*Briefings in Bioinformatics*, 2024, **25(1)**, 1–11
https://doi.org/10.1093/bib/bbad506
**Problem Solving Protocol**

# DLTKcat: deep learning-based prediction of temperature-dependent enzyme turnover rates

Sizhe Qiu, Simiao Zhao and Aidong Yang
Corresponding author: Aidong Yang, Department of Engineering Science, University of Oxford, OX1 3PJ, UK. E-mail: aidong.yang@eng.ox.ac.uk

## Abstract

The enzyme turnover rate, $k_{cat}$, quantifies enzyme kinetics by indicating the maximum efficiency of enzyme catalysis. Despite its importance, $k_{cat}$ values remain scarce in databases for most organisms, primarily because of the cost of experimental measurements. To predict $k_{cat}$ and account for its strong temperature dependence, DLTKcat was developed in this study and demonstrated superior performance (log10-scale root mean squared error = 0.88, $R$-squared = 0.66) than previously published models. Through two case studies, DLTKcat showed its ability to predict the effects of protein sequence mutations and temperature changes on $k_{cat}$ values. Although its quantitative accuracy is not high enough yet to model the responses of cellular metabolism to temperature changes, DLTKcat has the potential to eventually become a computational tool to describe the temperature dependence of biological systems.

***Keywords***: deep learning; compound–protein interaction; enzyme turnover rate; temperature dependence; genome-scale metabolic modeling

## INTRODUCTION

In the age of synthetic biology, more and more chemical processes are being catalyzed by enzymes [1, 2], and therefore, the quantitative study of enzyme kinetics becomes an important topic. The enzyme turnover rate, $k_{cat}$, is one of the most important parameters in describing enzyme kinetics, which quantifies the maximum efficiency of an enzyme in catalyzing a specific reaction [3]. In spite of its importance, there currently exists a huge gap of measured $k_{cat}$ for most organisms in commonly used enzyme databases [4], i.e. BRENDA [5] and SABIO-RK [6]. Also, measuring $k_{cat}$ values via enzyme assays is expensive and labor intensive [4], which means that it is hard to obtain $k_{cat}$ values in a high-throughput manner. The limited availability of $k_{cat}$ in databases and the indispensable requirement for $k_{cat}$ in the study of enzyme kinetics and other fields, such as metabolic modeling [7], fuel the impetus behind the development of computational methods to predict $k_{cat}$ values.

There are two main methods to predict $k_{cat}$ values: (1) estimating $k_{cat}$ based on apparent catalytic rate ($k_{app}$) with proteomic and fluxomic profiling and (2) predicting $k_{cat}$ using the compound–protein interaction (CPI) deep learning model. The first method obtains the $k_{cat}$ value by dividing the measured reaction flux by the quantified protein abundance [8, 9]. Although this method has been proved successful in resource allocation models of various microorganisms [10–13], fluxomics and proteomics are costly to measure, making this method difficult to implement.

CPI deep learning models have already been developed to predict biological parameters such as binding affinities ($K_d$) [14], Michaelis–Menten constants ($K_m$) [15] and enzyme turnover rates ($k_{cat}$) [16]. The inputs are usually simplified molecular-input line-entry system (SMILES) strings of compounds and subsequences of proteins. Compound and protein features are extracted by graph neural network, recurrent neural network or convolutional neural network (CNN), and then concatenated for the regression of the target value, such as $k_{cat}$ or $K_m$ [17]. For better performance, attention layers are added to capture the interaction between compound and protein features [18, 19]. DLKcat [16], the first CPI deep learning model for $k_{cat}$ prediction, can predict $\log_{10}(k_{cat})$ with the root mean squared error (RMSE) score below 1 and Pearson's $r$ = 0.71 for the test data set. However, one limitation of DLKcat and most other CPI models is that they do not account for experimental conditions like temperature, pH or ionic strength. As $k_{cat}$ has a strong dependence on temperature [20] and temperature is widely available in databases, developing a deep learning model that takes compound, protein and temperature features together as inputs are both necessary and approachable.

