KcatNet: advancing genome-wide enzyme K_{cat} prediction through structural enzymatic characterization

Tong Pan[1], Xin Cui[2], Huan Yee Koh[3], Yue Bi[1], Xiaoyu Wang[1], Yumeng Zhang[1], Shantong Hu[2], Geoffrey I. Webb[3], Lukasz Kurgan[4,*], Guimin Zhang[2,*], Jiangning Song[1,*]

[1]Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia

[2]College of Life Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

[3]Department of Data Science and Artificial Intelligence, Monash University, VIC 3800, Australia

[4]Department of Computer Science, Virginia Commonwealth University, Richmond 23284, USA

*To whom correspondence should be addressed: Jiangning Song, Email: jiangning.song@monash.edu; Guimin Zhang, Email: 2021500017@buct.edu.cn; Lukasz Kurgan, Email: lkurgan@vcu.edu.

Abstract

Enzyme turnover numbers (K_{cat}) are fundamental kinetic constants that quantify enzymatic efficiency. Systematic studies of K_{cat} are essential for characterizing the mechanisms underlying proteomic composition and cellular metabolism. However, experimental measurements of K_{cat} remain limited and prone to noise. To address this, we present KcatNet, a geometric deep learning model designed for high-throughput prediction of K_{cat} in metabolic enzymes across all organisms, leveraging paired enzyme sequence and substrate representations. KcatNet consistently outperforms existing predictors, particularly for enzymes with high catalytic efficiency, and demonstrates strong generalization to enzymes that are dissimilar to those in the training set. The model generates interpretable importance scores for enzyme residues involved in catalysis, thereby enabling a quantitative assessment of their contributions to catalytic activity. Furthermore, KcatNet uncovers structural mechanisms and interaction patterns within enzyme-substrate complexes, providing insights into architectural principles that are inaccessible with existing methods. We applied KcatNet to genome-scale K_{cat} prediction across diverse yeast species, improving proteome allocation predictions by integrating its outputs into metabolic models. Experimental validation further confirmed the model's ability to identify enzyme mutants with enhanced activity. By bridging the gap between sequence, structure, and function, KcatNet establishes a robust foundation for advancing understanding of molecular-level mechanisms and accelerating enzyme engineering efforts.

Introduction

The enzyme turnover number, K_{cat}, quantifies the number of substrate molecules converted to product per enzyme molecule per unit of time under substrate-saturated conditions (*1*). As a fundamental parameter of catalytic efficiency, K_{cat} provides critical insights into cellular metabolism and enzyme evolution (*2-4*). Incorporating organism-specific K_{cat} values into genome-scale metabolic models (GEMs) has advanced our understanding of protein allocation and cellular growth by accounting for the protein cost associated with specific reaction fluxes (*5, 6*). Furthermore, the diverse catalytic efficiencies of enzymes are shaped by evolutionary and biophysical constraints imposed by natural selection (*7, 8*). Therefore, investigating the differences in catalytic turnover rates underlying complex enzymatic mechanisms is essential for exploring biological systems from an evolutionary perspective (*9*). Comparative analyses of catalytic activities across diverse enzymes and their mutants offer valuable insights into enzyme evolution, informing strategies for metabolic engineering (*10*).

K_{cat} is typically determined through experimental measurements. However, obtaining a complete K_{cat}-ome for an organism via *in vitro* characterization faces significant challenges, including difficulty in purifying specific enzymes, limited substrate availability, and incomplete knowledge of required cofactors (*11*). Currently, there are no high-throughput methods for experimentally measuring K_{cat} values and the evaluation process remains time-consuming and labor-intensive (*12*). As a result, the number of annotated K_{cat} values for enzymes of individual species available in public databases such as BRENDA (*13*) and SABIO-RK (*14*) is limited and far below proteome-scale. For example, even in *Escherichia coli*, one of the most extensively studied model organisms, *in vitro* K_{cat} values are available for only ~12% of enzymatic reactions (*15*). Additionally, experimental K_{cat} values can span several orders of magnitude and vary widely due to differences in assay conditions (*16*). This variability, combined with the sparse annotation of K_{cat} values, highlights the urgent need for computational tools capable of reliably predicting these values for enzymes on a proteome-wide scale. Such methods would enable exploration of enzyme activities and accelerate progress in biomolecular design (*17*).

Recent advances in artificial intelligence (AI) have spurred the development of novel predictors of K_{cat} values (*18*). Methods such as DLkcat (*19*), TurNuP (*20*), and UniKP (*21*) leverage enzyme sequences and substrate features to predict K_{cat} values across diverse organisms, employing deep learning models or machine learning classifiers like gradient boosting (*22*) and extra trees (*23*). Despite these advancements, these approaches face challenges in model training (*24*), limiting their ability to generalize effectively and accurately predict high K_{cat} values, which are underrepresented in experimental datasets. While experimentally measured K_{cat} data encompass a diverse range of enzyme sequences and substrates, they often lack detailed structural information about enzyme-substrate interactions. Manual efforts are typically required to correlate structural data with catalytic efficiency to uncover the structural mechanisms driving catalytic activity (*25*). Predicting K_{cat} values using structural data remains a significant challenge (*26*), despite progress for specific organisms like *Escherichia coli*. For instance, Heckmann et al. (*15*) proposed a K_{cat} predictor that incorporates manually curated features, including catalytic site structures and reaction fluxes derived from parsimonious flux balance analysis. Although this approach demonstrates the potential of integrating enzyme structural data for K_{cat} prediction, its applicability is limited to a small subset of enzymatic reactions in well-characterized organisms like *E. coli*. Nevertheless, this remains a promising research direction, especially given recent breakthroughs in protein structure prediction, which have significantly expanded the availability of high-quality putative structural data for enzymes (*27*).

Motivated by the lack of broadly applicable and efficient methods that leverage structural data for K_{cat} prediction, we introduce KcatNet, an innovative tool based on a geometric deep learning model. KcatNet effectively captures structural information using paired enzyme sequences and substrate SMILES representations as input, enabling rapid and accurate K_{cat} prediction at the genomic scale. We demonstrate that KcatNet can capture the geometric contexts of enzyme-substrate interactions, even for enzymes without experimentally resolved structures, by working in conjunction with structure prediction methods. KcatNet outperforms existing methods in predicting *in vitro* K_{cat} values, exhibiting superior generalizability and accuracy, particularly for high K_{cat} values, while remaining agnostic to protein families. Additionally, KcatNet povideds interpretability by generating importance scores for residues involved in catalytic activity. It identifies spatially contiguous residue clusters proximal to or within the catalytic pocket, potentially linking these residues to functional roles beyond direct catalysis. We applied KcatNet to predict K_{cat} values across various yeast species, enabling the construction of the enzyme-constrained metabolic models that improve phenotype and proteome predictions. Finally, we used KcatNet to study the directed evolution of α-glucosidase. Among nine single-point mutations predicted by KcatNet to enhance catalytic efficiency, seven were experimentally validated to exhibit K_{cat} values exceeding those of the wild-type enzyme. Collectively, these results underscore KcatNet's potential to advance enzyme engineering and metabolic modelling.

Results

Model overview

The KcatNet framework is illustrated in Fig. 1. KcatNet is a custom-designed deep learning model that predicts K_{cat} values at the genomic scale using enzyme-substrate pairs as inputs (Fig. 1a). The model's design is grounded in the principle that enzymes catalyze reactions by binding substrate molecules in specialized regions known as catalytic pockets (*28, 29*). Active residues within these pockets facilitate enzymatic reactions by recognizing and positioning substrates, contributing to both specificity and efficiency (*30*). KcatNet is engineered to identify catalytic pockets and infer catalytic efficiency by analyzing enzyme substructures and uncovering associations between substrates and these pockets. The model comprises three main components, as shown in Fig. 1b: 1) an enzyme-substrate encoder module; 2) a graph-based residue-level partition module, and 3) and an interaction module that captures enzyme-substrate interaction patterns.

In the enzyme-substrate encoder module (Fig. 1d), representations of the enzyme and substrate are combined with complementary information generated by modern large language models. Specifically, we extract residue-level features from the enzyme sequence using ProtT5 (*31*) and ESM2 (*27*), two widely used deep learning-based protein language models (PLMs) trained on billions of amino acid sequences. These features encapsulate information about individual enzyme residues and their relationships within the context of the entire enzyme sequence. We concatenate these residue-level features to produce a comprehensive enzyme-wise sequence representation. For substrate representation, we use the SMILES format, a standard representation for encoding molecular structures. Substrates are encoded at two distinct levels: (i) Atomic level: Each atom is initialized with type and property embeddings, and (ii) Molecular level: The entire molecule is processed using a pretrained SMILES transformer (*32*). In addition to the PLM-based enzyme sequence representations and SMILES-based substrate encodings, we encode local enzyme substructures through the graph-based residue-level partition module (Fig. 1d). Here, the enzyme structure is modeled as a graph that captures the spatial arrangement of its amino acids. We employ a graph convolutional neural network (GCN) to process these enzyme structure graphs (*33*), refining residue-specific representations by propagating embeddings among spatially proximal residues. For enzymes lacking experimentally resolved structures, we use ESM2 to generate residue contact maps, which quantify spatial relationships between residues. These residues are then grouped into distinct clusters using graph partitioning (*34*), a method that decomposes the graph into smaller, strongly connected subgraphs. The resulting residue partitions, represented as weighted graphs, undergo max pooling operations to produce partition-wise representations.

We process the three complementary inputs, which represent the substrate structure and the enzyme structure and sequence, using a Multilayer Perceptron (MLP) network to generate K_{cat} predictions. In essence, this network captures spatially local interaction patterns between substrates and catalytic pockets within the graph-based enzyme structure, contextualized by the PLM-based enzyme sequence space (Fig. 1d). An attention mechanism evaluates the interaction propensity of substrates with different residue regions to identify catalytic pockets. During training, the model updates both the substrate and enzyme partition-wise representations based on their contributions to the enzyme-substrate reaction. Message-passing operations are applied to both substrates and enzyme substructures, where each enzyme residue receives messages from its corresponding enzyme substructure, and each substrate atom receives messages from the molecular-level embedding. Through iterative training updates relying on the message passing, the final enzyme and substrate embeddings accurately represent substrate-catalytic pocket interactions. Pooling operations are then used to project these embeddings into scalar probability values representing turnover number predictions.

KcatNet performance for enzyme turnover number prediction

We evaluated the accuracy of KcatNet for K_{cat} predictions on a benchmark set, as outlined in the Methods section, enabling a comparative assessment of our model's performance against existing methods. Predictions were assessed using several performance metrics, including Pearson's correlation coefficient (PCC), coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), by comparing predicted K_{cat} values with experimentally observed data (Refer to the "Evaluation metrics" section in the Supplementary File).

We compared K_{cat} values predicted by KcatNet with experimental measurements for the same enzyme-reaction pairs, as shown in Fig. 2a. The results demonstrate a strong correlation between KcatNet's predictions and experimentally measured K_{cat} values, achieving a PCC of 0.84 on the test set (Fig. 2a) and 0.94 on the entire dataset (Supplementary Fig. S4), respectively. We further compared KcatNet's predictive performance with two recently published tools, namely DLKcat (*19*) and UniKP (*21*), as shown in Fig. 2b. KcatNet achieved a substantially lower

RMSE of 0.78 on the test dataset, outperforming DLKcat (****, $P = 4.35 \times 10^{-10}$) and UniKP (***, $P = 3.15 \times 10^{-4}$). Notably, compared to the second-best method (UniKP), KcatNet improved R^2 by 18% ($R^2 = 0.69$, ***, $P = 4.07 \times 10^{-4}$). Additionally, KcatNet yielded a lower MAE of 0.55, outperforming both DLKcat (****, $P = 2.39 \times 10^{-11}$) and UniKP (****, $P = 1.32 \times 10^{-7}$). Pairwise *P*-values for these comparisons are provided in Supplementary Table S3. These findings indicate that KcatNet's predictions surpass those of previously published methods. Furthermore, when categorizing enzymes by their metabolic roles, KcatNet predicted higher K_{cat} values for enzymes involved in primary central and energy metabolism compared to those in intermediary and secondary metabolism (Fig. 2c). This observation aligns with previous findings (*35*) and provides additional evidence of the model's accuracy.

