GotEnzymes2: expanding coverage of enzyme kinetics and thermal properties

Nucleic Acids Research, 2025, gkaf1053
https://doi.org/10.1093/nar/gkaf1053
Database issue

GotEnzymes2: expanding coverage of enzyme kinetics and thermal properties

Bingxue Lyu[1,2,†], Ke Wu[1,2,†], Yuanyuan Huang[3], Mihail Anton[4,5], Xiongwen Li[1,2], Sandra Viknander[4], Danish Anwer[4], Yunfeng Yang[6], Diannan Lu[7], Eduard Kerkhoven[4,8,9], Aleksej Zelezniak[4,10,11], Dan Gao[1,*], Yu Chen[3,*], Feiran Li[1,2,*]

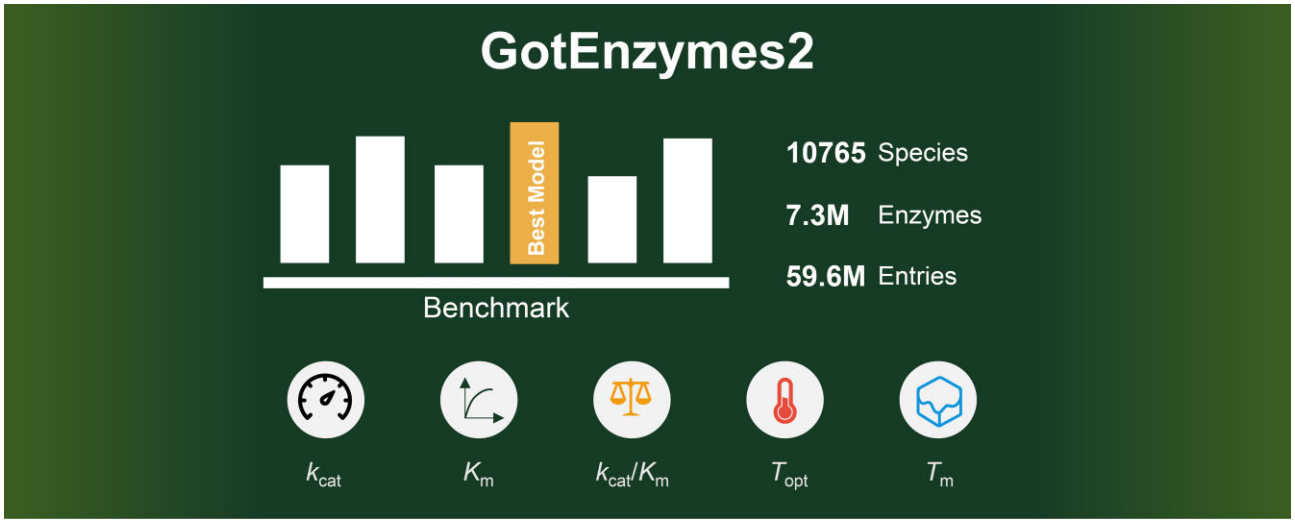
[1]Institute of Biopharmaceutical and Health Engineering, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
[2]Key Laboratory for Industrial Biocatalysis, Ministry of Education, Institute of Biochemical Engineering, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China
[3]State Key Laboratory of Quantitative Synthetic Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[4]Department of Life Sciences, Chalmers University of Technology, Gothenburg SE-412 96, Sweden
[5]ELIXIR, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, United Kingdom
[6]Institute of Environment and Ecology, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
[7]Department of Chemical Engineering, Tsinghua University, Beijing 100084, China
[8]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Kongens 2800, Denmark
[9]SciLifeLab, Chalmers University of Technology, GothenburgSE-412 96, Sweden
[10]Randall Centre for Cell & Molecular Biophysics, King's College London, Guy's Campus, London SE1 1UL, United Kingdom
[11]Institute of Biotechnology, Life Sciences Centre, Vilnius University, Vilnius Sauletekio al. 7 LT10257, Lithuania

*To whom correspondence should be addressed. Email: feiranli@sz.tsinghua.edu.cn
Correspondence may also be addressed to Yu Chen. Email: y.chen3@siat.ac.cn
Correspondence may also be addressed to Dan Gao. Email: gao.dan@sz.tsinghua.edu.cn
†The first two authors should be regarded as Joint First Authors.

Abstract

Enzyme kinetics are fundamental for understanding metabolism, yet experimentally measured parameters remain scarce. To address this gap, we introduce GotEnzymes2, a substantially expanded resource covering 10 765 species, 7.3 million enzymes, and 59.6 million unique entries. Compared with the first version, GotEnzymes2 now integrates both catalytic and thermal parameters, enabling unified predictions of k_{cat}, K_m, k_{cat}/K_m, optimal temperature, and melting temperature. This expansion markedly broadens species and enzyme coverage, creating the most comprehensive database of enzyme kinetic and stability parameters to date. To construct the resource, we systematically benchmarked state-of-the-art models for catalytic and thermal parameter prediction, and incorporated the best-performing strategies to ensure accuracy and generalizability. Altogether, GotEnzymes2 provides the community with a powerful resource for data-driven enzyme discovery, design, and engineering, with broad applications in systems biology, metabolic engineering, and synthetic biology. GotEnzymes2 is publicly accessible at https://metabolicatlas.org/gotenzymes.

Graphical abstract



Received: August 9, 2025. **Revised:** September 14, 2025. **Accepted:** September 19, 2025

Introduction

Enzymes, the primary biological catalysts in living organisms, play an essential role in metabolic processes and cellular function [1, 2]. Quantitative characterization of their catalytic efficiency and thermal stability is of both significant theoretical and practical importance for understanding biological metabolism [3, 4], guiding enzyme engineering [5], optimizing industrial bioprocesses, and advancing the field of synthetic biology [6].

Catalytic efficiency is defined by three core kinetic parameters: k_{cat} (turnover number), which represents the maximum number of substrate molecules converted by an enzyme active site per unit time; K_m (Michaelis constant), which represents the substrate concentration required to achieve half of the maximum catalytic rate and measures substrate affinity; and k_{cat}/K_m (catalytic efficiency), which estimates overall catalytic performance. In addition, enzymes are characterized by their thermal properties. An enzyme's optimal temperature (T_{opt}) defines the temperature at which an enzyme exhibits peak activity. Thermal stability is often characterized by the melting temperature, T_m, which measures the enzyme's resistance to denaturation at elevated temperatures. Both T_{opt} and T_m are crucial for understanding enzyme function across diverse environments and are particularly important for industrial applications. However, existing databases that record enzyme kinetic parameters and thermal properties, such as BRENDA [7], SABIO-RK [8], and UniProt [9], have limited coverage of kinetic and thermal properties due to scarcity of the experimental data, posing a significant barrier to the *in silico* rational selection and engineering of enzymes for diverse applications [10]. To address this gap, various computational models have been developed in recent years (Table 1). For kinetic parameter prediction, models including DLKcat [11], TurNuP [12], DLTKcat [13], DeepEnzyme [14], Kroll *et al.*'s model (referred to as Boost_KM) [15], UniKP [16], EITLEM-Kinetics [17], and CataPro [18] have been developed. In parallel, models including TOMER [19] and Seq2Topt [20] have been developed for predicting enzyme thermal properties. These diverse methods have significantly advanced the field of enzyme property prediction, yet challenges remain in benchmarking and generalizability across diverse biological contexts.