TurNuP [21], a CPI model for $k_{cat}$ with enhanced performance than DLKcat, included temperature as a feature in a case study to predict $k_{cat}$ for *Escherichia coli* (*E. coli*), but it was not a general predictive model for temperature-dependent $k_{cat}$. EF-UniKP and Revised UniKP [22] were developed to predict temperature-dependent $k_{cat}$ values. They considered $k_{cat}$ values at different

**Sizhe Qiu** is a PhD student at the Department of Engineering Science, University of Oxford. His research interests are in bacterial metabolism, bioinformatics and deep learning.
**Simiao Zhao** is a PhD student of Radcliffe department of Medicine, University of Oxford. His research interests include applying and developing machine learning tools to analyze biological data.
**Aidong Yang** is a Professor of Engineering Science at the Department of Engineering Science, University of Oxford. His research interests include mathematical modelling of biological systems such as metabolic networks and microbial communities.

**Received:** October 19, 2023. **Revised:** November 29, 2023. **Accepted:** December 8, 2023

temperatures and include the temperature value as a feature. However, the feature importance of temperature in those two models was not assessed, and no case studies were conducted to show the model's ability to predict the effect of temperature on $k_{cat}$ values. Also, the R-squared ($R^2$) scores of predictions by those two models were reported to be below 0.5.

With the aim to construct a deep learning model on $k_{cat}$ prediction that is more accurate than previously published models, this study developed DLTKcat. DLTKcat is a bidirectional attention CPI model with molecular graphs converted from SMILES strings, 3-mer subsequences of proteins and temperature features as inputs. It showed superior performance (log10-scale RMSE = 0.88, $R^2 = 0.66$) than previously published models (e.g. EF-UniKP), and demonstrated the feature importance of temperature. Then, DLTKcat exhibited its potential application in enzyme sequence design by predicting the effect of amino acid substitutions on $k_{cat}$ at different temperatures. Finally, we incorporated temperature-dependent proteome constraints in bacterial metabolic modeling with predicted $k_{cat}$ at different temperatures, to explore the possibility of using DLTKcat to make metabolic modeling sensitive to temperature changes.

# METHODS

## Data set preparation

The data set used to construct the deep learning model was extracted from the BRENDA and SABIO-RK databases. Enzyme class (EC) number, substrate name, organism name, protein identifier (UniProt ID), enzyme type, temperature and $k_{cat}$ values were queried from SABIO-RK via application programming interface (API). The data in BRENDA were fetched using BRENDApyrser [23]. The canonical SMILES string [24] of the substrate, which describes the molecular structure of chemical species, was obtained by querying the PubChem compound database [25] via API. The amino acid sequence of each enzyme protein was queried from the UniProt database [26] based on the UniProt ID also via API. The sequences of wild-type (WT) enzymes were mapped directly. For mutants caused by amino acid substitutions, amino acids at mutated locations were changed based on mutation information from BRENDA and SABIO-RK. Entries with other types of mutations were removed. All API codes can be found at https://github.com/SizheQiu/DLTKcat.

After SMILE strings and amino acid sequences were obtained, the data set filtered out all redundant entries with the same SMILE string, amino acid sequence, temperature and $k_{cat}$ value. For entries with the same SMILE string, amino acid sequence, temperature but different $k_{cat}$ values, only the entry with the largest $k_{cat}$ value was kept, as done in Li *et al.* [16]. Finally, 4383 entries from SABIO-RK and 11 866 entries from BRENDA remained. In all, 10 556 entries' enzymes were WTs and 5693 entries' enzymes were mutants (Figure S1). $k_{cat}$ values of 87 EC (numbers) were found to have significant correlations with temperature, which covered 2430 entries (Figure S2). Considering the uneven distribution of temperature values in the data set, oversampling was performed to append two times of entries at low ($T < 20°C$) and high ($T > 40°C$) temperature ranges by randomly duplicating existing entries at those temperature ranges (Figure S3). Because previously published CPI deep learning models have shown that additional features, such as enzyme molar mass or the octanol–water partition coefficient of substrate, could not improve model performance [15, 21], the finalized data set of this study only contained SMILES strings of substrates, amino acid sequences of enzyme proteins and temperature values.

## Construction of the deep learning model

Similar to other CPI deep learning models, DLTKcat uses Graph Attention Network (GAT) and CNN to extract features from the substrate molecular graph and enzyme protein sequence, respectively (Figure 1). The use of bidirectional attention, adopted from BACPI by Li *et al.* [27], and integration of temperature and inverse temperature values capture the temperature-dependent interactions between atoms of the compound and residues of the protein. Finally, the concatenated features of compound, protein and temperature are fed into several dense layers (fully connected layers) to predict the $\log_{10}(k_{cat})$ value.