We observe that the distribution of K_{cat} values in the test dataset is skewed, with approximately 71% of values falling between 0.1 and 100, while only 14% exhibit high K_{cat} values (>100) (Extended Data Fig. 2a). Accurate identification of enzymes with high K_{cat} values is particularly important for enzymology and synthetic biology (*5*). However, we find that existing models systematically overestimate low K_{cat} values and underestimate high ones (Supplementary Fig. S3, S5). To evaluate KcatNet's performance in predicting high K_{cat} values, we stratified the analysis across different K_{cat} ranges (Fig. 2d). The results show that KcatNet provides accurate predictions, with RMSEs ranging from 0.61 to 1.01 across the entire spectrum of K_{cat} values, indicating predictions within one order of magnitude of experimental values. Notably, KcatNet consistently outperforms the two recently published methods, particularly excelling in predicting both very low and very high K_{cat} values. Furthermore, while the K_{cat} values in the test dataset originate from diverse organisms, we partitioned the test set into organism-specific subsets to evaluate KcatNet's performance for individual organisms. We find that KcatNet delivers strong predictive performance across a wide variety of organisms, including those underrepresented in the training dataset, such as *Drosophila melanogaster* (Extended Data Fig. 3c, 3d).

KcatNet captures enzyme homolog-independent patterns for K_{cat} prediction

To evaluate performance of KcatNet across protein families, we obtained protein family annotations from the CATH protein family database (*36*) and detailed the abundance of protein families in our test dataset in Supplementary Fig. S8. We found that KcatNet outperforms the two recently published methods across various protein families, including those with limited data, such as Golgi alpha-mannosidase II (Fig. 2e). This suggests that our model captures fundamental mechanisms of enzyme-substrate reactions rather than overfitting family-specific patterns.

Previous studies have shown that predictive performance is often influenced by the similarity between target enzyme sequences and those in the training dataset, as similar enzymes tend to share similar functions (*1*). As expected, KcatNet performs exceptionally well in predicting K_{cat} for test enzymes highly similar to those in the training set, achieving a PCC of 0.90 (Supplementary Fig. S6). To assess KcatNet's generalization ability, we tested its accuracy on unseen enzymes not included in the training set (Extended Data Fig. 2b). KcatNet maintained a strong correlation between the predicted and experimental K_{cat} values, with a PCC of 0.84 ($N = 1116$). In this setting, KcatNet outperformed existing methods (Extended Data Fig. 2c), achieving a lower RMSE of 0.90 compared to DLKcat (****, $P = 7.98 \times 10^{-5}$) and UniKP (*, $P = 2.13 \times 10^{-2}$). Additionally, it achieved a higher R^2 of 0.57, outperforming DLKcat (***, $P = 2.07 \times 10^{-4}$) and UniKP(*, $P = 2.25 \times 10^{-2}$), confirming the statistical significance of KcatNet's superior performance. Pairwise *P*-values for these comparisons are provided in Supplementary Table S4. To further mitigate potential overfitting caused by similar enzyme entries, we evaluated the model's robustness on enzymes dissimilar with those in the training dataset. As shown in Fig. 2f, we divided the test set into subsets based on their maximal enzyme sequence identity compared to the training set and calculated model performance at different levels of sequence identity. KcatNet generated accurate predictions even for test enzymes with low sequence identity to those in the training set.

Similarly, KcatNet demonstrated excellent performance in predicting K_{cat} for chemical reactions included in the training set (Supplementary Fig. S7). However, even after excluding reactions present in both the test and training datasets, the model maintained high predictive accuracy, with a PCC of 0.78 ($N = 575$) (Extended Data Fig. 2b). Notably, for previously unseen reactions not present in the training dataset, KcatNet achieved a lower RMSE of 0.78 compared to DLKcat (****, $P = 6.38 \times 10^{-6}$) and UniKP (**, $P = 1.39 \times 10^{-3}$). It also demonstrated a significantly higher R^2 of 0.72 than DLKcat (****, $P = 1.19 \times 10^{-5}$) and UniKP (***, $P = 8.10 \times 10^{-4}$) (Extended Data Fig. 2c). These results indicate that KcatNet can reliably predict K_{cat} even for previously unseen enzymes and reactions not present in the training dataset. We found that enzyme characteristics exerted a greater influence on K_{cat} prediction than substrate properties (Extended Data Fig. 3a), likely due to the overlapping information in enzyme representations, which inherently capture functional details and substrate specificity constraints.

KcatNet identifies active sites involved in catalysis

A key strength of KcatNet lies in its ability to leverage enzyme structural features to provide interpretable context for catalytic efficiency predictions. By utilizing residue graphs, KcatNet is capable of identifying active sites within enzyme-substrate complexes that drive catalytic reactions, even without the prior knowledge of substrate-binding information.

We demonstrate that KcatNet can pinpoint spatially contiguous regions of residues proximal to or within the catalytic pocket that play critical roles in catalysis. As shown in Fig. 3a, using an enolase from *Pelagibaca bermudensis* HTCC2601 (PDB: 4H2H, chain A) as an example (*37*), we investigated the enzyme's catalytic efficiency in catalyzing L-4-Hydroxyproline betaine. We visualized these predictions at the residue level by calculating the model's activation scores (normalized attention weights) and mapping them onto the enzyme's three-dimensional structure, with binding and active sites highlighted in red as retrieved from UniProt. KcatNet accurately identified all catalytic residues directly involved in the reaction by assigning them high activation scores. Additionally, all residues assigned high importance scores, though located at distinct sequence positions, were spatially proximal to the catalytic residues (Supplementary Fig. S9).

To quantify the spatial distribution of key residues identified by KcatNet as crucial for enzyme-substrate catalytic reaction efficiency, we analyzed 11 enzymes from our test dataset with detailed catalytic residue annotations in the MCSA database (*38*). As shown in Fig. 3b, for each enzyme, we calculated the spatial distance between annotated catalytic residues and the active residues deemed important for turnover number prediction—those assigned high activation scores by KcatNet—as well as non-active residues (low activation scores). Our analysis revealed a distinct spatial distribution between active and non-active residues identified by KcatNet, with a significant portion of critical residues located near catalytic sites (Supplementary Fig. S9).

We hypothesize that some residues near or within the catalytic pocket play functional roles beyond direct catalysis. In some cases, the model identifies a small subset of residues with high activation scores that are involved in substrate binding. For example, as shown in Fig. 3c, we examined dTDP-glucose 4,6-dehydratase (PDB: 1BXK, chain A) from *E. coli* (*39*), which converts dTDP-glucose to dTDP-4-keto-6-deoxyglucose. The model identified a small group of residues with high activation scores that, while not directly participating in catalysis, are involved in substrate binding. Collectively, these results suggest that KcatNet effectively identifies regions near substrate binding and catalytic sites by calculating the functional importance of residues in predicting enzyme-substrate reaction efficiency, offering valuable insights into the structural basis of enzyme-substrate interactions.

KcatNet accurately predicts K_{cat} values for mutant and wild-type enzymes

Next, we evaluated KcatNet's ability to elucidate the effects of missense variants on K_{cat} values, a critical aspect of protein engineering. We separated catalytic reaction entries into wild-type and mutant enzymes and independently assessed the model's predictions for each group. We observed a strong correlation between KcatNet's predicted and experimentally measured K_{cat} values for both mutant enzymes (PCC = 0.85 for the test set, Fig. 4a; PCC = 0.94 for the entire dataset, Supplementary Fig. S10) and wild-type enzymes (PCC = 0.78 for the test set, Fig. 4b; PCC = 0.94 for the entire dataset, Supplementary Fig. S10). Compared to the two recently published methods, DLKcat and UniKP, KcatNet achieved lower RMSEs for both mutant enzymes (RMSE = 0.78 for the test set, Fig. 4c; RMSE = 0.52 for the entire dataset, Supplementary Fig. S11) and wild-type enzymes (RMSE = 0.77 for the test set, Fig. 4c; RMSE = 0.40 for the entire dataset, Supplementary Fig. S11). Similarly, KcatNet achieved higher R^2 values between predicted and experimentally measured K_{cat} values for mutant enzymes (R^2 = 0.71 for the test set, Fig. 4c; and R^2 = 0.88 for the entire dataset, Supplementary Fig. S12) and wild-type enzymes (R^2 = 0.59 for the test set, Fig. 4c; and R^2 = 0.89 for the entire dataset, Supplementary Fig. S12), outperforming both DLKcat and UniKP.

Fig. 4d presents a subset of K_{cat} values for six mutant enzymes, each with at least 25 unique single or multiple amino acid substitutions. KcatNet accurately predicted K_{cat} values for these enzymes across 192 mutation entries, achieving a high PCC of 0.98. We further categorized the mutation data into wild-type and "reducing" mutations, defining the latter as those with K_{cat} values less than 50% of the wild-type enzyme. Fig. 4e illustrates that KcatNet effectively distinguishes between wild-type enzymes and those with reducing mutations, with wild-type enzymes exhibiting higher K_{cat} values. We also demonstrated that the residual-level importance scores identified by KcatNet, which reflect the extent to which residues influence catalytic efficiency through roles such as catalysis and substrate binding, are reliable and potentially biophysically significant. For example, by mapping the residue activation scores of the enzyme PNP (PDB: 2A0W, chain A) onto its three-dimensional structure (*40*), we observed a correlation between these activation scores and catalytic efficiency. Nearly all mutations at residues identified by KcatNet as highly important resulted in reduced K_{cat} values (Fig. 4f). This finding aligns well with

the observation that functionally important residues in enzymes are typically conserved and that mutations in these residues often impair enzyme function (*41*).

KcatNet's predictions improve enzyme-constrained genome-scale metabolic models

Enzyme-constrained genome-scale metabolic models (ecGEMs) (*42*) simulate proteome allocation underlying key cellular properties such as growth rate, and rely heavily on genome-wide K_{cat} values, as protein costs in metabolic processes are determined by an enzyme's effective turnover rate. The K_{cat} values predicted by KcatNet can facilitate the reconstruction of ecGEMs, enabling the investigation of proteome allocation patterns (*43*). Li et al. *(19)* reconstructed original-ecGEMs for all 343 yeast/fungi species by assigning K_{cat} values from other substrates or organisms and, when necessary, introducing wildcards in the EC number to account for missing data. The original-ecGEMs reconstruction pipeline relied heavily on accurate EC number annotations and the availability of experimentally measured K_{cat} values in databases, ultimately yielding K_{cat} values for around 40% of enzymes. In contrast, deep learning–predicted K_{cat} values based on enzyme sequences and substrate SMILES significantly improved coverage, providing K_{cat} values for about 80% of enzymes *(19)*. The remaining gaps primarily involved generic substrates lacking defined SMILES representations, such as phosphatidate and thioredoxin (*44*). Beyond achieving higher coverage, we applied the same original-ecGEMs pipeline but parameterized it with KcatNet-predicted K_{cat} values to assess their impact on proteome prediction accuracy. As shown in Fig. 5a, integrating KcatNet-predicted K_{cat} values into ecGEMs improved performance over the original ecGEMs, yielding more accurate growth rate predictions for four yeast species across diverse carbon sources and oxygen conditions. Compared to models parameterized with DLKcat-predicted K_{cat} values, our predictions reduced the mean squared error between the predicted growth rates and published quantitative data in 16 out of 22 environment-species combinations.

Harnessing KcatNet for enzyme engineering

Enzyme engineering aims to develop more efficient enzyme variants through directed evolution. However, identifying effective evolutionary pathways is challenging, as it requires an understanding of molecular mechanisms, such as gene expression and protein folding, while navigating biological and physical constraints. We applied KcatNet in a prospective study to guide the directed evolution of α-glucosidase (*45*), a key enzyme in the amylolytic pathways of various organisms. α-Glucosidase catalyzes the hydrolysis of α-D-glucosides and facilitates transglucosylation, enzymatic activities extensively used in the synthesis of oligosaccharides and α-D-glucosides. Despite its importance, progress in designing α-glucosidase mutants with enhanced efficiency has been limited.