Benchmarking enzyme prediction models is difficult due to inconsistent datasets, heterogeneous evaluation metrics, and the variable ability of models to generalize across biologically relevant conditions. Existing approaches often lack rigorous assessment of performance on low-homology sequences and in predicting mutational effects, which are two critical aspects for enabling broader applicability. The absence of standardized evaluations across these aspects has hindered both methodological refinement and real-world deployment. To address this, we propose a three-step strategy: first, we retrain existing models on all kinetic parameters (k_{cat}, K_m, k_{cat}/K_m) and thermal properties (T_{opt}, T_m) using a unified dataset to assess the accuracy, generalizability, and mutational prediction capability, respectively; second, we combine diverse feature representations (e.g. pretrained protein language models) with machine or deep learning model architectures to optimize prediction performance; third, we apply the best-performing models to systematically update and expand the GotEnzymes database [21] with large-scale predictions of kinetic and thermal properties across a diverse set of enzymes and organisms, thereby creating a comprehensive resource for enzyme research and engineering.

Materials and methods

Dataset acquisition

The EITLEM-Kinetics dataset contains kinetic data for multiple enzyme-substrate reactions, including 34 429 k_{cat}, 28 664 K_m, and 13 388 k_{cat}/K_m. These data provide important support for reproducing the DLKcat, UniKP (k_{cat}, K_m, k_{cat}/K_m), and Boost_KM models. During data processing, for reactions as inputs in TurNuP, we used EC numbers annotated in the EITLEM-Kinetics datasets to fill in the reaction completeness, ensuring data accuracy and consistency. For DLTKcat, which requires temperature information, we referenced the temperature data included in the BRENDA [7] and SABIO-RK [8] databases to fill in the necessary temperature parameters. For entries lacking thermal parameters in BRENDA and SABIO-RK, we excluded them from the dataset. For protein structure information, we predicted the 3D structures of all protein sequences using ESMFold [22]. The T_{opt} dataset ($n = 2917$) was obtained from the GitHub repository of TOMER, which was originally obtained from the BRENDA database. To address the T_{opt} imbalance, we doubled the entries with high T_{opt} ($\geq 80^{\circ}C$) by randomly duplicating existing points in this range. This creates a more balanced dataset, reducing bias toward lower T_{opt} values and improving predictions for high-temperature enzymes [19, 20]. The training and test datasets of thermal stability (T_m) were obtained from DeepTM [23] and Meltome Atlas [24]. The T_m training and test datasets had 25 399 and 6350 entries, respectively.

Calculation of protein identity and substrate similarity

We used the MMseqs2 [25] to calculate the identity of protein sequences and the *FingerprintSimilarity* function from *RDKit* to calculate the similarity between substrates.

Results

Comparison of different enzyme kinetic and thermal property prediction models on unified datasets

We began by collecting kinetic parameter prediction models with available code for both enzyme kinetic parameters (k_{cat}, K_m, k_{cat}/K_m) and thermal properties (T_{opt}, T_m), which exhibited significant differences in their original datasets and reported performance (Fig. 1A and B). To benchmark the performance of kinetic prediction models, we adopted the EITLEM-Kinetics datasets to retrain them, since it is currently the largest in scale, integrating relevant data from UniProt [9], BRENDA [7], and SABIO-RK [8]. This dataset contains 34 429 enzyme–substrate pairs for k_{cat}, 28 664 enzyme–substrate pairs for K_m, and 13 388 enzyme–substrate pairs for k_{cat}/K_m (Fig. 1A). In all three datasets, mutants account for ∼40% of all entries (Supplementary Fig. S1A), enabling evaluation of model sensitivity to sequence perturbations. These datasets cover 8000 protein types and 3000 substrates (Supplementary Fig. S1B and C). The k_{cat}, K_m, and k_{cat}/K_m values follow a log-normal distribution (Supplementary Fig. S1D). To evaluate T_{opt} and T_m prediction models, we used the datasets from TOMER [19], DeepTM [23], and Meltome Atlas [24], which contain 2917 T_{opt} entries and 31 749 T_m entries (Supplementary Fig. S1E).

Most existing kinetic models are trained using protein sequence and substrate inputs, which allows for direct retrain-

Table 1. Enzyme kinetic and thermal properties prediction model

Model	Parameters	Input	Characteristics
DLKcat [11]	k_{cat}	Protein sequence and substrate	k_{cat} ($R^2 = 0.49$), integrated in GECKO 3.0
TurNuP [12]	k_{cat}	Protein sequence and reaction fingerprint	k_{cat} ($R^2 = 0.44$) of an entire reaction, unable to differentiate the k_{cat} for each substrate in multi-substrate reactions
DLTKcat [13]	k_{cat}	Protein sequence, substrate, and temperature	k_{cat} at different temperatures ($R^2 = 0.66$)
DeepEnzyme [14]	k_{cat}	Protein sequence, substrate, and protein 3D structure	k_{cat} ($R^2 = 0.58$) utilizing protein 3D structure
Boost_KM [a] [15]	K_m	Protein sequence and substrate	K_m ($R^2 = 0.53$)
UniKP [16]	k_{cat}, K_m, k_{cat}/K_m	Protein sequence and substrate	k_{cat} ($R^2 = 0.67$), K_m ($R^2 = 0.60$), and k_{cat}/K_m ($R^2 = 0.56$), supports temperature and pH inputs
EITLEM-Kinetics [17]	k_{cat}, K_m, k_{cat}/K_m	Protein sequence and substrate	k_{cat} ($R^2 = 0.72$), K_m ($R^2 = 0.69$), and k_{cat}/K_m ($R^2 = 0.68$) utilizing transfer learning
CataPro [18]	k_{cat}, K_m, k_{cat}/K_m	Protein sequence and substrate	k_{cat} (PCC = 0.497), K_m (PCC = 0.633), and k_{cat}/K_m (PCC = 0.413), training on hard set, exhibiting strong robustness
TOMER [19]	T_{opt}	Protein sequence and optimal growth temperature (OGT)	T_{opt} ($R^2 = 0.632$)
Seq2Topt [20]	T_{opt}, T_m	Protein sequence	T_{opt} ($R^2 = 0.57$) and T_m ($R^2 = 0.64$)

[a]Here, we use Boost_KM to refer to the model developed by Kroll et al.

ing with the EITLEM-Kinetics datasets. However, models such as DLTKcat require temperature information, while DeepEnzyme depends on structural information. To accommodate these requirements, we collected the corresponding temperature data through databases (i.e. UniProt [9], BRENDA [7]) and structural data from protein structure prediction models (i.e. ESMFold [22]). In addition, TurNuP was trained using protein sequences and reaction fingerprints, requiring us to extend the dataset with reaction data from BRENDA [7]. Similarly, for T_{opt} and T_m models, we retrained models only when the original training code was available and inputs were limited to either protein sequence alone or in combination with OGT.