### *Compound representation*

RDKit [28] converts the SMILES string into the molecular graph of the substrate with atoms as vertices, and chemical bonds as edges. The graph, along with the initial embeddings of its vertices, is fed into the graph attentional layer of GAT. A linear learnable transformation converts the embeddings ($v_i^{init} \in R^{H_c}$, $H_c$=80) into higher-level features of the compound ($v_i^{'} \in R^{H_c^{'}}$, $H_c^{'}$=50). The multi-head attention mechanism in GAT concatenates output features from three independent graph attentional layers to increase the stability of the self-attention learning process. Finally, a single-layer neural network transforms concatenated features into the compound space. The final output features are atom features ($v_i \in R^{H_c}$) (Figure 1). Extended Connectivity Fingerprints (ECFPs) [29] of length 1024, computed by RDKit, are also used to represent the compound. A multilayer neural network transforms ECFPs into the compound space ($f \in R^{H_c}$).

### *Protein representation*

To capture diverse protein residue patterns, the protein sequence is split into overlapping 3-mer subsequences. 3-mer subsequences are then translated to randomly initialized embeddings ($r_i^{init} \in R^{H_p}$, $H_p$=80). Through four convolutional layers with leaky ReLU [30] as the activation function, embeddings are transformed to higher-level features of the protein sequence that can capture the complex relationships of residues. The final output features are residue features ($r_i \in R^{H_p}$) (Figure 1).

### *Bidirectional attention and integration of temperature*

The bidirectional attention mechanism is used to represent the interactions between atoms of the compound and residues of the protein. Residue, atom features and fingerprints are transformed into vectors ($c_i \in R^d$, $p_i \in R^d$, $hf \in R^d$), and a soft alignment matrix ($A \in R^{N_v \times N_r}$) indicates the interaction strengths. $d$ is the unified latent dimension ($d$ = 40, 64). The weighted information is extracted from the soft alignment matrix, and attention weights are computed in both atom-to-residue ($\alpha_{a2r} \in R^{N_v}$) and residue-to-atom ($\alpha_{r2a} \in R^{N_r}$) directions. The outputs are compound ($h_c \in R^d$) and protein ($h_p \in R^d$) features (Figure 1). To improve learning stability and representation capacity, a multi-head attention model (number of heads = 3) is used to capture diverse aspects of CPI ($h_c^{final} \in R^d$, $h_p^{final} \in R^d$).

Inspired by the Arrhenius equation ($k_{cat} = Ae^{-\frac{E_a}{RT}}$) [20], temperature ($T$) and inverse temperature ($\frac{1}{T}$) are first normalized ($\frac{x-x_{min}}{x_{max}-x_{min}}$), and then concatenated with compound and protein features output by the bidirectional attention process. The inverse of temperature best represents the linear relationship between $\frac{1}{T}$ and $\log_{10}(k_{cat})$. The concatenated features ($h_c^{final} \| h_f \| h_p^{final} \| [T, \frac{1}{T}]$ $\|$, is the concatenation operation) are then fed into several dense layers (layer number = 3–6), with leaky ReLU as the activation function, for the regression of the $\log_{10}(k_{cat})$ value.