We used KcatNet to filter α-glucosidase mutations with enhanced catalytic efficiency (Fig. 5b). We generated all single-point α-Glucosidase variants by substituting each amino acid in its 555-residue sequence with one of the other 19 canonical amino acids, resulting in 19*555 = 10,545 variants. Initial filtering of these variants was performed using two *in silico* tools: the sequence-based ESM-1v (*46*) and the structure-based ProteinMPNN (*47*), both experimentally validated by Sean R. Johnson *et al*. (*48*) for predicting the expression and folding of enzyme sequences into soluble, active forms. This filtering process yielded 242 mutations with scores higher than those of the wild-type α-glucosidase. Subsequently, we performed *in silico* screening of these 242 variants using KcatNet to predict their corresponding K_{cat} values. Our analysis identified nine residues where multiple single-point mutations were predicted to enhance K_{cat} values (Fig. 5c). For each residue, we randomly selected one mutation for follow-up experimental validation. Of the nine mutants predicted to have increased catalytic efficiency, seven exhibited K_{cat} values higher than the wild-type enzyme (Fig. 5d). These results demonstrate that KcatNet could effectively aid in enzyme engineering by identifying promising variants for experimental evaluation.

Discussion

A systematic and comprehensive investigation of enzyme efficiency is essential for unraveling their molecular catalytic mechanisms and evolutionary trajectories, enabling the design of enzymes to address current challenges in medicine, engineering, and planetary health (*49*). However, predicting enzyme kinetic constants on a genome-wide scale remains a significant challenge due to variability in organismal fitness and the diversity of biochemical reactions (Supplementary Fig. S2). In this study, we introduce KcatNet, a sophisticated deep learning-based framework that accurately predicts the kinetic constant K_{cat} on a genome-wide scale. KcatNet processes enzyme sequences and substrate SMILES representations as inputs, consistently delivering accurate K_{cat} predictions by effectively capturing site-specific interaction patterns between substrates and enzyme active sites. Our model identifies spatially contiguous enzyme regions, including residues directly involved in catalysis and those critical

for substrate binding, providing a detailed mapping of enzyme-substrate interactions and catalysis that is not accessible with existing approaches.

KcatNet generates accurate predictions of *in vitro* K_{cat} values for enzymes across a diverse collection of organisms, achieving PCCs > 0.80 (Extended Data Fig. 3d). We empirically demonstrate that KcatNet outperforms DLKcat, a leading deep learning model that incorporates substrate information and protein sequences, and UniKP, a recent tree-based ensemble model, in predicting K_{cat} for both mutant and wild-type enzymes. KcatNet also exhibits strong generalizability, accurately predicting K_{cat} for enzymes and reactions absent from the training set (Extended Data Fig. 2b, 2c). Even when evaluated on enzymes with low similarity to those in the training set, KcatNet maintains a low RMSE between measured and predicted K_{cat} values. Additionally, KcatNet effectively distinguishes catalytic efficiency differences for enzymes in various metabolic contexts, predicting higher turnover numbers for enzymes in high-flux pathways of primary central and energy metabolism compared to those involved in intermediary and secondary metabolism.

Identifying enzymes with high catalytic efficiency is a central goal in enzymology and synthetic biology (*10*). KcatNet demonstrates superior performance in predicting high K_{cat} values, particularly in the context of highly imbalanced data where such high instances are rare. Traditional machine learning models have historically struggled with this imbalance, often failing to extract crucial information from limited high-k_{cat} data. By categorizing our dataset according to enzyme class (first-level EC numbers) and comparing the average measured K_{cat} values with the average predicted values for each class, we found that KcatNet consistently predicted higher K_{cat} values for faster EC classes and lower values for slower ones (Extended Data Fig. 4b). This result underscores KcatNet's effectiveness in capturing the intrinsic catalytic properties of enzymes, as evidenced by its accurate predictions of K_{cat} values across different enzyme classes (Extended Data Fig. 4c).

We attribute KcatNet's superior performance in part to the enzyme representations generated by pretrained large language models. Our analysis revealed that such representations play a critical role in predicting K_{cat} values (Extended Data Fig. 3a). Notably, even when predictions were based solely on enzyme information—without accounting for the catalyzed reaction—a respectable coefficient of determination was achieved. This finding suggests that the intrinsic properties of enzymes significantly influence turnover numbers, likely reflecting varying selection pressures on catalytic turnover driven by enzyme utilization. Interestingly, additional substrate data might yield only moderate performance improvements if the model struggles to capture key interactions essential for inferring reaction efficiency, particularly due to the absence of explicit structural enzyme-substrate binding information. The role of active sites in catalysis—whether through substrate binding, catalytic transformation, or direct participation in the reaction—has been extensively studied in structural and biochemical research (*50*). However, existing models for predicting K_{cat} often overlook the detailed molecular interactions between enzyme active sites and substrates, largely because most enzyme-substrate complex structures are unavailable. KcatNet addresses this limitation by leveraging structural characterizations of enzymes to improve the interpretability of catalytic efficiency predictions. Our study demonstrates that KcatNet, when applied to multiple enzymes, effectively captures structurally significant site-specific interactions between enzymes and substrates. It uncovers spatially contiguous regions of residues near active sites linked to key functional roles, such as ligand-binding sites and catalytic residues. This approach offers valuable biological insights into enzyme-substrate interactions and catalysis, providing a detailed map of the underlying architecture—something that existing methods struggle to achieve, even without prior ligand-binding information.

Despite KcatNet delivering reliable K_{cat} predictions with performance surpassing existing methods, several challenges remain. One significant challenge is the high variance in K_{cat} measurements for the same enzyme-reaction pairs across different studies. Although we use a manually curated K_{cat} dataset for model construction and evaluation, this variance persists. It likely stems not only from database errors and differences in experimental procedures or assay conditions, such as temperature and pH. Another challenge arises with reactions involving multiple substrates or heteromeric enzyme complexes. KcatNet currently takes one substrate and one protein as input, while multiple different substrates or proteins may be involved in such reactions. This can result in multiple predicted K_{cat} values for a single reaction. In these cases, a more refined approach to predict a single K_{cat} value for multi-substrate or heteromeric enzymes would be ideal. Moreover, the current model does not account for environmental factors, a critical limitation when simulating real experimental conditions. K_{cat} values can vary considerably depending on factors like pH and temperature. Since experimental conditions are rarely documented in enzyme kinetic parameter databases, such information was unavailable as inputs for our prediction model. It is also important to note that other kinetic parameters, such as the Michaelis constant (K_m), do not improve KcatNet's performance in predicting K_{cat}. Future work on developing methods to simultaneously predict K_{cat} and K_m values,

aligning with experimentally determined K_{cat}/K_m ratios, could enable comprehensive parameterization of enzyme kinetics and provide valuable insights into cellular physiology.

A key application of KcatNet is to facilitate the reconstruction of enzyme-constrained metabolic models (ecGEMs) for virtually any organism by predicting genome-scale K_{cat} profiles as kinetic parameters. We employed a Bayesian approach to adjust the predicted *in vitro* kinetic parameters to align with existing experimental growth data. Our results demonstrate that KcatNet provides comprehensive K_{cat} predictions that significantly improve proteome allocation predictions. Additionally, KcatNet accurately captures changes in K_{cat} due to single amino acid substitutions, identifying mutations in catalytic residues that are likely to result in loss of enzyme function. The application of KcatNet to α-Glucosidase in directed evolution highlights its transformative potential for synthetic biology and biochemistry research. KcatNet effectively identified high-activity α-Glucosidase variants, with seven out of nine single-point mutations predicted to enhance catalytic efficiency, exhibiting K_{cat} values exceeding those of the wild-type enzyme. In conclusion, KcatNet can be broadly applied for monitoring biocatalysis, guiding enzyme engineering, and advancing our understanding of enzyme evolution and metabolic pathways.

Methods

Training and test datasets

We curated a dataset of K_{cat} measurements for both wild-type and mutant enzymes, which we used to construct and evaluate our predictive model. For the K_{cat} values of wild-type enzymes, we utilized the dataset curated by Kroll *et al*. This dataset derived the K_{cat} values from three databases—BRENDA, UniProt, and Sabio-RK—linking K_{cat} valus with corresponding enzyme sequences, reactant IDs, and reaction equations. To minimize noise in the data, such as errors from transcription or unit misplacements, more than half of the data points were manually verified to ensure K_{cat} values matched those reported in the original papers.

For the K_{cat} values of mutant enzymes, we collected data from the BRENDA (Release 2023.1) and SABIO-RK databases, focusing on entries with mutation annotations in their descriptions. Each dataset was linked to substrate details, organism information, UniProt IDs, and/or enzyme EC numbers. When multiple K_{cat} values were available for the same enzyme-substrate pair, we excluded values less than 1% of the maximum and calculated the geometric mean of the remaining values. Substrate SMILES, representing the substrate structures, were retrieved from the PubChem database by matching substrate names. Entries without corresponding SMILES were excluded. For protein sequences, we obtained entries with UniProt IDs directly from the UniProt database using their mapping service. For entries without UniProt IDs, sequences were retrieved from both the UniProt and BRENDA databases based on EC number and organism information. Entries returning multiple sequences were omitted.

Given that most K_{cat} values in the BRENDA and SABIO-RK databases lacked experimental condition details such as pH and temperature, we omitted this information from our curated dataset. As a result, we obtained 11,757 final entries for model training and testing purposes, including 11,288 unique reactions and 7,441 unique enzymes. All K_{cat} values were converted to a logarithmic scale. This dataset was randomly divided into an 80% training dataset for model construction and a 20% test dataset for performance evaluation and comparison.

Enzyme sequence and structure embedding

The model takes as input an enzyme sequence paired with the SMILES representation of a molecular substrate and outputs the predicted K_{cat} value. We represent an enzyme sequence $S = \{\mathrm{r}_1, \mathrm{r}_2, \ldots, \mathrm{r}_n\}$, consisting of n residues, as an attributed graph. This graph representation of the enzyme, denoted as $ENZ = (V, A, X)$, embodies the three-dimensional structure of the input enzyme. In this representation, the residue set $V \subseteq S$ servers as the graph's nodes, the adjacency matrix A (of size $n \times n$) quantifies the connectivity of these nodes based on their spatial distance, and the residue feature matrix $X \in \mathcal{R}^{n\times \mathrm{d}}$ encapsulates the properties of the residues.

For an enzyme sequence S, we generate residue-level feature embeddings using two pretrained protein language models: ProtT5 and the model proposed by Bepler and Berger, resulting in embeddings $E_1 \in \mathcal{R}^{n\times d_1}$ and $E_2 \in \mathcal{R}^{n\times d_2}$, where d_1 defaults to 1024 and d_2 defaults to 1280, respectively. ProtT5 is a deep learning-based language model pre-trained on datasets comprising 393 billion amino acids. Bepler and Berger's model is a bidirectional long short-term memory (Bi-LSTM) neural network trained on three types of data: 1) The protein's SCOP classification based on its general structure; 2) The self-contact map of the protein's structure, and 3) Sequence alignments of homologous proteins. These enzyme vectors encapsulate information about individual residues and broader protein-

level characteristics. Additionally, we represent the enzyme's spatial structure as an $n \times n$ residue adjacency matrix, generated by Bepler and Berger's model, which encodes the 2D structure of the enzyme S.

Substrate molecular embedding

Substrate molecules are represented using the SMILES format. For a substrate molecule Sub={r_1,r_2,...,r_m} consisting of m atoms, the representation is embedded at two levels: the atomic level and the molecular level. At the atomic level, each atom is initialized with type and property embeddings, forming the atomic feature matrix $S_{\text{atom}} \in \boldsymbol{\mathcal{R}}^{\text{m}\times 43}$, which encapsulates the atom's characteristics. Atom type features are derived from one of nine possible element groups: B, C, N, O, P, S, Se, halogens, and mMetals. The physicochemical properties include atom degree, the number of implicit hydrogens, atom hybridization, and chirality. At the molecular level, we obtain a comprehensive substrate representation $S_{\text{mol}} \in \boldsymbol{\mathcal{R}}^{1\times 1024}$ using a pretrained SMILES transformer, which captures the overall properties of the substrate.