After retraining, UniKP (k_{cat}) and EITLEM-Kinetics (k_{cat}) performed the best for k_{cat} prediction, achieving Coefficient of Determination (R^2) values of 0.674 and 0.628, respectively (Fig. 1C). For K_m prediction, the retrained Boost_KM, UniKP, EITLEM-Kinetics, and CataPro achieved R^2 values of 0.607, 0.662, 0.579, and 0.598, respectively (Fig. 1C). For k_{cat}/K_m prediction, the retrained UniKP (k_{cat}/K_m) outperformed EITLEM-Kinetics (k_{cat}/K_m) and CataPro (k_{cat}/K_m), with R^2 values of 0.589, 0.556, and 0.502, respectively (Fig. 1C). The overall better performance of the k_{cat} prediction compared to K_m and k_{cat}/K_m may be attributed to its larger dataset size compared to those for K_m and k_{cat}/K_m. Additionally, the R^2 values of Boost_KM and TurNuP showed improvement compared to their original reports, increasing by 0.08 and 0.17 (compared to original report), respectively, further showing the positive impact of dataset expansion on model accuracy.

For T_{opt} prediction, TOMER [19] and Seq2Topt [20] were chosen due to the code availability for retraining. TOMER is a machine learning model that takes both sequence and OGT as input features, while Seq2Topt is a deep learning model that uses only sequences as input (Fig. 1D). For T_m prediction, only Seq2Topt was retrained (Fig. 1D), and its results outperformed the originally reported performance in its publication. The performance of the retrained models was evaluated using the R^2, Pearson's Correlation Coefficient (PCC), Mean Absolute Error, Spearman Correlation, and Root Mean Square Error, as shown in Supplementary Table S1.

Evaluation of the generalization ability of enzyme kinetics parameters and thermal property prediction models

We systematically evaluated the generalization ability of models for predicting enzyme kinetics and thermal properties, uniquely assessing performance across both protein sequence identity and substrate similarity. For kinetics models, as can be expected, performance declined with decreasing similarity on both axes, with retrained UniKP and Boost_KM showing the most robust generalization for k_{cat}/K_m and K_m predictions, respectively (Fig. 2A and B and Supplementary Fig. S2). We therefore propose that this dual-axis evaluation should become a standard for assessing generalization. In contrast, models predicting thermal properties (T_{opt} and T_m) demonstrated stable performance across a wide range of sequence identities, indicating strong generalization even to distant proteins (Fig. 2C) and in different OGT ranges (Fig. 2D).

Evaluation of enzyme kinetic parameter prediction models on mutants

To assess the models' utility for enzyme engineering, we evaluated their performance on predicting the kinetic parameters of mutants. The retrained UniKP model was superior, achieving high R^2 values on the mutant dataset for k_{cat} ($R^2 = 0.743$), K_m ($R^2 = 0.787$), and k_{cat}/K_m ($R^2 = 0.667$) (Fig. 2E). This high performance was maintained even as the number of mutation sites increased (Fig. 2F). Critically, leading models could also accurately predict the directional impact of mutations on activity; for instance, UniKP (k_{cat}) predicted whether a mutation would increase or decrease k_{cat} with 87.3% accuracy. These findings validate the models' robustness for variant prediction and highlight their potential to guide rational enzyme design. Further details on comparative performance and directional accuracy are available in the supplementary materials (Supplementary Fig. S3).

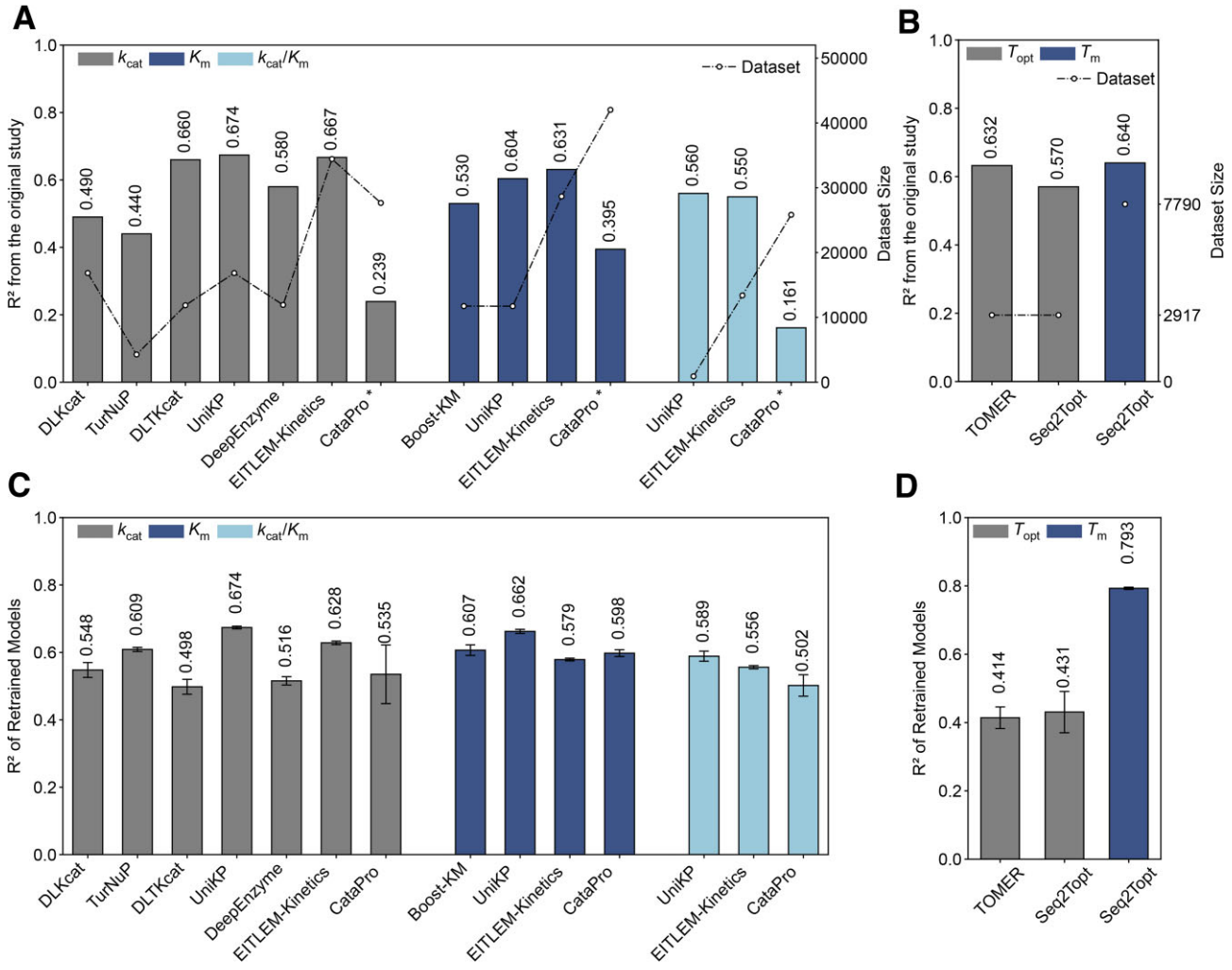


Figure 1. Performance of retrained enzyme kinetic parameter and thermal properties prediction models on unified datasets. **(A)** The dataset size of different enzyme kinetics prediction models and their reported R^2 values. It should be noted that the R^2 for CataPro was calculated using predictions on the test dataset and the corresponding labels. **(B)** The dataset size of different enzyme thermal property prediction models and their reported R^2 values. **(C)** R^2 of different retrained kinetic parameter prediction models on the EITLEM-Kinetics datasets. **(D)** R^2 of different retrained thermal properties prediction models on the T_{opt} and T_m datasets. Error bars represent the standard deviation of the test performance over five random train-test splits of the dataset ($n = 5$). ∗ To be noted here, CataPro employs an unbiased dataset and splits the training and test sets under protein sequence similarity control. This more challenging strategy results in lower R^2 values compared to random splitting of other models.