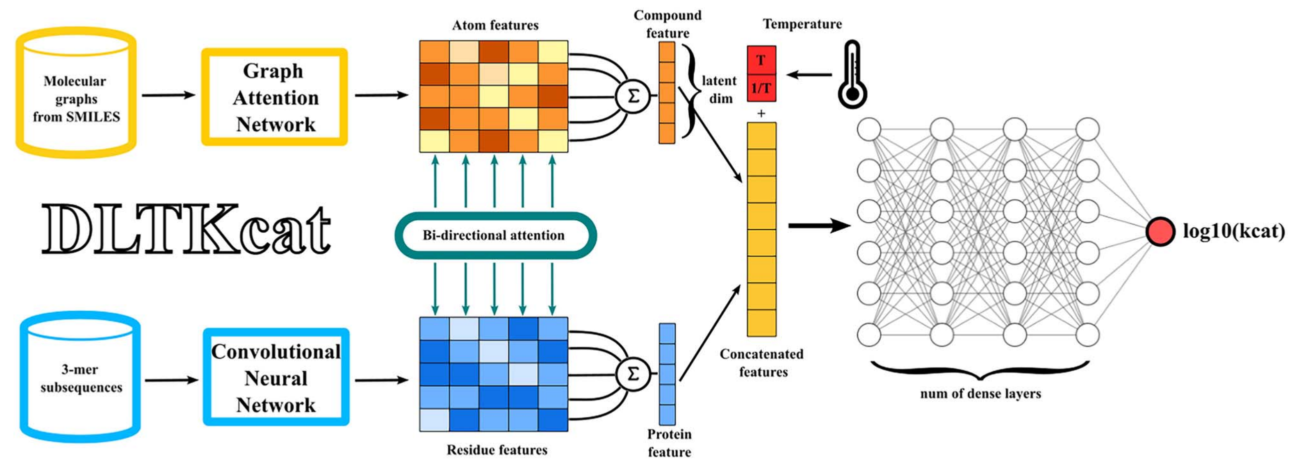


**Figure 1.** The overview of DLTKcat. With a pair of substrate and enzyme as the input, a GAT and a CNN learn the representations of the atom and residue from the compound molecular graph and protein sequence. Next, atom and residue representations are fed into the bidirectional attention neural network to integrate the representations and capture the important regions of compounds and proteins. Then, temperature ($T$) and inverse temperature ($\frac{1}{T}$) are integrated into the concatenated features. Finally, the concatenated features are used to predict the $\log_{10}(k_{cat})$ value.

### *Model training*

Because of the large size of the data set, batch training was used with a batch size of 32. Adam optimization algorithm [31] was used to update neural network weights iteratively. The loss function was mean squared error (MSE). The initial learning rate was 0.001, and the learning rate decayed by 50% for every 10 epochs to prevent overfitting. For details of software and hardware, please see Section S1.1 of the Supplementary Information.

## Interpretation of attention weights on protein residues

The bi-direction attention mechanism in section Bidirectional Attention and Integration of Temperature assigns attention weights to protein subsequences and atoms of the substrate. A higher attention weight of one residue means that residue is more important for the enzyme kinetics toward a certain substrate. The residue attention weights ($\alpha_{r2a}$) can be computed based on the intermediate output in the deep learning model.

$$c_i = LeakyReLU\,(W_v v_i) \tag{1}$$

$$p_i = LeakyReLU\,(W_r\, r_i) \tag{2}$$

$$I_p = A^T \tanh(CW_{a2r}) \tag{3}$$

$$\alpha_{r2a} = softmax\left(\left[PW_p \,\middle\|\, I_p\right] a_{r2a}\right) \tag{4}$$

$v_i$ and $r_i$ are atom and residue feature vectors (sections Compound Representation and Protein Representation), $W_v \in R^{d\times H_c}$ and $W_r \in R^{d\times H_p}$ transform $v_i$ and $r_i$ to $c_i$ and $p_i$, respectively (Equations (1 and 2)). $d$ is the latent dimension in the bidirectional attention mechanism. $C = [c_1, c_2, \ldots, c_{N_v}]$, $P = [p_1, p_2, \ldots, p_{N_r}]$, $W_{a2r} \in R^{d\times d}$, $A = \tanh(CUP^T) \in R^{N_v\times N_r}$ is a pairwise interaction matrix for atoms and residues, and $I_p \in R^{N_v\times d}$ represents information from atoms to residues (Equation (3)). $W_p \in R^{d\times d}$, $\|$ is the concatenation operation, $a_{r2a} \in R^{2d}$, and the vector of residue attention weights ($\alpha_{r2a}$) is protein attention weights normalized by the softmax function (Equation (4)).

## Proteome constrained flux balance analysis with predicted $k_{cat}$

Flux balance analysis has been used to estimate metabolic fluxes and cellular growth rates for decades [32]. The basic required inputs are the stoichiometric matrix (S) from the genome-scale metabolic model (GSMM) [32] and growth medium parameters that set upper bounds for nutrient uptake rates. Flux balance analysis computes metabolic fluxes ($v_i$) by maximizing an objective function (Equation (5)), which is usually the growth function [$v_{growth}$, biomass formation rate normalized to 1 gram dry weight (gDW) of biomass], via linear optimization in a constrained solution space of mass conservation (Equation (6)) and lower/upper bounds ($v_{lb}$, $v_{ub}$) of reaction fluxes (Equation (7)). Flux balance analysis was conducted using COBRApy [33] in this study.