Overview of the KcatNet model

The KcatNet model leverages enzyme sequences and one of the substrates to accurately predict genome-scale enzyme turnover numbers. This is achieved through a structure-characterized deep learning framework that analyzes enzyme substructures and uncovers correlations between substrates and catalytic pockets to infer catalytic efficiency. This framework consists of three major components: 1) An enzyme-substrate representation encoder module embeds enzymes using a geometrically invariant representation and employs a multi-scale substrate representation at both the atomic and molecular levels, facilitating message propagation between these scales. 2) A graph-based residue-level partition module broadcasts embeddings among spatially proximal residues, grouping closely interacting residues into distinct partitions, and 3) A catalytic complex turnover number predictor captures the interaction patterns between enzyme residue partitions and their corresponding substrates, enhancing the prediction of enzyme kinetic turnover numbers.

The enzyme-substrate representation encoder

We employed a graph encoder based on Principal Neighborhood Aggregation (PNA), a variant of graph convolutional network, to compute and propagate residue-level representations of enzymes. This encoder enhances message passing between spatially proximal residues within the enzyme structure by incorporating multiple aggregators and degree scalers, accounting for the characteristic residue density at protein functional sites.

Initially, we generate the enzyme embedding $E_{ini} \in \mathcal{R}^{n\times \text{h}}$ by concatenating and reducing the dimensionality of enzyme representations to a hidden dimension h using a multilayer perceptron (MLP) network, with h set to 200 by default:

$$E_{ini} = Norm(MLP([MLP(E_1), MLP(E_2)]))$$

where Norm refers to layer normalization applied to the input features to stabilize the learning process, and the brackets [] indicate the concatenation operation. We use the Rectified Linear Unit (ReLU) activation function.

The initial enzyme representation $E_{ini} \in \mathcal{R}^{n\times \text{h}}$, along with the adjacency matrix A of size $n \times n$, is fed into a PNA module. This module produces residue-level representations $E_{res}^{(i)} \in \mathcal{R}^{n\times d_{\text{i}}}$ through message passing between spatially proximal residues, where d_{i} denotes the embedding dimension of the ith convolutional layer:

$$E_{res}^{(i)} = \text{PNA}\left(A, E_{res}^{(i-1)}\right)$$

For the substrate, we generate the embedding $S_{ini} \in \mathcal{R}^{1\times \text{h}}$ by summing individual atom-level embeddings $X_{atom} \in \mathcal{R}^{\text{m}\times 43}$ into a single ligand representation $X_{lig} \in \mathcal{R}^{1\times \text{h}}$ based on the atoms' importance scores and concatenating global features of the molecular compound $X_{mol} \in \mathcal{R}^{1\times 1024}$ using an MLP network, with h set to 200 by default:

$$S_{ini} = Norm\left(MLP\left(\left[MLP\left(S_{lig}\right), MLP(S_{mol})\right]\right)\right)$$

We aggregate ligand atoms into a unified ligand representation $S_{lig} \in \mathcal{R}^{1\times \text{h}}$ by summing atom embeddings $S_{atom} \in \mathcal{R}^{\text{m}\times d_0}$ according to their importance scores:

$$S_{lig} = softmax(MLP(S_{atom}))^T S_{atom}$$

The graph-based residue-level partition module

The graph-based residue-level partition module groups residues into several partitions based on their spatial proximity. For a given graph defined by the adjacency matrix $A \in \{0,1\}^{n\times n}$ and the residue-level feature matrix $E_{res} \in \mathcal{R}^{n\times h}$, our model applied a graph convolutional network (GCN) to produce an enzyme partition assignment matrix $A_{region} \in \mathcal{R}^{\mathrm{k}\times n}$, where k represents the number of enzyme regions for the ith convolutional layer. The enzyme partition features $E_{region} \in \mathcal{R}^{\mathrm{k}\times \mathrm{h}}$ is subsequently derived by pooling residue features according to their assigned regions.

$$A_{region}^{(i)} = \mathrm{GCN}\left(A, E_{res}^{(i)}\right)$$

$$E_{region}^{(i)} = A_{region}^{(i)} E_{res}^{(i)}$$

The catalytic complex turnover number predictor

The catalytic complex turnover number predictor consistently estimates K_{cat} values by modeling the local interaction patterns between substrates and enzyme pockets. Using the substrate molecular representation $S \in \mathcal{R}^{1\times \mathrm{h}}$ and the enzyme partition features $E_{region} \in \mathcal{R}^{\mathrm{k}\times \mathrm{h}}$ as inputs, the model infers the propensity of specific enzyme regions for enzyme-substrate interactions by calculating the relationship between the substrate molecule and various enzyme partitions through an attention mechanism. The substrate molecular representation and enzyme partition representations are then updated as follows:

$$Att = softmax\left(\frac{SW^Q \times (E_{region}W^K)^T}{\sqrt{d_K}}\right)$$

$$\hat{S} = Norm(S + MLP\left((Att \times E_{region}W^V)\, W^O\right))$$

$$\hat{E}_{region} = Norm(E_{region} + MLP((1_k^T S) \odot Att^T))$$

where W^Q, W^K, W^V, and W^o are learnable parameters, d_K is a constant scalar factor, and *Norm* refers to layer normalization.

A message-passing operation is then applied to independently facilitate information exchange within the substrate and enzyme substructures. Specifically, substrate representations at both the atomic and global molecular levels, as well as enzyme representations at the residue and partition levels, undergo information exchange and updating through a multilayer perceptron (MLP) network as follows:

$$\hat{S}_{atom} = Norm(S_{atom} + MLP(\hat{S}))$$

$$\hat{E}_{res} = E_{res} + MLP((A_{region})^T \hat{S})$$

Through iterative updates of residue-level and atomic-level representations, the final enzyme and substrate embeddings are derived using a max-pooling operation. These embeddings are then projected into scalar probability values, facilitating spatially consistent turnover number predictions:

$$Y_{kcat} = MLP([max\left(\hat{S}_{atom}\right), max\left(\hat{E}_{res}\right)])$$

Given the experimental K_{cat} annotations and the predicted K_{cat} values, we minimized the root mean squared error (RMSE) loss to optimize the supervised objective. Additionally, we employed the MinCUT loss to learn an unsupervised partitioning of residues, resulting in the following overall loss function:

$$\mathcal{L} = \mathcal{L}_{rmse} + \mathcal{L}_{MinCUT}$$

Experimental materials

Plasmid construction

The gene encoding GH13_31 α-Glucosidase (BspAG13_31A) from *Bacillus sp* AHU2216 (*51*) was codon-optimized, synthesized and cloned into the pET28a vector by Beijing Tsingke Biotech Co. Ltd., resulting in the

plasmid pET28a-31A (see Supporting materials). The BspAG13_31A mutants (D69A, H89K, R230Q, W287F, D396K, T408N, Q468D, H528A, H544N) were generated using whole plasmid PCR with the 2x Phanta Max Mix (Vazyme, Nanjing) and pET28a-31A as the template. The primers used are listed in Table S1. The PCR products were digested with DpnI and transformed into *Escherichia coli* DH5a. Correct mutants were verified by sequencing and subsequently transformed into *E. coli* BL21(DE3) for protein expression.

Protein expression and purification

The mutants were individually inoculated into 2 mL of LB medium containing 50 μg/mL kanamycin and cultured overnight at 37 °C and 220 rpm. The overnight culture (1 mL) was transferred into 100 mL of LB medium with 50 μg/mL kanamycin in 100 mL shaking flasks. The cultures were incubated at 37 °C and 220 rpm for 2 - 3 h until the OD600 reached 0.6-0.8. Isopropyl β-D-1-thiogalactopyranoside (IPTG) was then added to a final concentration of 0.2 mM, and the cultures were further incubated at 18 °C overnight. After induced expression, the cells were harvested by centrifugation at 4,000 rpm nd 4 °C for 10 minutes. The cell pellets were resuspended in 10 mM phosphate buffer (pH 7.0) containing 0.5 M NaCl and disrupted by sonication. The lysate was centrifuged at 12,000 rpm and 4 °C for 10 minutes, and the supernatant was collected and loaded onto a nickel affinity column (GE Healthcare) for purification. The proteins were desalted using a PD-10 Desalting Column (Cytiva) and analyzed by SDS-PAGE. Protein concentrations were determined using the Bradford Protein Assay Kit (Beyotime, China).

Determination of kinetic parameters

Enzymatic assays to determine kinetic parameters were performed in 50 mM NaH2PO4-Na2HPO4 buffer (pH 7.0) at 37 °C using p-nitrophenyl-α-D-glucopyranoside (pNPG) at concentrations ranging from 0 to 5 mM (0 mM, 0.5 mM, 1 mM, 1.25 mM, 2 mM, 2.5 mM, 3.75 mM, 5 mM). Enzyme solutions (0.01~0.02 μM) were used for different variants. Following the literature (*51*), reactions were incubated at 37 °C for 15 minutes and then terminated with an equal volume of 100 mM NaCO3. Each reaction was performed in duplicate. Absorbance values were measured at 400 nm, and a standard curve was established to correlate p-nitrophenol (pNP) concentration with optical density (OD400) (Fig. S18). Initial reaction velocities (V0) were calculated from the rate of pNP production per second due to enzymatic hydrolysis at different pNPG concentrations. The V0 values and their corresponding pNPG concentrations were imported into the GraphPad Prism software, which automatically calculated the kinetic parameters (V_{max}, K_m, and K_{cat}) using the Michaelis-Menten equation.

Data availability: All raw and benchmark data resources used in this study are publicly available via Zenodo at (https://doi.org/10.5281/zenodo.14997549). The curated datasets were obtained from the following databases: BRENDA (https://www.brenda-enzymes.org/), SABIO-RK (http://sabio.h-its.org/), Protein Data Bank (PDB) (https://www.rcsb.org), UniProt (https://www.uniprot.org), and M-CSA (https://www.ebi.ac.uk/thornton-srv/m-csa/). All source code required to evaluate the conclusions in the paper is provided in the paper and/or the Supplementary Materials and can be accessed at https://github.com/BioColLab/KcatNet.

Code availability: All source code for KcatNet is available at https://github.com/BioColLab/KcatNet.

Author contributions:

T.P. and J.S. conceived the ideas. T.P., X.C. and G.Z. designed the experiments. T.P. and X.C. performed the experiments. T.P. analysed the data and prepared figures. T.P. and J.S. wrote the manuscript. G.I.W. and L.K. provided guidance on data analyses. G.Z., G.I.W. and J.S. supervised the project. All authors contributed ideas to the work and assisted in manuscript editing and revision.

Competing interests: All authors declare no competing interests.

Acknowledgements

We acknowledge financial support from National Health and Medical Research Council of Australia (grant nos. APP1127948, APP1144652, APP2036864 to J.S.). We also acknowledge financial support from the Major and Seed Inter-Disciplinary Research projects awarded by Monash University (J.S.). We thank Robin B. Gasser for valuable insights during the manuscript revision.