Optimal module combinations for enzyme kinetic and thermal parameter prediction

To identify the most effective predictive models, we performed a systematic combinatorial screen of key modules, including protein representations, substrate representations, and model architectures (Table 2). For enzyme kinetics, an extensive benchmark of 216 unique configurations revealed that a machine learning architecture (ExtraTrees) paired with large language model representations (ProtT5 for proteins, MolGen for substrates) surpassed existing deep learning models at the current data scale (Fig. 3A–D). This optimal combination, ProtT5&MolGen&ExtraTrees, demonstrated superior performance over all retrained published models, particularly in predicting the parameters of mutants (Fig. 4A–C). Applying a similar strategy to enzyme thermal properties, we identified the combination of ProtT5 and the Seq2Topt architecture as the top performer, which improved R^2 by 0.09 for T_{opt} (compared to retrained result) and 0.20 for T_m over previous state-of-the-art models (Fig. 4D and E).

Table 2. Common protein and substrate representations and model architectures

	Substrate representation	Protein representation	Model architecture
Approach	RDKitFP ECFP MACCSkeys FP	ESM-1b[26]	UniKP (Extra-TreesRegressor)
	Mole-BERT [27]	ESM-1v[28]	DLKcat [attention based multilayer perceptron (MLP)]
	ChemBERTa-2 [29]	ESM2 [22]	EITLEM-Kinetics (attention based MLP)
	UniMol V1 [30] UniMol V2 [31] MolGen [33]	ESM C[a] ProtT5 [32] ProLLaMA [34]	CataPro (MLP)
	SMILES Transformer [35]		

[a]ESM C was from https://www.evolutionaryscale.ai/blog/esm-cambrian

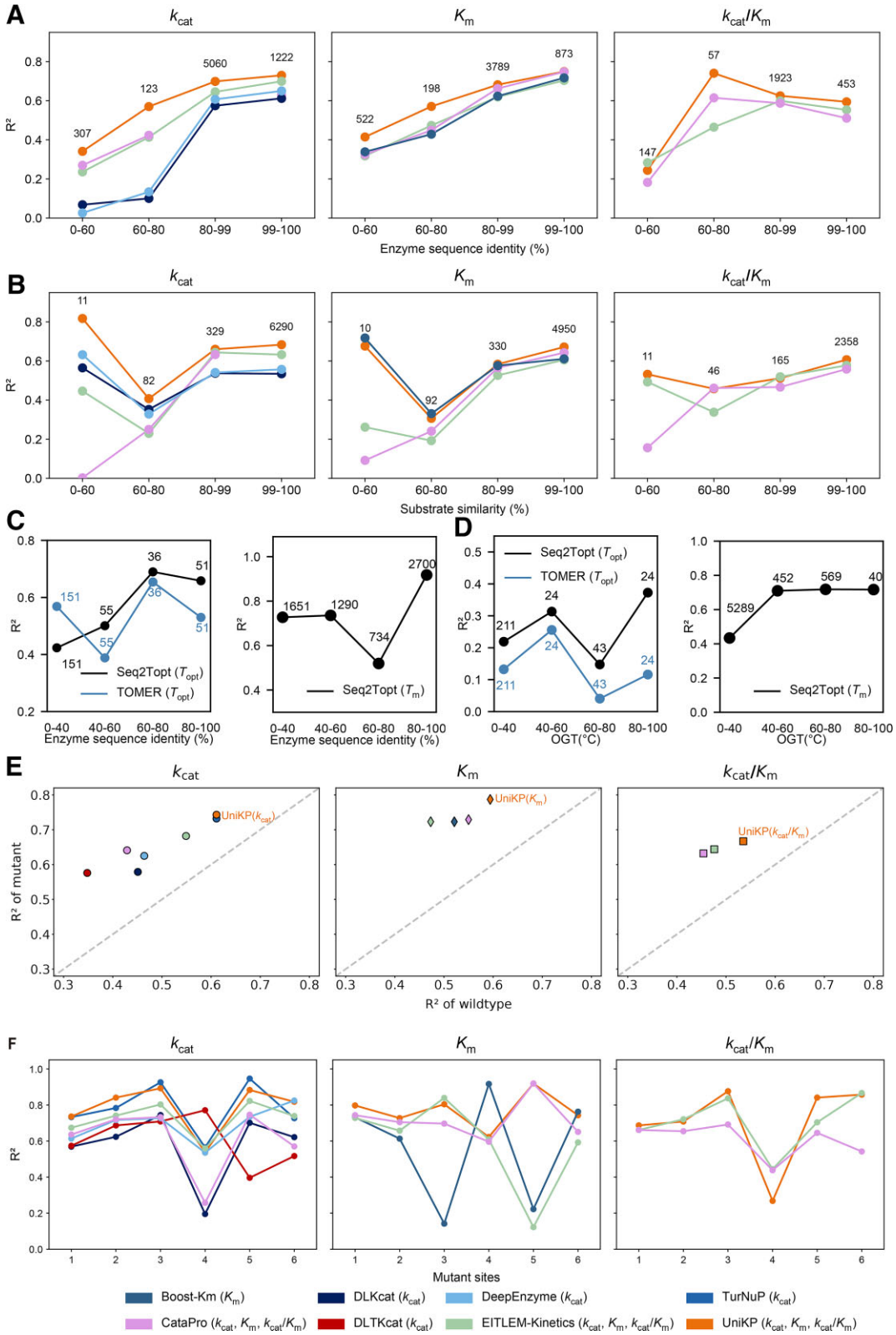


Figure 2. Generalization capabilities of the retrained enzyme kinetic parameter and thermal properties prediction models in the dimensions of protein identity and substrate similarity and performance of the enzyme kinetic parameter models in predicting mutants. Generalization ability of the retrained k_{cat}, K_m, and k_{cat}/K_m prediction models evaluated across **(A)** enzyme sequence identity and **(B)** substrate similarity. **(C)** Generalization ability of the retrained T_{opt} and T_m prediction model. **(D)** Performance of the retrained T_{opt} and T_m prediction model in different OGT intervals. **(E)** R^2 of the retrained model predictions for wild-type and mutants on the test set. Here, circles represent the k_{cat} model, diamonds represent the K_m model, and squares represent k_{cat}/K_m. **(F)** R^2 of the retrained model predictions for mutants with varying numbers of mutation sites on the test set.