$$\text{Max } v_{growth} \tag{5}$$

$$S * v = 0 \tag{6}$$

$$v_{lb} \le v_i \le v_{ub} \tag{7}$$

Proteome constrained flux balance analysis tightens the solution space by integrating proteome constraints of reactions into conventional flux balance analysis [34]. The reaction flux ($v_i$, $\frac{mmol}{hr*gDW}$) is constrained by the enzyme capacity ($k_i\,[E_i]$ or $a_i\,(MW_i * [E_i])$) (Equation (8)). $k_i$ is the $k_{cat}$ of reaction $i$ and $[E_i]$ is the enzyme molar concentration ($\frac{mmol}{gDW}$). $a_i$ ($\frac{\mu mol}{\min * mg\ E}$) is the enzyme-specific activity, defined as the micro moles of products formed by an enzyme in a given amount of time per milligram of the enzyme protein. $MW_i$ is enzyme molar mass ($\frac{g}{mol}$). Proteome was divided into sectors of inflexible housekeeping (Q), anabolism (A), transportation (T) and catabolism (C). The upper bound of all flexible sectors (i.e. C, A, T) combined was assumed to be 50% of the total proteome (Equation (9)) [35–37].

$$v_i \le k_i\,[E_i] \;\; or \;\; v_i \le a_i\,(MW_i * [E_i]) \tag{8}$$

$$\phi_Q\,(50\%) + \phi_C + \phi_A + \phi_T \le 100\% \tag{9}$$

$$\phi_A * P_{TOT} = MW_{ribosome} * [E_{ribosome}] = \frac{v_{growth}}{a_{ribosome}} \tag{10}$$

