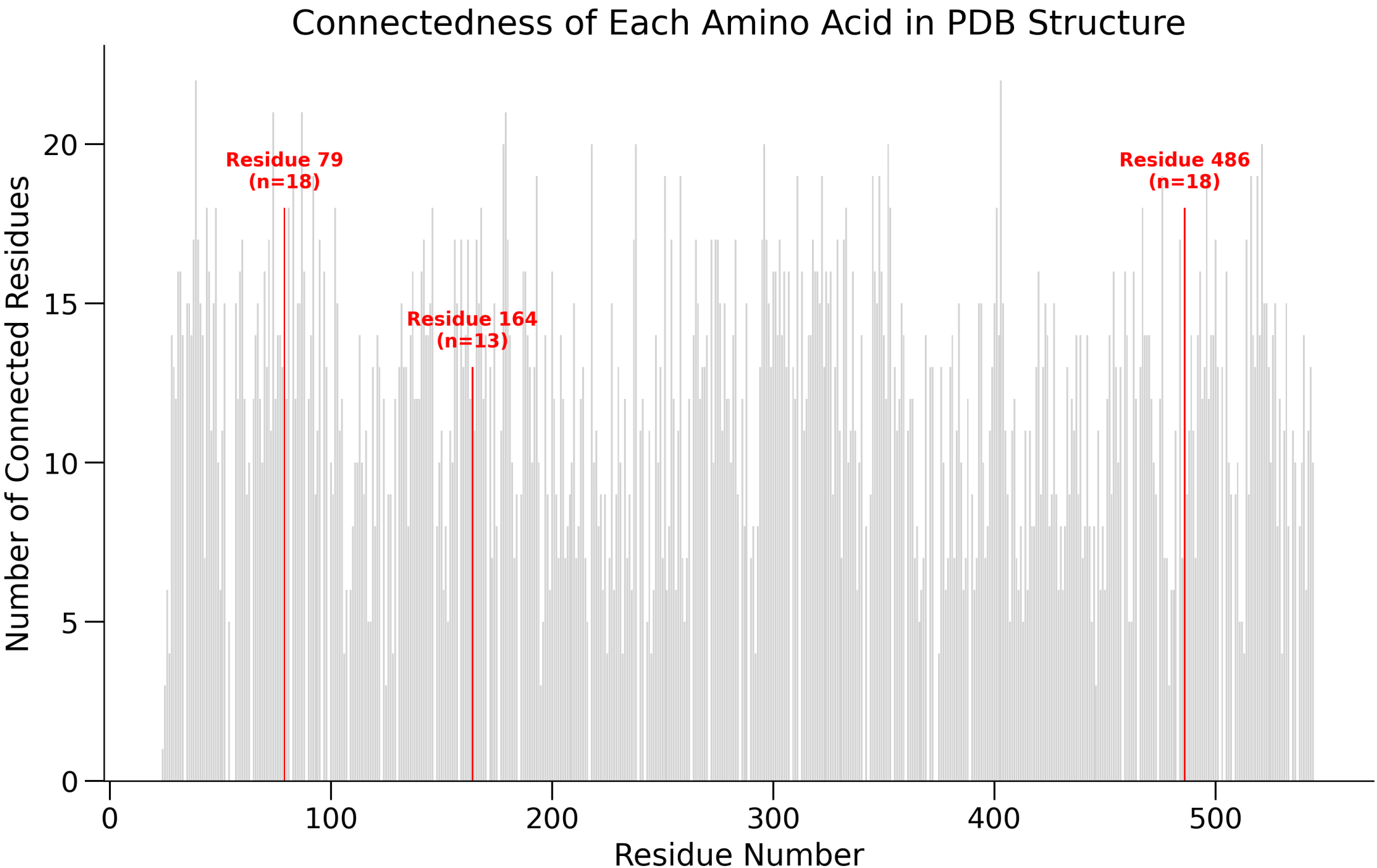

A
WT k_{cat}
MD k_{cat} (Catalytic Residue → Ala)
True (dashed)
Pred (solid)
Pred ±1 band
Sample 15
Sample 13
Sample 22
Sample 5
Sample 16
Sample 9
Sample 19
Sample 20
Sample 8
Sample 21
Sample 4
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
Sample 9
Sample 10
Sample 11
Sample 12
Sample 13
Sample 14
Sample 15
Sample 16
Sample 17
Sample 18
Sample 19
Sample 20
Sample 21
Sample 22
B
WT K_M
MD K_M (Catalytic Residue → Ala)
True (dashed)
Pred (solid)
Pred ±1 band
Sample 15
Sample 16
Sample 19
Sample 22
Sample 1
Sample 21
Sample 20
Sample 9
Sample 4





Substrates

Sub #1: nad+
Sub #2: s-adenosyl-l-methionine
Sub #3: trans,trans-farnesyl diphosphate
Sub #4: (2s,3r)-3-hydroxybutane-1,2,3-tricarboxylate
Sub #5: udp-galactose
Sub #6: atp
Sub #7: 3'-azido-3'-deoxythymidine monophosphate
Sub #8: bradykinin
Sub #9: gastrin 17
Sub #10: n-carbamoyl-l-aspartate
Sub #11: 5-hydroxy-2-oxo-4-ureido-2,5-dihydro-1h-imidazole-5-carboxylate
Sub #12: fe2+
Sub #13: d-glyceraldehyde 3-phosphate


Connectedness of Each Amino Acid in PDB Structure
Residue 79
(n=18)
Residue 164
(n=13)
Residue 486
(n=18)
Number of Connected Residues
0
5
10
15
20
0
100
200
300
400
500
Residue Number