References

1. C. Markin *et al.*, Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. *Science* **373**, eabf8761 (2021).
2. M. Basan *et al.*, Overflow metabolism in Escherichia coli results from efficient proteome allocation. *Nature* **528**, 99-104 (2015).
3. J. Zhu, C. B. Thompson, Metabolic regulation of cell growth and proliferation. *Nature reviews Molecular cell biology* **20**, 436-450 (2019).
4. V. Nguyen *et al.*, Evolutionary drivers of thermoadaptation in enzyme catalysis. *Science* **355**, 289-294 (2017).
5. I. Domenzain *et al.*, Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. *Nat Commun* **13**, 3766 (2022).
6. X. Fang, C. J. Lloyd, B. O. Palsson, Reconstructing organisms in silico: genome-scale models and their emerging applications. *Nat Rev Microbiol* **18**, 731-743 (2020).
7. E. Campbell *et al.*, The role of protein dynamics in the evolution of new enzyme function. *Nat Chem Biol* **12**, 944-950 (2016).
8. J.-B. Lalanne *et al.*, Evolutionary convergence of pathway-specific enzyme expression stoichiometry. *Cell* **173**, 749-761. e738 (2018).
9. E. Z. Eisenmesser *et al.*, Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **438**, 117-121 (2005).
10. K. Chen, F. H. Arnold, Engineering new catalytic activities in enzymes. *Nat Catal* **3**, 203-213 (2020).
11. P. Wendering, M. Arend, Z. Razaghi-Moghadam, Z. Nikoloski, Data integration across conditions improves turnover number estimates and metabolic predictions. *Nat Commun* **14**, 1485 (2023).
12. J. O. Park *et al.*, Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. *Nat Chem Biol* **12**, 482-489 (2016).
13. I. Schomburg, A. Chang, D. Schomburg, BRENDA, enzyme data and metabolic information. *Nucleic Acids Res* **30**, 47-49 (2002).
14. U. Wittig *et al.*, SABIO-RK—database for biochemical reaction kinetics. *Nucleic Acids Res* **40**, D790-D796 (2012).
15. D. Heckmann *et al.*, Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nat Commun* **9**, 5252 (2018).
16. J. A. Adams, Kinetic and catalytic mechanisms of protein kinases. *Chemical reviews* **101**, 2271-2290 (2001).
17. R. Otten *et al.*, Rescue of conformational dynamics in enzyme catalysis by directed evolution. *Nat Commun* **9**, 1314 (2018).
18. J. Planas-Iglesias *et al.*, Computational design of enzymes for biotechnological applications. *Biotechnol. Adv.* **47**, 107696 (2021).
19. F. Li *et al.*, Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction. *Nat Catal* **5**, 662-672 (2022).
20. 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. *Nat Commun* **14**, 4139 (2023).
21. H. Yu, H. Deng, J. He, J. D. Keasling, X. Luo, UniKP: a unified framework for the prediction of enzyme kinetic parameters. *Nat Commun* **14**, 8211 (2023).
22. C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review* **54**, 1937-1967 (2021).
23. P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees. *Machine learning* **63**, 3-42 (2006).
24. H. Y. Koh, A. T. Nguyen, S. Pan, L. T. May, G. I. Webb, Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data. *Nature Machine Intelligence*, 1-15 (2024).
25. R. Mitra *et al.*, Geometric deep learning of protein–DNA binding specificity. *Nat Methods* **21**, 1674-1683 (2024).

26. S. Choudhury, B. Narayanan, M. Moret, V. Hatzimanikatis, L. Miskovic, Generative machine learning produces kinetic models that accurately characterize intracellular metabolic states. *Nat Catal* **7**, 1086-1098 (2024).
27. Z. Lin *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
28. R. B. Leveson-Gower, C. Mayer, G. Roelfes, The importance of catalytic promiscuity for enzyme design and evolution. *Nature Reviews Chemistry* **3**, 687-705 (2019).
29. T. Pan *et al.*, SCREEN: A Graph-based Contrastive Learning Tool to Infer Catalytic Residues and Assess Enzyme Mutations. *Genomics Proteomics Bioinformatics*, qzae094 (2025).
30. L. Falivene *et al.*, Towards the online computer-aided design of catalytic pockets. *Nat. Chem.* **11**, 872-879 (2019).
31. A. Elnaggar *et al.*, ProtTrans: Towards cracking the language of Life's code through self-supervised deep learning and high performance computing. arXiv 2020. *arXiv preprint arXiv:2007.06225*, (2007).
32. S. Honda, S. Shi, H. R. Ueda, Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738*, (2019).
33. G. Corso, L. Cavalleri, D. Beaini, P. Liò, P. Veličković, Principal neighbourhood aggregation for graph nets. *Advances in Neural Information Processing Systems* **33**, 13260-13271 (2020).
34. Z. Jia, S. Lin, M. Gao, M. Zaharia, A. Aiken, Improving the accuracy, scalability, and performance of graph neural networks with roc. *Proceedings of Machine Learning and Systems* **2**, 187-198 (2020).
35. I. Elia, M. C. Haigis, Metabolites and the tumour microenvironment: from cellular mechanisms to systemic metabolism. *Nature metabolism* **3**, 21-32 (2021).
36. S. Das *et al.*, Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. *Bioinformatics* **31**, 3460-3467 (2015).
37. S. Zhao *et al.*, Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature* **502**, 698-702 (2013).
38. A. J. M. Ribeiro *et al.*, Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res* **46**, D618-D623 (2018).
39. A. D. Hegeman, J. W. Gross, P. A. Frey, Probing catalysis by Escherichia coli dTDP-glucose-4, 6-dehydratase: identification and preliminary characterization of functional amino acid residues at the active site. *Biochemistry* **40**, 6598-6610 (2001).
40. A. S. Murkin *et al.*, Neighboring group participation in the transition state of human purine nucleoside phosphorylase. *Biochemistry* **46**, 5038-5049 (2007).
41. L. Gerasimavicius, B. J. Livesey, J. A. Marsh, Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nat Commun* **13**, 3895 (2022).
42. G. Li *et al.*, Bayesian genome scale modelling identifies thermal determinants of yeast metabolism. *Nat Commun* **12**, 190 (2021).
43. Y. Chen, J. Nielsen, In vitro turnover numbers do not reflect in vivo activities of yeast enzymes. *Proceedings of the National Academy of Sciences* **118**, e2108391118 (2021).
44. G.-S. Han, G. M. Carman, Characterization of the human LPIN1-encoded phosphatidate phosphatase isoforms. *J Biol Chem* **285**, 14628-14638 (2010).
45. W. Auiewiriyanukul, W. Saburi, K. Kato, M. Yao, H. Mori, Function and structure of GH 13_31 α-glucosidase with high α‐(1→4)-glucosidic linkage specificity and transglucosylation activity. *FEBS letters* **592**, 2268-2281 (2018).
46. J. Meier *et al.*, Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems* **34**, 29287-29303 (2021).
47. J. Dauparas *et al.*, Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49-56 (2022).
48. S. R. Johnson *et al.*, Computational scoring and experimental evaluation of enzymes generated by neural networks. *Nat Biotechnol*, 1-10 (2024).

49. S. Cocco, L. Posani, R. Monasson, Functional effects of mutations in proteins can be predicted and interpreted by guided selection of sequence covariation information. *Proceedings of the National Academy of Sciences* **121**, e2312335121 (2024).
50. M. M. Mader, P. A. Bartlett, Binding energy and catalysis: the implications for transition-state analogs and catalytic antibodies. *Chemical reviews* **97**, 1281-1302 (1997).
51. M. Kobayashi *et al.*, Calcium ion-dependent increase in thermostability of dextran glucosidase from Streptococcus mutans. *Bioscience, biotechnology, and biochemistry* **75**, 1557-1563 (2011).

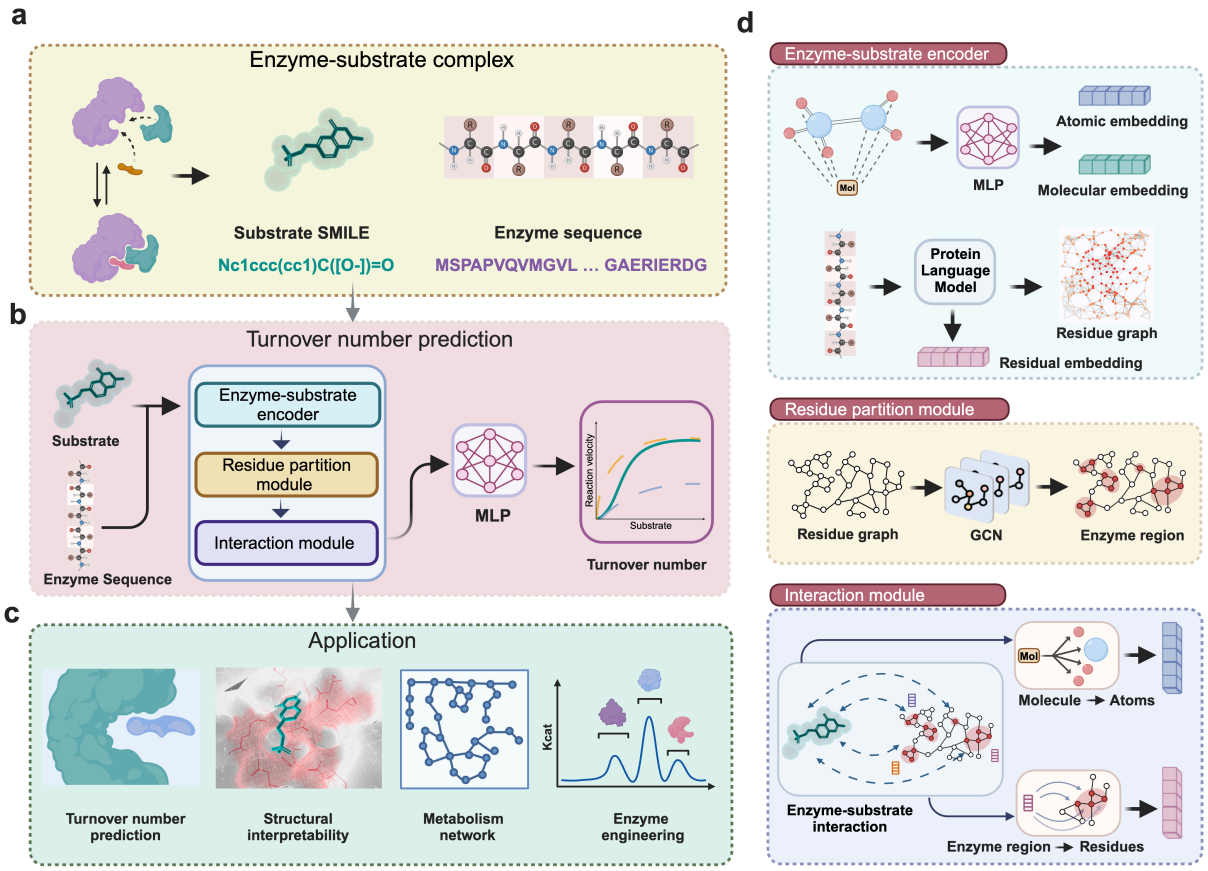


Fig. 1. Overview of KcatNet. a KcatNet input: KcatNet takes enzyme sequences paired with substrate SMILES representations as inputs. **b** Architecture schematic: High-level visualization of the KcatNet model architecture. **c** Application: Predicted K_{cat} values generated by KcatNet are applied to downstream analyses. **d** Detailed components: Enzyme-substrate encoder: Substrates are represented at two levels—atomic (using atom type and property embeddings) and molecular (processed via a pretrained SMILES transformer to extract molecular fingerprints). Enzymes are modeled as residue-level graphs, where residue features are derived from pretrained language models (ProtT5 and ESM2) and linked through ESM2-derived contact maps; Residue partition module: Neighboring residues within the enzyme graph are identified and organized into local regions using graph convolutional networks (GCNs); Interaction module: Molecular fingerprints and enzyme region embeddings are iteratively refined using attention networks to model enzyme-substrate interactions. Substrate fingerprints are decomposed into atomic features, while enzyme regions are broken down into updated residue embeddings. This process is repeated, refining both enzyme regions and molecular fingerprints to produce the final K_{cat} prediction.