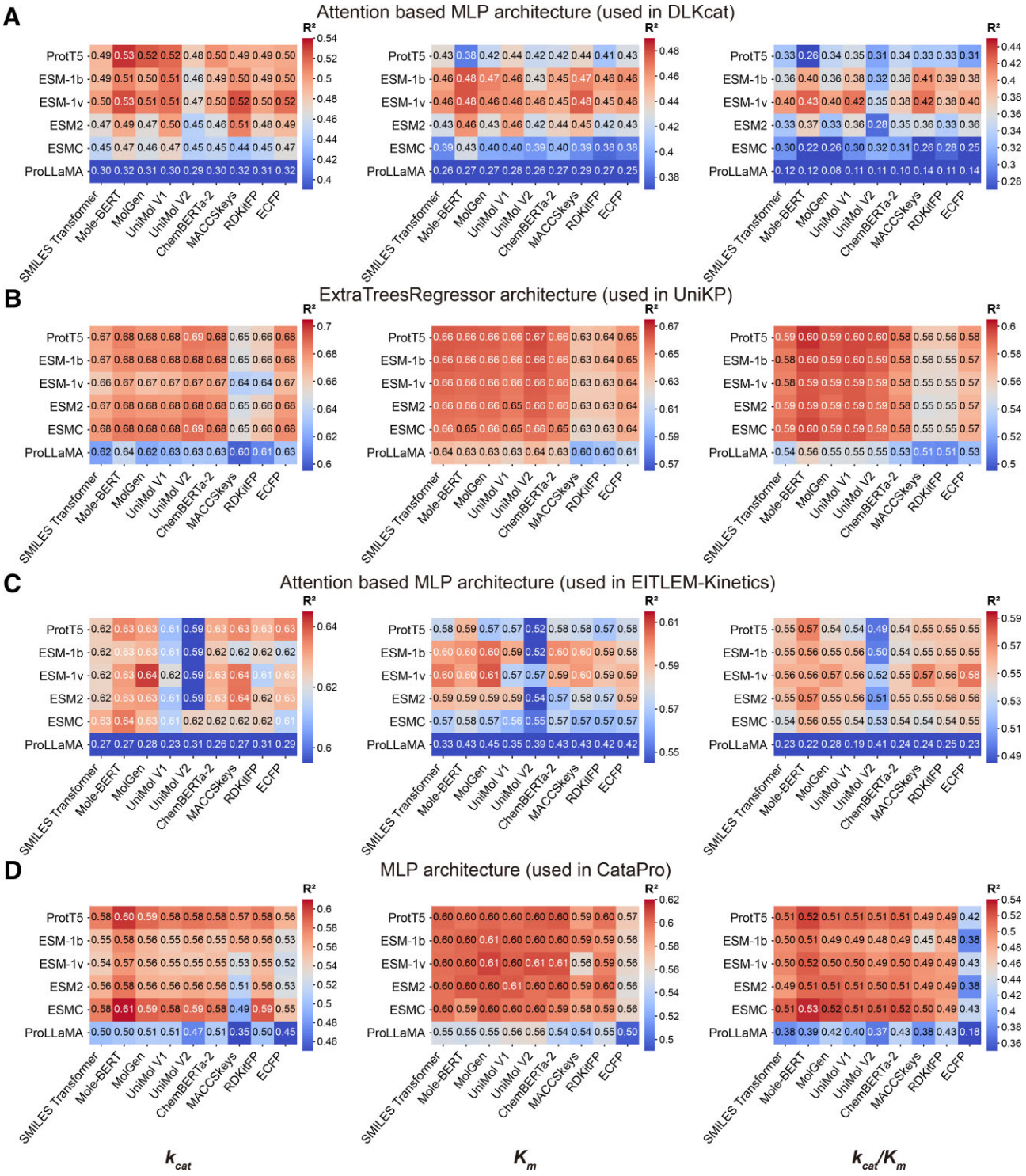


Figure 3. Performance comparison of 216 model configurations. Heatmap showing the R^2 values on the test set for k_{cat}, K_m, and k_{cat}/K_m prediction across all combinations of protein representations, substrate representations, and model architectures. **(A)** Attention-based MLP architecture (used in DLKcat). **(B)** ExtraTreesRegressor architecture (used in UniKP). **(C)** Attention-based MLP architecture (used in EITLEM-Kinetics). **(D)** MLP architecture (used in CataPro).