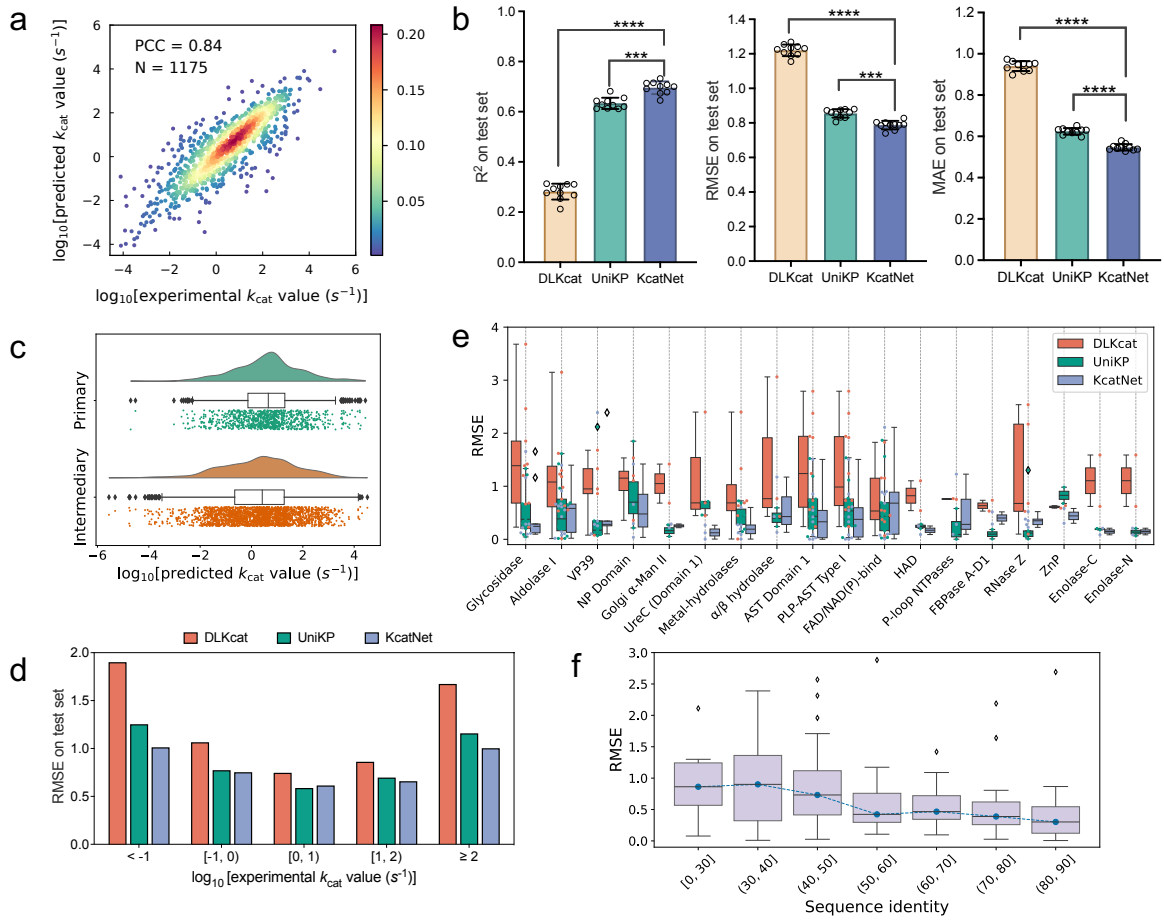


Fig. 2. Predictive performance of KcatNet. **a** Scatter plot showing the Pearson coefficient correlation (PCC) between experimentally measured K_{cat} values and KcatNet-predicted K_{cat} values for the test dataset (n =1175). The color gradient represents the density of data points, ranging from blue (low density) to red (high density). **b** Performance comparison of KcatNet with baseline models DLKcat and UniKP in terms of the Root Mean Square Error (RMSE) (left) and Coefficient of Determination (R2) (right) between experimentally measured K_{cat} values and predicted K_{cat} values. Error bars represent the standard deviation of the mean based on 10 independent runs. Significance levels are indicated (two-sided t-test results: $***$, $P \leq 0.001$; $****$, $P \leq 0.0001$). **c** Distinct KcatNet predictions for enzymes in central energy metabolism versus intermediary and secondary metabolism. **d** Evaluation of KcatNet predictions for enzymes with varying K_{cat} values using RMSE scores on the test datasets. **e** Performance comparison of KcatNet and baseline models DLKcat and UniKP across various protein families based on RMSE. **f** RMSE for the test sets for KcatNet for different levels of maximal enzyme sequence identity compared to enzymes in the training set. For the box plots in **e** and **f**, the lower limit represents the lower quartile, the middle line represents the median, and the upper limit represents the upper quartile, respectively.

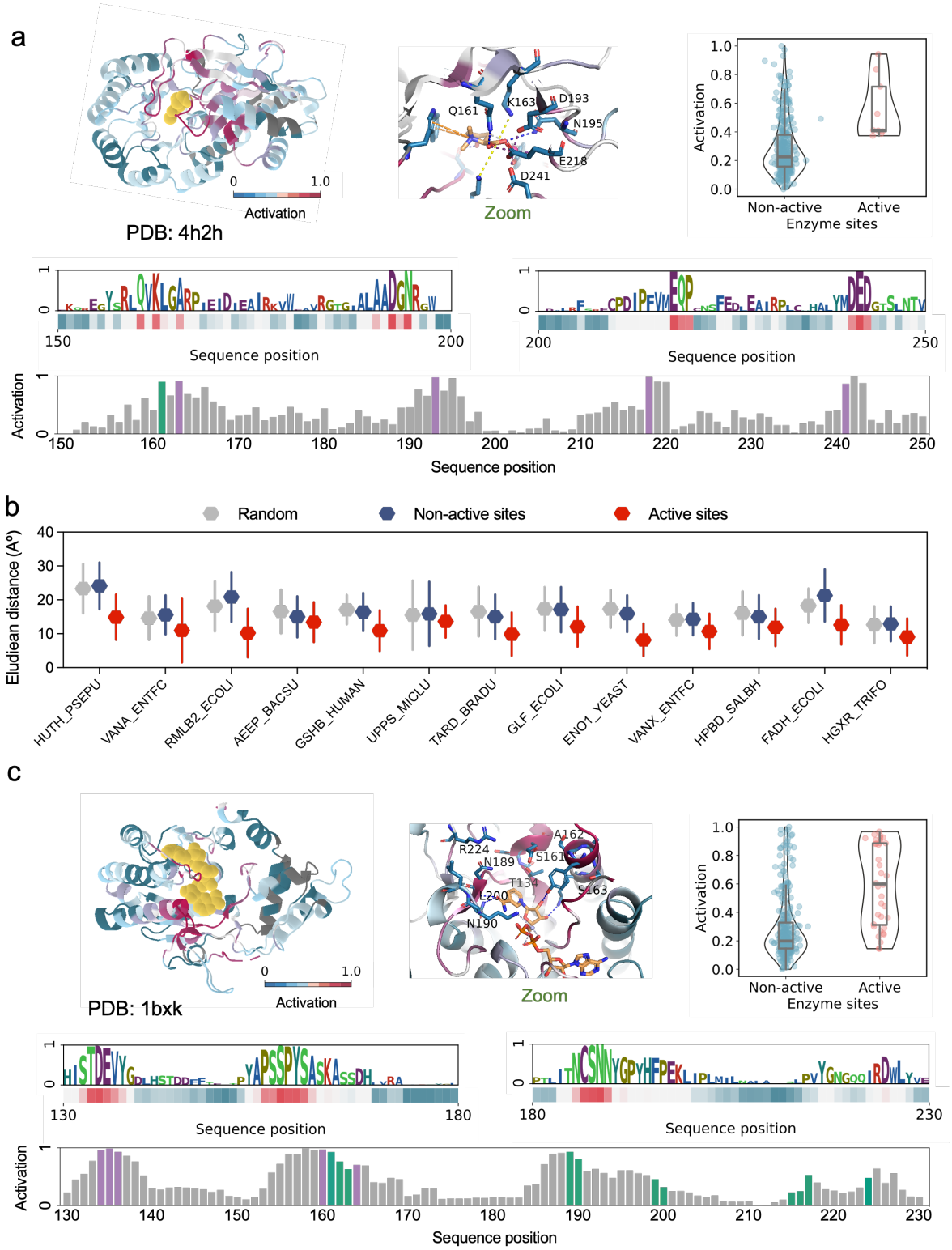


Fig. 3. Analysis and interpretation of putative K_{cat} values generated by KcatNet. a Examples of important residues identified by KcatNet, showing consistency with enzyme functional sites for enolase (PDB ID: 4H2H, Chain A). The normalized attention weights (activation scores) are mapped onto the enzyme's three-dimensional structure (upper panel), highlighting important residues in red and non-important residues in green. The substrate is represented in yellow. Key regions are magnified

to display their detailed cartoon representation. Activation scores for both active and non-active sites are calculated, showing that important sites are typically assigned high activation scores (upper right). The sequence logo of residue importance scores is plotted (medium panel), where the color gradient represents the scores, ranging from blue (low activation scores) to red (high activation scores). Activation scores across the enzyme sequence are plotted (lower panel), with catalytic residues highlighted in purple, substrate binding sites marked in green, and the remaining sequence positions shown in gray. **b** Important residues identified by KcatNet with high activation scores are generally located close to enzyme catalytic sites. For each protein, we show the distribution of distances between the identified active sites (with high activation scores) and non-active sites (with low activation scores) and the nearest active site residue (from M-CSA). **c** Examples of important residues identified by KcatNet, showing consistency with enzyme functional sites for dTDP-glucose 4,6-dehydratase from *E. coli* (PDB ID: 1bxk, Chain A). Similar to **a,** activation scores across the enzyme sequence are plotted (lower panel), with catalytic residues highlighted in purple and substrate binding sites marked in green.

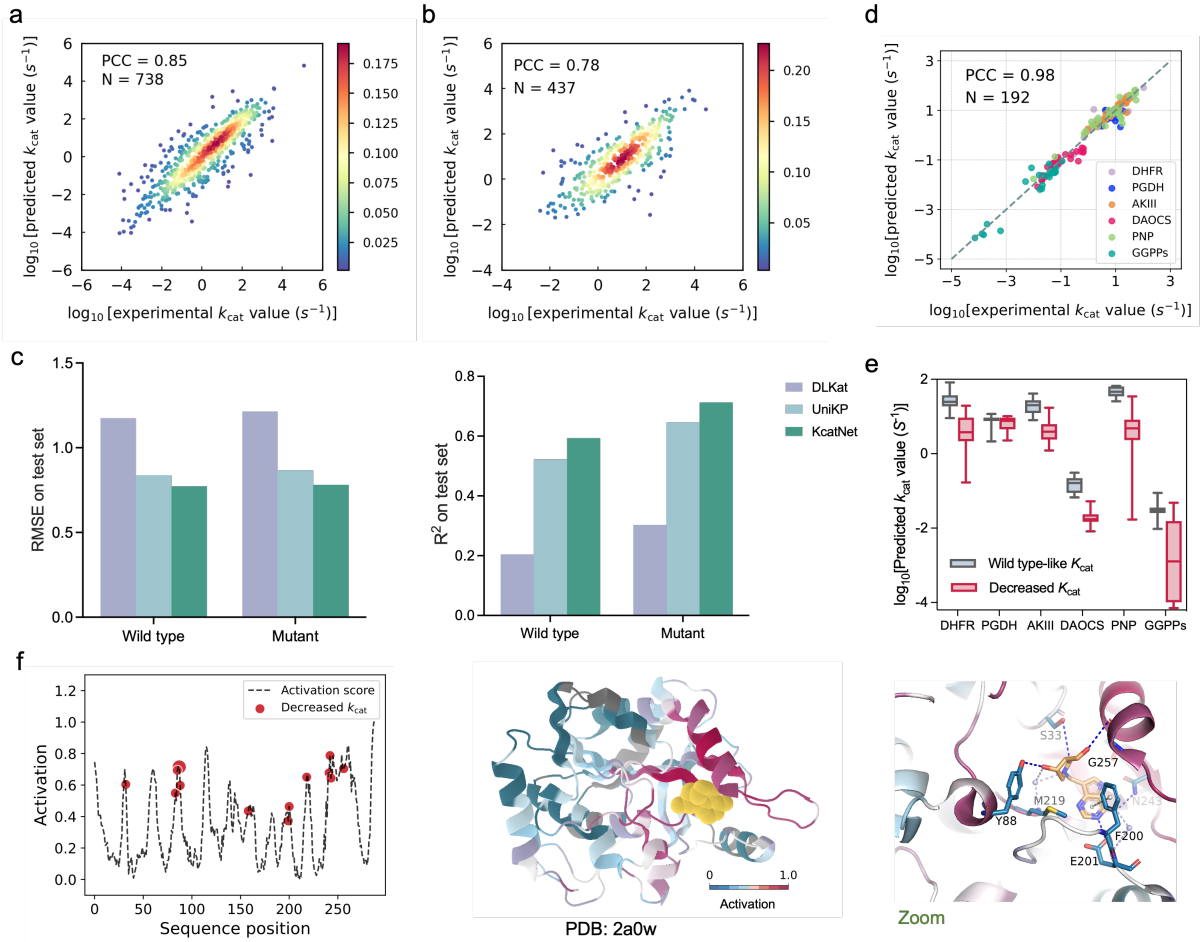


Fig. 4. KcatNet discriminates K_{cat} values of enzymes and their mutants. a-b Scatter plots illustrating the Pearson coefficient correlation (PCC) between experimentally measured and predicted K_{cat} values for wild-type enzymes (**a**) (*N*= 738) and mutant enzymes (**b**) (*N*= 437). **c** Comparison of KcatNet with baseline models DLKcat and UniKP based on the Root Mean Square Error (RMSE) (left), and Coefficient of Determination (R2) (right) between experimentally measured and predicted K_{cat} values for wild-type and mutant enzymes. **d** Comparison between predicted and measured K_{cat} values for several well-studied enzyme–substrate pairs with extensive experimental mutagenesis data. Enzyme abbreviations: DHFR, dihydrofolate reductase; PGDH, d-3-phosphoglycerate dehydrogenase; AKIII, aspartokinase III; DAOCS, deacetoxycephalosporin C synthase; PNP, purine nucleoside phosphorylase; GGPPs, geranylgeranyl pyrophosphate synthase. Substrate abbreviations: G3P, glycerate 3-phosphate; L-Asp, l-aspartate; IPP, isopentenyl Diphosphate. **e** Comparison of predicted K_{cat} values for several mutated enzyme–substrate pairs between enzymes with wild-type-like K_{cat} and decreased K_{cat}. **f** Normalized attention weights (activation scores) of wild-type PNP enzyme (PDB ID: 2A0W, Chain A) using inosine as the substrate (left). Red dots correspond to mutated residues with decreased K_{cat}, marked on the curve according to their residue positions. Dot size indicates the number of mutant enzymes with mutations at that residue. Activation scores of residues predicted by KcatNet are mapped onto the enzyme structure (middle). Residues with high activation scores are highlighted in red (important), while others are shown in green (non-important). The substrate is depicted in purple. Important regions are magnified to show the cartoon representation (right).

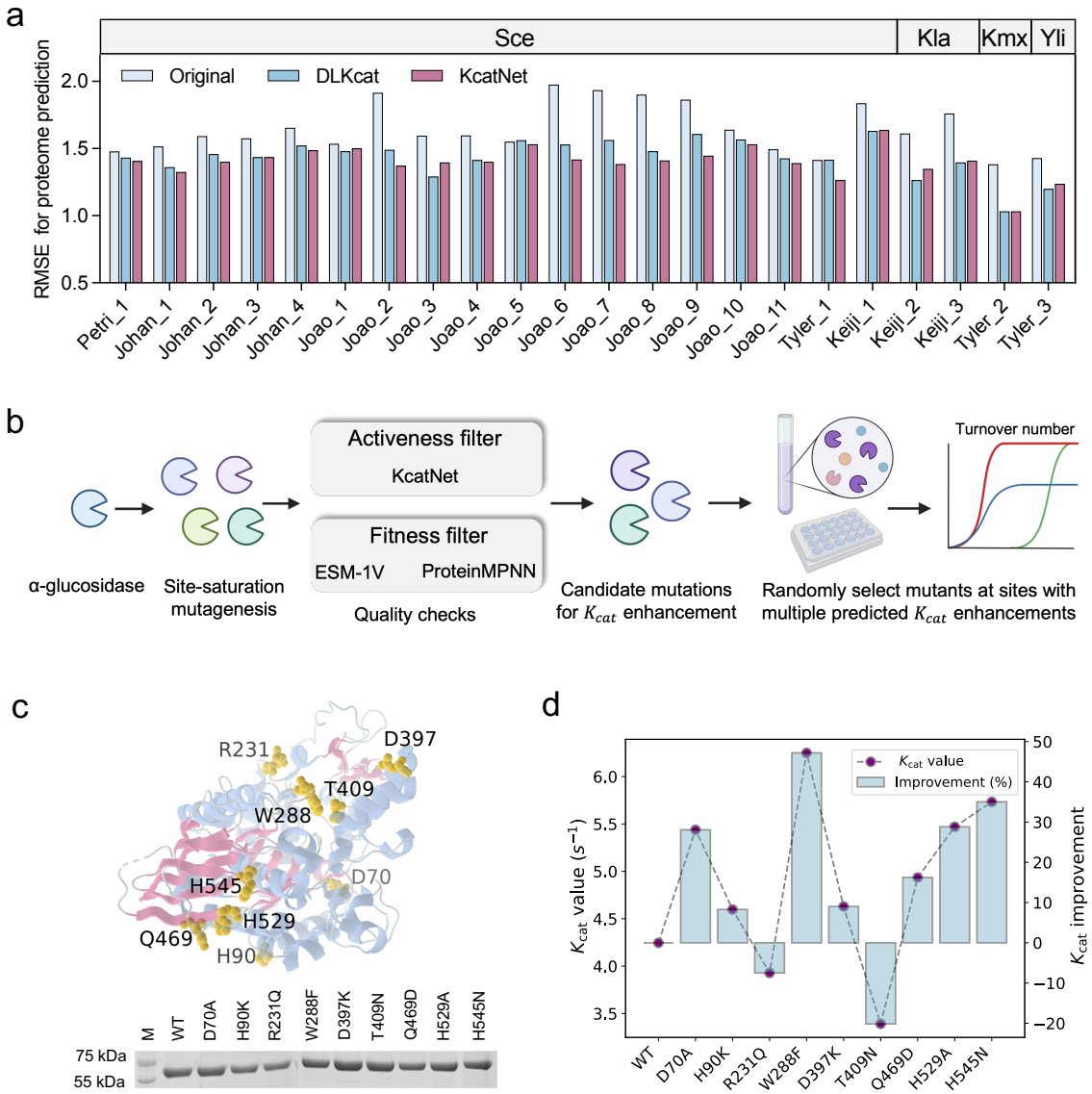
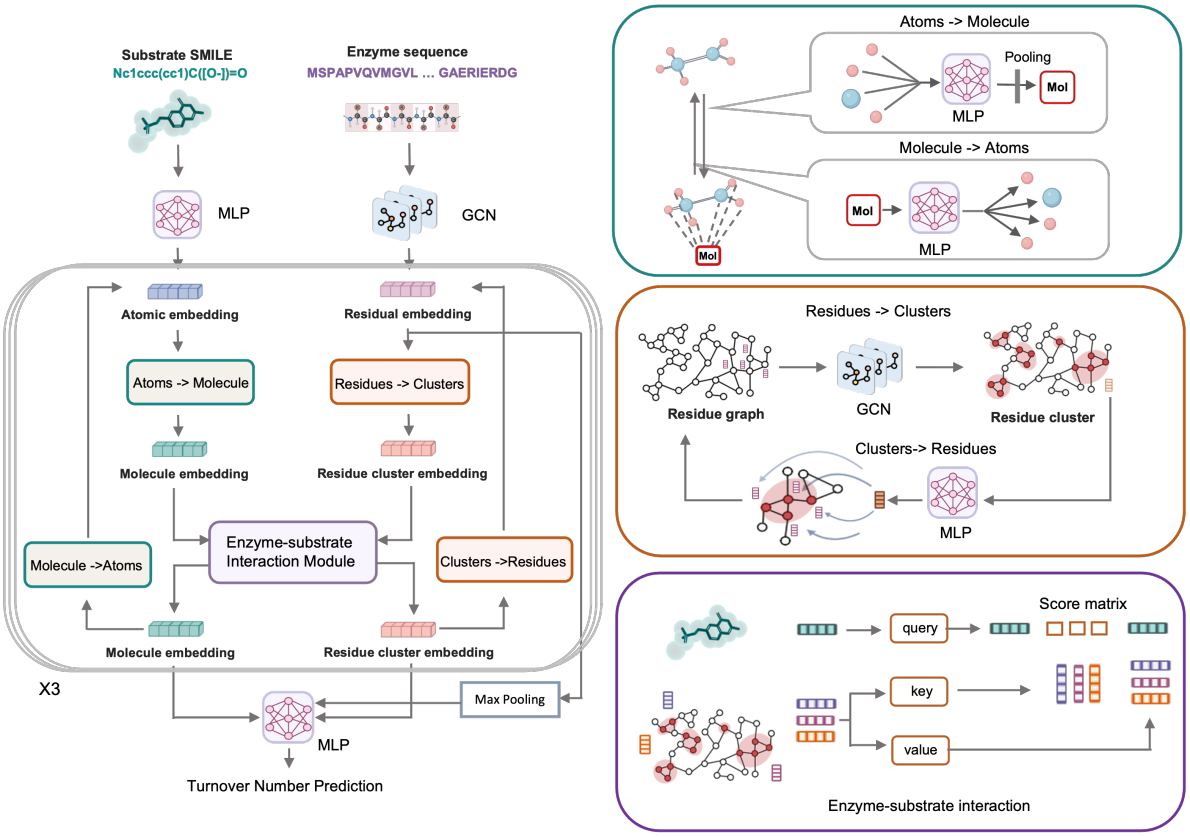
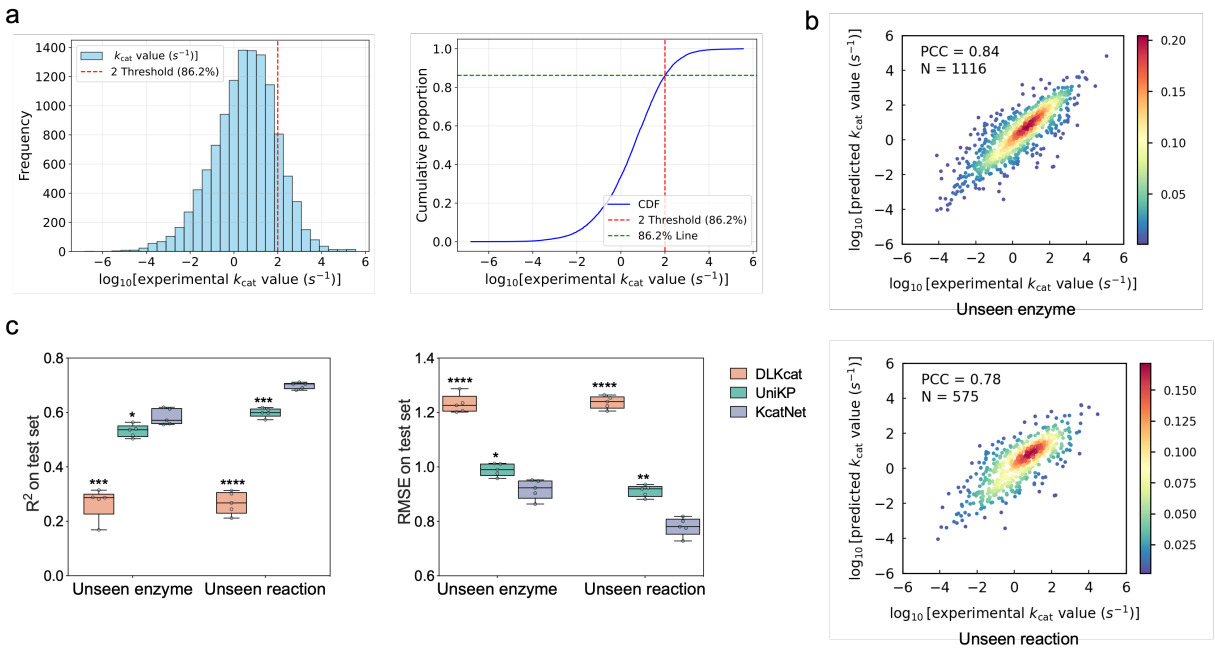