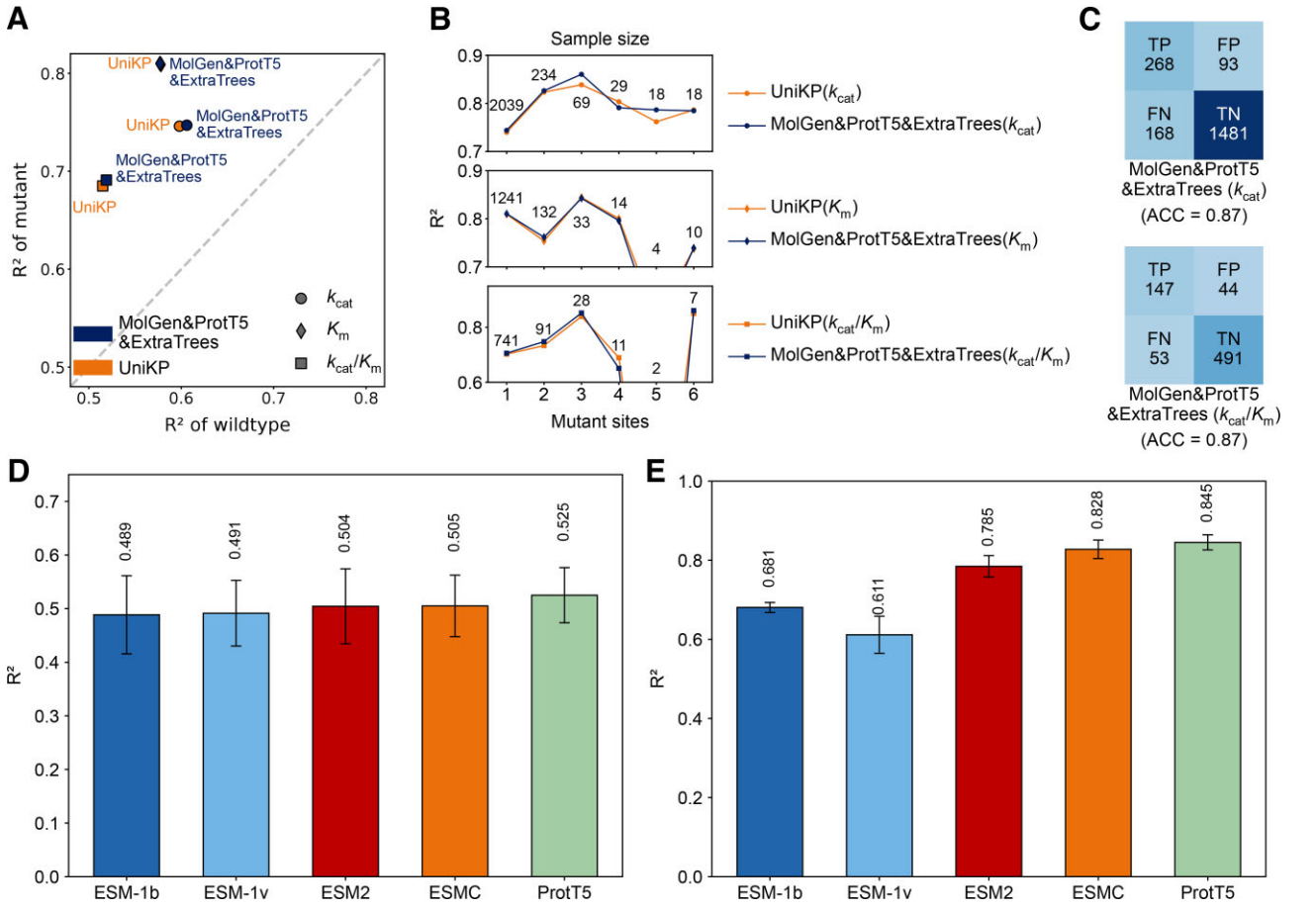


Figure 4. Performance of the optimal combined model. **(A)** Comparison of the R^2 for mutation predictions between the optimal combined model and retrained UniKP (k_{cat}, K_m, k_{cat}/K_m) on the test set. Here, circles represent the k_{cat} model, diamonds represent the K_m model, and squares represent k_{cat}/K_m. **(B)** Comparison of mutation prediction performance across different numbers of mutation sites between the optimal combined model and retrained UniKP (k_{cat}, K_m, k_{cat}/K_m) on the test set. **(C)** Comparison of mutation direction prediction performance between the optimal combined model and retrained UniKP (k_{cat}, K_m, k_{cat}/K_m). **(D)** Performance of combinations of T_{opt} model. **(E)** Performance of T_m model combinations. Error bars represent the standard deviation of the test performance over five random train-test splits of the dataset ($n = 5$).

Table 3. Comparison between GotEnzymes and GotEnzymes2

	GotEnzymes	GotEnzymes2
Species	8099	10 765
Enzymes (million)	5.8	7.3
Entries (million)	25	59.6
Parameters	k_{cat}	k_{cat}, K_m, k_{cat}/K_m, T_{opt}, and T_m

Expansion of the GotEnzymes database

The original GotEnzymes database encompassed predicted k_{cat} values for 25 million enzyme–substrate pairs, covering 5.8 million enzymes from 8099 species. To further expand the dataset, we updated the species list based on the latest KEGG [36] database, increasing the total number of species to 10 765, the number of enzymes to 7.3 million, and the number of enzyme–substrate pairs to 59.6 million in GotEnzymes2 (Table 3). Additionally, we substantially enriched the range of annotated properties. Using our optimal combined enzyme kinetic model (ProtT5&MolGen&ExtraTrees), we extended predictions to include k_{cat}, K_m, and k_{cat}/K_m parameters. For enzyme thermal properties, we employed the best-performing model (ProtT5&Seq2Topt) to predict T_{opt} and T_m (Fig. 5A). These updates transform GotEnzymes2 into a comprehensive and multi-parameter enzyme property resource, facilitating downstream applications in metabolic engineering, enzyme design, and synthetic biology.

Global analysis of enzyme thermal properties

For our global analysis of enzyme thermal properties, to classify the enzymes into thermal categories, we used the OGT of their respective source organisms. This OGT information was sourced from the GOSHA database [37] and linked to our dataset via organism name mapping between GOSHA and KEGG. The sample sizes of organisms were $n = 19$ for psychrophiles, $n = 5696$ for mesophiles, $n = 253$ for thermophiles, and $n = 61$ for hyperthermophiles. As shown in Fig. 5B, the distributions of optimal reaction temperature (T_{opt}) and melting temperature (T_m) for these enzyme groups are clearly distinct. Enzymes from psychrophiles and mesophiles, which are adapted to colder environments, exhibit lower thermal characteristics. Specifically, psychrophilic enzymes display the lowest temperature profiles, while mesophilic enzymes typically have T_{opt} values clustered in the 30°C–50°C range with correspondingly moderate T_m values. While enzymes from thermophilic and hyperthermophilic organisms possess significantly higher T_{opt} and T_m values. Their T_{opt} values are generally above 70°C, with

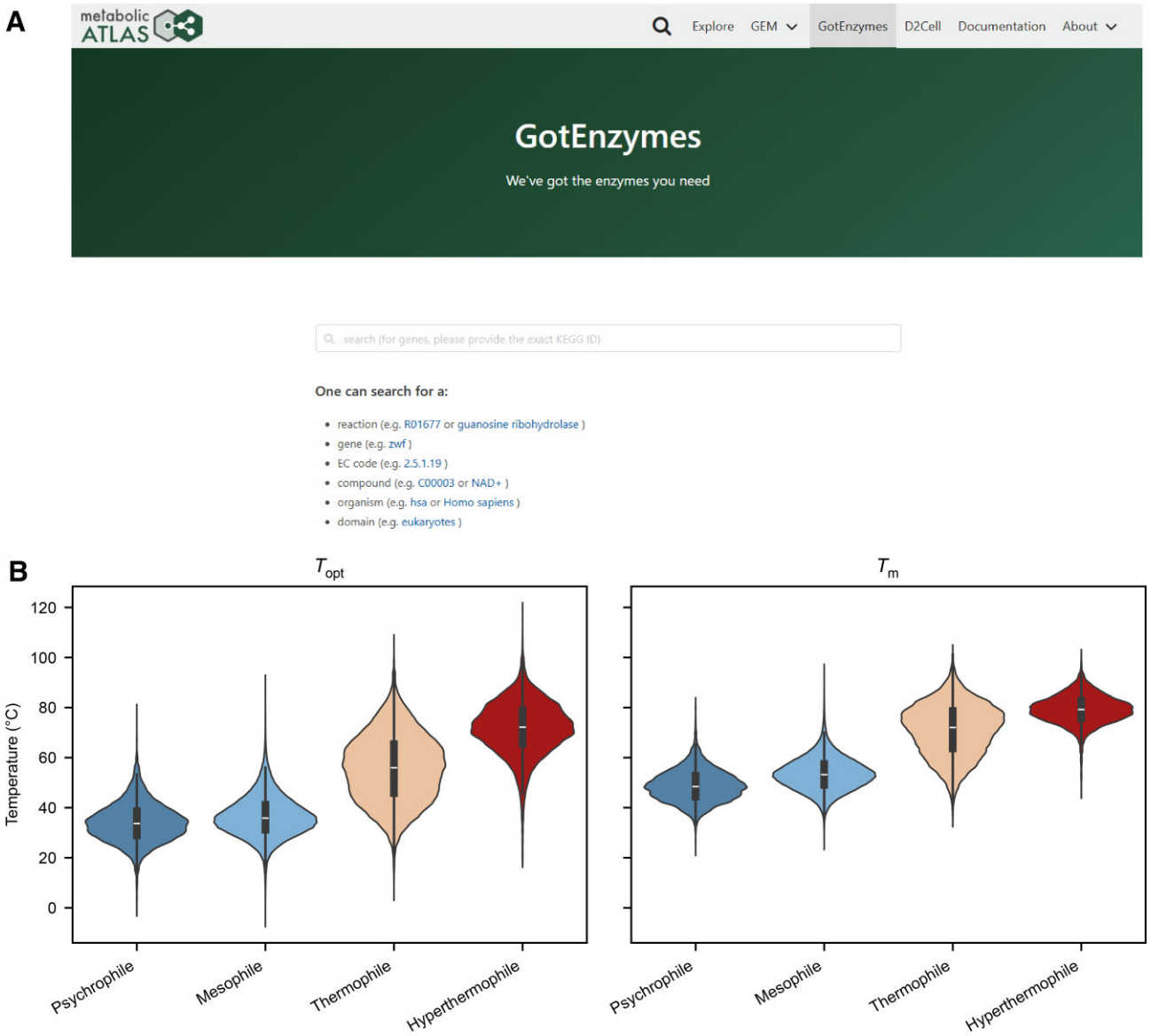


Figure 5. Overview of the GotEnzymes2 database. **(A)** User interface of GotEnzymes2. **(B)** Global analysis of enzyme thermal properties. The sample sizes of organisms were $n = 19$ for psychrophiles, $n = 5696$ for mesophiles, $n = 253$ for thermophiles, and $n = 61$ for hyperthermophiles. The inner box represents the interquartile range (from lower to upper quartile). The central line is the median, and whiskers extend to 1.5× the interquartile range.

some hyperthermophilic enzymes showing peak activity near 100°C, and their elevated T_m values reflect their enhanced thermal stability.