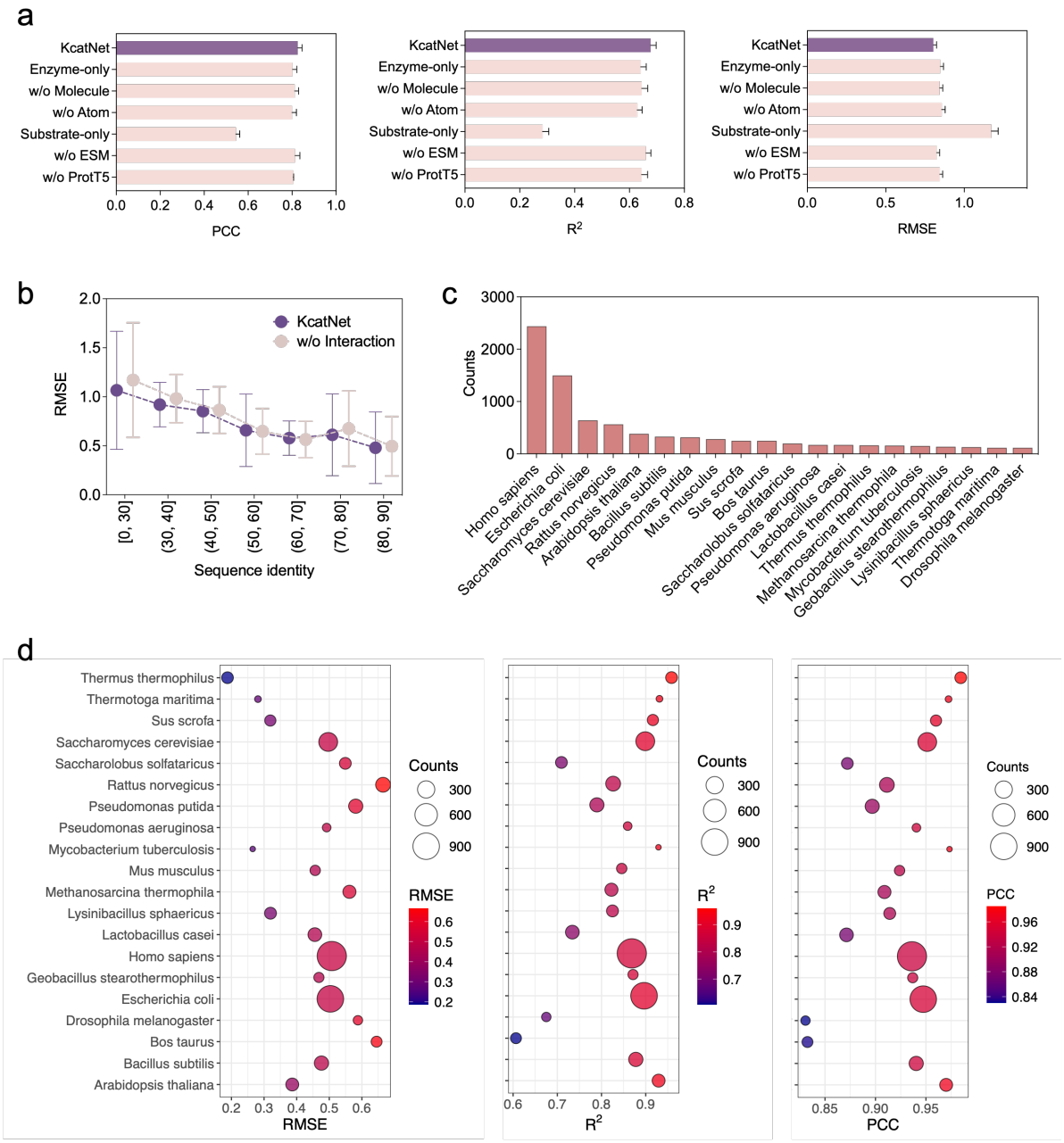
Fig. 5. Application of KcatNet to ecGEMs and enzyme engineering. **a** Comparison of the original-ecGEM with ecGEMs parameterized using K_{cat} values from KcatNet and DLKcat for predicting quantitative proteome data, evaluated against experimental results using root mean square error (RMSE). Proteome predictions were generated for four yeast species (Sce, *Saccharomyces cerevisiae*; Kla, *Kluyveromyces lactis*; Kmx, *Kluyveromyces marxianus*; Yli, *Yarrowia lipolytica*) across distinct culture conditions. **b** KcatNet-based *in silico* screening of 10,545 α-glucosidase missense mutants to identify potential variants with improved K_{cat} for experimental validation. **c** Experimental impact of α-glucosidase missense variants: The nine mutated residues are plotted on the 3D structure, along with their protein abundance assessed by western blot (left) and experimentally measured K_{cat} values of the variants (right).



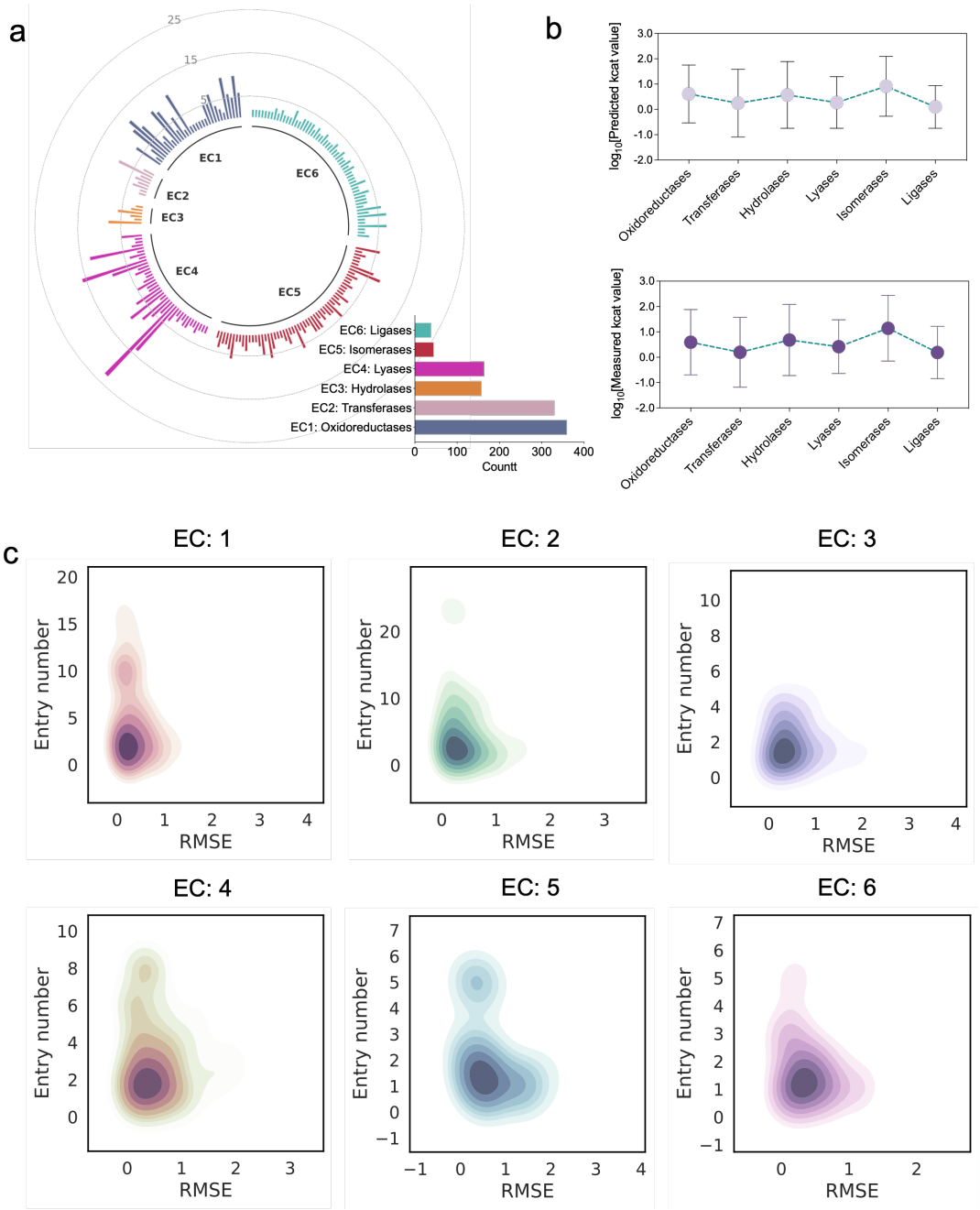
Extended Data Fig. 1. Schematic diagram of KcatNet architecture for predicting enzyme turnover numbers. The substrate representations are processed through a Multi-Layer Perceptron (MLP) to update atomic embeddings, while the enzyme representations are passed through a graph convolutional network (GCN) to obtain updated residue embeddings. These embeddings are further refined through the Atoms-to-Molecule and Residues-to-Clusters modules to generate molecule and residue cluster embeddings, respectively. The enzyme-substrate interaction module facilitates message passing between the substrate and residue clusters, refining the molecule and residue cluster embeddings. Additional transformations, including Molecule-to-Atom and Residue Cluster-to-Residue mappings, are applied before the embeddings are fed into a final MLP for turnover number prediction, with max pooling applied to the residue-level embeddings. Substrate representation is embedded at both atomic and molecular levels, with the Atoms-to-Molecule module passing messages from atomic to molecular embeddings, weighted by importance scores calculated through an MLP. The Molecule-to-Atom module allows the reverse process. The Residues-to-Clusters module groups residues into clusters based on spatial proximity using GCN graph pooling, while the Cluster-to-Residue mapping unpools the clustered regions. The enzyme-substrate interaction module uses cross-attention to capture interactions between enzyme residue clusters and substrates. Refer to the Method section for details.



Extended Data Fig. 2. Distribution of K_{cat} values and model performance on unseen enzymes and reactions. a Distribution of K_{cat} values in the entire dataset (left). Cumulative distribution function (CDF) of the experimental K_{cat} values (right). The green dashed line represents the 86.2% threshold for reference, and the red dashed line indicates the K_{cat} values corresponding to the 86.2% threshold of the data. b Scatter plot showing the Pearson coefficient correlation (PCC) between experimentally measured K_{cat} values and KcatNet-predicted K_{cat} values for novel enzymes (upper, $\boldsymbol{n}$ = 1116) and novel reactions (lower, $\boldsymbol{n}$ = 575) in the test dataset. c Comparison of KcatNet with baseline models DLKcat and UniKP based on the Coefficient of Determination (R2) (left) and Root Mean Square Error (RMSE) (right) for novel enzymes and novel reactions in the test dataset. For boxplots, the center line represents the median, while the top and bottom edges mark the first and third quartiles, respectively. Statistical significance levels are indicated (two-sided ***t***-test: ∗, $\boldsymbol{P} \leq 0.05$; ∗∗, $\boldsymbol{P} \leq 0.01$; ∗∗∗, $\boldsymbol{P} \leq 0.001$; ∗∗∗∗, $\boldsymbol{P} \leq 0.0001$).



Extended Data Fig. 3. Ablation study of KcatNet and application of KcatNet to enzymes in various organisms. a Importance of different input features for KcatNet performance on the test datasets. Feature importance was assessed by removing specific features and evaluating their impact on model performance in terms of PCC (left), R^2 (medium), and RMSE (right). Considered features included enzyme representation-based features (embeddings from ProtT5 and ESM models) and substrate representation-based features (atom-level features using atom type and property embeddings, and molecule-level features processed via a pretrained SMILES transformer). **b** Comparison of model performance at varying sequence identities, where the enzyme-substrate interaction module of KcatNet is excluded (denoted as w/o Interaction). Instead, substrate and enzyme features are concatenated directly. Excluding the interaction module leads to reduced performance, highlighting the importance its importance in the model. **c** Enzymes in the test dataset are grouped by their organisms, showing the enzyme counts within each organismic category. **d** Performance of KcatNet in predicting *in vitro* K_{cat} values for enzymes across diverse organisms, evaluated using RMSE, R^2. and PCC.



Extended Data Fig. 4. Application of KcatNet to enzymes with varying functions. **a** Primary EC number-based categorization of enzymes: Enzymes in the test dataset are grouped by their primary EC number, showing the enzyme counts within each functional category. **b** Predicted K_{cat} values by KcatNet for enzymes within each functional category (upper panel) and experimental K_{cat} values (lower panel) for enzymes across all EC classes. **c** Effect of the number of identical catalytic reactions in the training dataset on predictive performance across enzyme function categories.