Case study: data-driven sourcing of a thermostable biocatalyst

The industrial modification of starch requires highly thermostable glycogen branching enzymes (GBE, EC 2.4.1.18), as many existing candidates exhibit insufficient stability at high temperatures. The GotEnzymes2 database is designed to address this challenge directly.

Instead of performing laborious literature searches, a user can simply query for EC number "2.4.1.18" within the database and sort the results by melting temperature (T_m) in descending order. This process rapidly generates a shortlist of top-ranking, hyper-thermostable enzymes, providing ideal starting points for protein engineering. This data-driven workflow can significantly accelerate a project's initial phase. For instance, the GBE with UniProt ID O50094 (top 0.2%) could be efficiently identified through this method and selected for subsequent directed mutagenesis [38].

Discussion

Recent years have witnessed substantial progress in the prediction of enzyme properties, including kinetic parameters (k_{cat}, K_m, k_{cat}/K_m) and thermal properties (T_{opt}, T_m), which are crucial for enzyme-constrained modeling and engineering. However, differences in datasets and model performance hinder reproducibility, benchmarking, and widespread adoption. Here, we addressed these limitations through a comprehensive benchmarking framework by retraining leading models on unified large-scale datasets. For kinetic predictions, retrained versions of UniKP and EITLEM-Kinetics emerged as top performers. For thermal properties, Seq2Topt outperformed others after retraining. To assess real-world appli-

cability, we evaluated model generalization to divergent sequences and substrates, as well as performance on mutant enzymes. Notably, retrained UniKP exhibited strong generalization and maintained high accuracy across both wild-type and mutant datasets. Importantly, UniKP, EITLEM-Kinetics, and DeepEnzyme accurately predicted mutation effects, a critical feature for enzyme design. Thermal models showed stable performance across low-homology sequences, suggesting an ability to capture more global determinants of thermostability. To optimize further, we combined advanced protein and molecular representations (e.g. ProtT5, MolGen) with different model architectures. The ProtT5&MolGen&ExtraTrees model improved kinetic predictions, especially for mutants, while ProtT5&Seq2Topt enhanced T_{opt} and T_m prediction. These advances enabled a major update to GotEnzymes, expanding species coverage from 8099 to 10 765 and enzyme–substrate pairs from 25 million to 59.6 million, now including k_{cat}, K_m, k_{cat}/K_m, T_{opt}, and T_m.

In conclusion, our study presents a unified benchmarking framework for enzyme property prediction, identifies optimal model configurations through extensive modular evaluation, and delivers a significantly expanded GotEnzymes2 database encompassing high-accuracy predictions for catalytic and thermal parameters across a broad phylogenetic landscape. However, several challenges remain despite significant advances. Model performance remains constrained by the quality of available data and the limited integration of structural information. Additionally, model outputs can vary substantially across architectures, posing a challenge for interpretability and reliability. Future efforts should prioritize the curation of higher-quality datasets, inclusion of underrepresented enzyme classes, and incorporation of structure-aware representations to drive more consistent and mechanistically grounded predictions. Ultimately, the continued expansion of publicly available, experimentally verified enzyme kinetic and thermal stability data will be the most crucial element for training next-generation models with even higher accuracy and broader applicability.

Acknowledgements

Author contributions: Bingxue Lyu (Data curation [equal], Formal analysis [equal], Methodology [equal], Resources [equal], Software [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Ke Wu (Conceptualization [equal], Data curation [equal], Methodology [equal], Writing—original draft [equal], Writing—review & editing [equal]), Yuanyuan Huang (Data curation [equal], Software [equal], Visualization [equal], Writing—review & editing [equal]), Mihail Anton (Software [equal], Writing—review & editing [equal]), Xiongwen Li (Data curation [equal], Writing—review & editing [equal]), Sandra Viknander (Conceptualization [equal], Software [equal]), Danish Anwer (Software [equal]), Yang Yunfeng (Supervision [supporting], Writing—review & editing [equal]), Diannan Lu (Supervision [supporting], Writing—review & editing [equal]), Eduard Kerkhoven (Supervision [equal], Writing—review & editing [equal]), Aleksej Zelezniak (Supervision [supporting]), Dan Gao (Supervision [equal], Writing—review & editing [equal]), Yu Chen (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), and Feiran Li (Funding acquisition [equal], Supervision [lead], Writing—original draft [equal], Writing—review & editing [equal]) RX

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

F. L. acknowledges financial support from the National Key R&D Program of China (2024YFA0920300), National Natural Science Foundation of China General Project (22478223), Shenzhen Medical Research Fund (A2403013), Tsinghua Shenzhen International Graduate School Cross-disciplinary Research and Innovation Fund (JC2024004), and Department of Chemical Engineering-iBHE Joint Cooperation Fund (DCE-iBHE-2023-1). A. Z. is supported by Biotechnology and Biological Sciences Research Council (BBSRC) grant number BB/Y000730/1, Marius Jakulis Jason Foundation, Swedish Research Council (Vetenskapsrådet) grant no. 2023-04254, 2019-05356, and Formas grant 2019-01403. Funding to pay the Open Access publication charges for this article was provided by National Key R&D Program of China.

Data availability

The unified dataset used for retraining the kinetic parameter prediction models is available via EITLEM-Kinetics (https://github.com/XvesS/EITLEM-Kinetics). The dataset used for retraining the thermal properties prediction models is obtained from DeepTM (https://github.com/liimy1/DeepTM), TOMER (https://github.com/jafetgado/tomer/), and Meltome Atlas (https://meltomeatlas.proteomics.wzw.tum.de/master_meltomeatlasapp/). The KEGG database (https://www.genome.jp/kegg/) was used for the GotEnzymes2 database (https://digitallifethu.com/gotenzymes). The authors declare that all data supporting the findings and enabling the reproduction of all figures in this study are available within the paper and its Supplementary Information. Source data are provided with this paper. All data used in this study can be accessed at https://github.com/LiLabTsinghua/GotEnzymes2. To facilitate further use, we have made all the codes and detailed instructions available in our GitHub repository, located at https://github.com/LiLabTsinghua/GotEnzymes2.

References

1. Knowles JR. Enzyme catalysis: not different, just better. *Nature* 1991;350:121–4. https://doi.org/10.1038/350121a0
2. Breaker RR. DNA enzymes. *Nat Biotechnol* 1997;15:427–31. https://doi.org/10.1038/nbt0597-427
3. Chen Y, Nielsen J. Energy metabolism controls phenotypes by protein efficiency and allocation. *Proc Natl Acad Sci USA* 2019;116:17592–7. https://doi.org/10.1073/pnas.1906569116
4. Sánchez BJ, Zhang C, Nilsson A *et al.* Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol Syst Biol* 2017;13:935. https://doi.org/10.15252/msb.20167411
5. Adadi R, Volkmer B, Milo R *et al.* Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. *PLoS Comput Biol* 2012;8:e1002575. https://doi.org/10.1371/journal.pcbi.1002575
6. Currin A, Swainston N, Day PJ *et al.* Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence

space intelligently. *Chem Soc Rev* 2015;44:1172–239. https://doi.org/10.1039/C4CS00351A
7. Chang A, Jeske L, Ulbrich S *et al.* BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res* 2021;49:D498–508. https://doi.org/10.1093/nar/gkaa1025
8. Wittig U, Rey M, Weidemann A *et al.* SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. *Nucleic Acids Res* 2018;46:D656–60. https://doi.org/10.1093/nar/gkx1065
9. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* 2023;51:D523–31. https://doi.org/10.1093/nar/gkac1052
10. Nilsson A, Nielsen J, Palsson BO. Metabolic models of protein allocation call for the kinetome. *Cell Syst* 2017;5:538–41. https://doi.org/10.1016/j.cels.2017.11.013
11. Li F, Yuan L, Lu H *et al.* Deep learning-based k_{cat} prediction enables improved enzyme-constrained model reconstruction. *Nat Catal* 2022;5:662–72. https://doi.org/10.1038/s41929-022-00798-z
12. Kroll A, Rousset Y, Hu XP *et al.* Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. *Nat Commun* 2023;14:4139. https://doi.org/10.1038/s41467-023-39840-4
13. Qiu S, Zhao S, Yang A. DLTKcat: deep learning-based prediction of temperature-dependent enzyme turnover rates. *Brief Bioinform* 2023;25:bbad506. https://doi.org/10.1093/bib/bbad506
14. Wang T, Xiang G, He S *et al.* DeepEnzyme: a robust deep learning model for improved enzyme turnover number prediction by utilizing features of protein 3D-structures. *Brief Bioinform* 2024;25:bbae409. https://doi.org/10.1093/bib/bbae409
15. Kroll A, Engqvist MKM, Heckmann D *et al.* Deep learning allows genome-scale prediction of Michaelis constants from structural features. *PLoS Biol* 2021;19:e3001402. https://doi.org/10.1371/journal.pbio.3001402
16. Yu H, Deng H, He J *et al.* UniKP: a unified framework for the prediction of enzyme kinetic parameters. *Nat Commun* 2023;14:8211. https://doi.org/10.1038/s41467-023-44113-1
17. Shen X, Cui Z, Long J *et al.* EITLEM-Kinetics: a deep-learning framework for kinetic parameter prediction of mutant enzymes. *Chem Catalysis* 2024;4:101094. https://doi.org/10.1016/j.checat.2024.101094
18. Wang Z, Xie D, Wu D *et al.* Robust enzyme discovery and engineering with deep learning using CataPro. *Nat Commun* 2025;16:2736. https://doi.org/10.1038/s41467-025-58038-4
19. Gado JE, Beckham GT, Payne CM. Improving enzyme optimum temperature prediction with resampling strategies and ensemble learning. *J Chem Inf Model* 2020;60:4098–107. https://doi.org/10.1021/acs.jcim.0c00489
20. Qiu S, Hu B, Zhao J *et al.* Seq2Topt: a sequence-based deep learning predictor of enzyme optimal temperature. *Brief Bioinform* 2025;26:bbaf114. https://doi.org/10.1093/bib/bbaf114
21. Li F, Chen Y, Anton M *et al.* GotEnzymes: an extensive database of enzyme parameter predictions. *Nucleic Acids Res* 2023;51:D583–6. https://doi.org/10.1093/nar/gkac831
22. Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;379:1123–30. https://doi.org/10.1126/science.ade2574
23. Li M, Wang H, Yang Z *et al.* DeepTM: a deep learning algorithm for prediction of melting temperature of thermophilic proteins directly from sequences. *Comput Struct Biotechnol J* 2023;21:5544–60. https://doi.org/10.1016/j.csbj.2023.11.006
24. Jarzab A, Kurzawa N, Hopf T *et al.* Meltome atlas—thermal proteome stability across the tree of life. *Nat Methods* 2020;17:495–503. https://doi.org/10.1038/s41592-020-0801-4
25. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;35:1026–8. https://doi.org/10.1038/nbt.3988
26. Rives A, Meier J, Sercu T *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA* 2021;118:e2016239118. https://doi.org/10.1073/pnas.2016239118
27. Xia J, Zhao C, Hu B *et al.* Mole-BERT: rethinking pre-training graph neural networks for molecules. In: *The Eleventh International Conference on Learning Representations (ICLR 2023)*. Kigali, Rwanda, 2023.
28. Meier J, Rao R, Verkuil R *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *Adv Neural Inf Process Syst* 2021;34:29287–303.
29. Ahmad W, Simon E, Chithrananda S *et al.* Chemberta-2: towards chemical foundation models. arXiv, https://arxiv.org/abs/2209.01712, 5 September 2022, preprint: not peer reviewed.
30. Zhou G, Gao Z, Ding Q *et al.* Uni-mol: a universal 3d molecular representation learning framework. In: *The Eleventh International Conference on Learning Representations (ICLR 2023)*. Kigali, Rwanda, 2023.
31. Ji X, Wang Z, Gao Z *et al.* Exploring molecular pretraining model at scale. In: *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS 2024)*. Vancouver, Canada, 2024.
32. Elnaggar A, Heinzinger M, Dallago C *et al.* ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2022;44:7112–27. https://doi.org/10.1109/TPAMI.2021.3095381
33. Fang Y, Zhang N, Chen Z *et al.* Domain-agnostic molecular generation with chemical feedback. In: *The Twelfth International Conference on Learning Representations (ICLR 2024)*. Vienna, Austriar, 2024.
34. Lv L, Lin Z, Li H *et al.* ProLLaMA: a Protein Large Language Model for Multi-Task Protein Language Processing. *IEEE Trans Artif Intell* 2025;1:1–12. https://doi.org/10.1109/TAI.2025.3564914
35. Honda S, Shi S, Ueda HR. SMILES Transformer: pre-trained molecular fingerprint for low data drug discovery. arXiv, https://arxiv.org/abs/1911.04738, 12 November 2019, preprint: not peer reviewed.
36. Kanehisa M, Furumichi M, Sato Y *et al.* KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 2021;49:D545–51. https://doi.org/10.1093/nar/gkaa970
37. Helena-Bueno K, Brown CR, Melnikov S. Gosha: a database of organisms with defined optimal growth temperatures. bioRxiv, https://doi.org/10.1101/2021.12.21.473645, 30 May 2023, preprint: not peer reviewed.
38. Zhu J, Long J, Li X *et al.* Improving the thermal stability and branching efficiency of *Pyrococcus horikoshii* OT3 glycogen branching enzyme. *Int J Biol Macromol* 2024;255:128010. https://doi.org/10.1016/j.ijbiomac.2023.128010

Received: August 9, 2025. **Revised:** September 14, 2025. **Accepted:** September 19, 2025