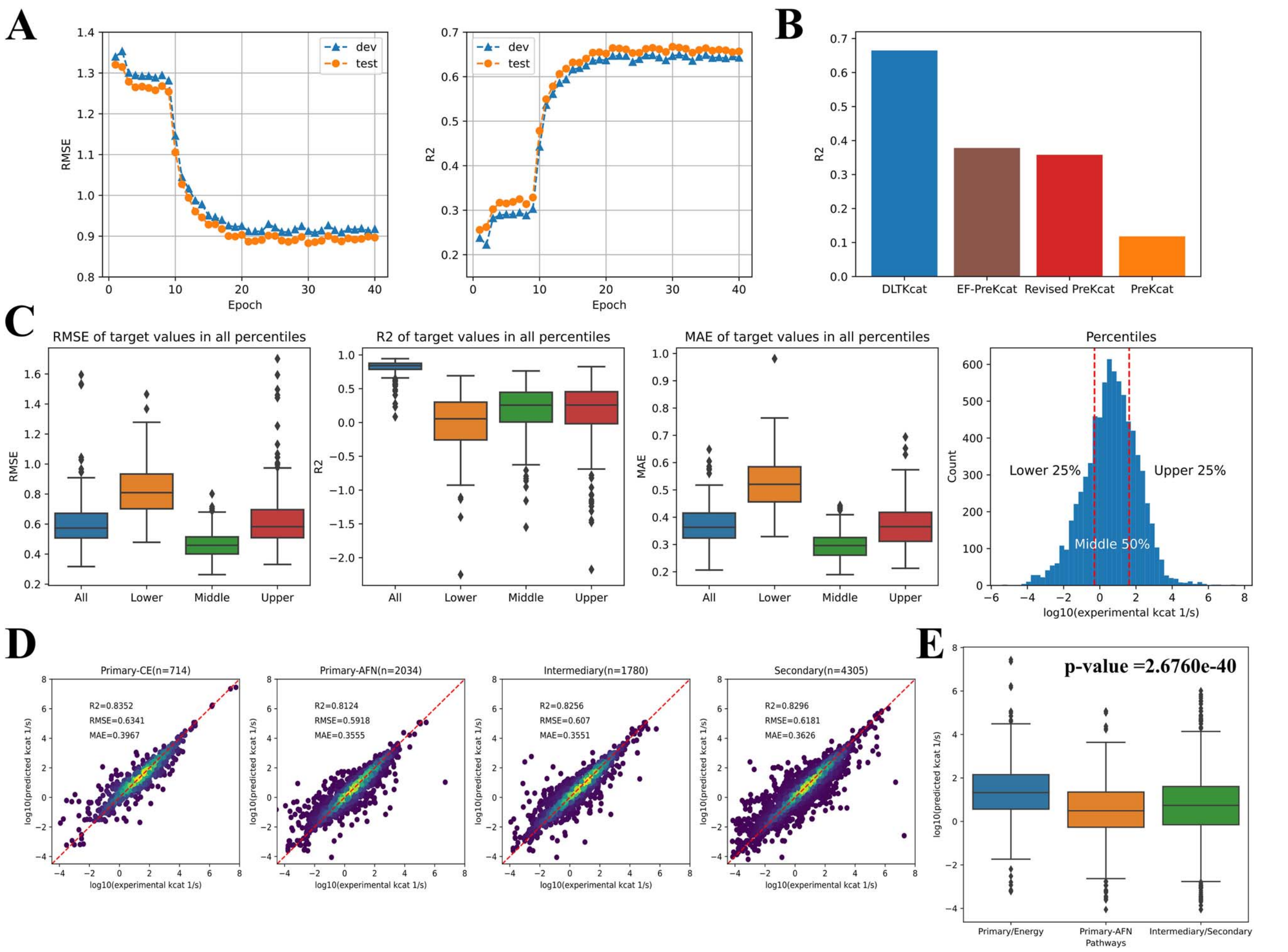


**Figure 2.** Assessment of the model performance. (**A**) The RMSE and $R^2$ scores of $\log_{10}(k_{cat})$ prediction during the training process. Test: the test set; dev: the validation set. The RMSE and $R^2$ of the test set at the end of training are 0.88 and 0.66. (**B**) Comparison of reported $R^2$ scores of DLTKcat, EF-UniKP, Revised UniKP and UniKP on $\log_{10}(k_{cat})$ prediction with temperature values. (**C**) The distributions of RMSE, $R^2$ and MAE scores of $\log_{10}(k_{cat})$ prediction for target values at lower 25%, middle 50% and upper 25% percentiles. (**D**) $R^2$, RMSE and MAE scores of $\log_{10}(k_{cat})$ prediction for enzymes in primary-CE, primary-AFN, intermediary and secondary metabolism. (**E**) The comparison of distributions of predicted $\log_{10}(k_{cat})$ values in primary-CE and other metabolic pathways ($P$-value < 0.001). Primary-AFN, primary metabolism—amino acid/fatty acid/nucleotide.

$$\phi_C * P_{TOT} = \sum MW_i * [E_i] = \sum \frac{v_i * MW_i}{k_i} \tag{11}$$

$\phi_x$ is the mass fraction of sector x for $x = A, C, T$. $P_{TOT}$ is the total mass of the proteome normalized to 1 gDW of biomass ($\frac{g}{gDW}$). The enzyme activity of the ribosome for the anabolism sector ($a_{ribosome}$) was set as 107.4 $\frac{mmol}{hr*g\ E}$ (Equation (10)) [36, 38]. $k_{cat}$ values were predicted for the catabolic sector (sector C), by DTLKcat (Equation (11)).

In this study, proteome constrained flux balance analysis was performed for *Lactococcus lactis MG1363* (LL) and *Streptococcus thermophilus LMG18311* (ST). The GSMMs used were obtained from the work of Flahaut *et al*. [39] and Pastink *et al*. [40]. Experimental data of LL and ST's growth rates at different temperatures were obtained from Chen *et al.* [41] and Vaningelgem *et al.* [42]. The carbon sources of LL and ST, in experiments, were glucose and lactose, respectively. Therefore, the enzyme activities ($a_{CT}$, CT stands for carbon source transportation) of glucose transport via phosphotransferase system and lactose: galactose antiporter were set as 361.14 $\frac{mmol}{hr*gE}$ [43] and 540 $\frac{mmol}{hr*gE}$ [44] (Equation (11)). Because both lactose and glucose were sufficient in the growth medium [41, 42], no Michaelis–Menten kinetics was needed for transporter proteins. Lactic and acetic acids were two major products of the central carbon metabolism of lactic acid bacteria, and the enzyme activity of acid exportation ($a_{AT}$) was set as 6360 $\frac{mmol}{hr*gE}$ [36, 38] (Equation (12)).

$$\phi_T * P_{TOT} = MW_{AT} * [E_{AT}] + MW_{CT} * [E_{CT}] = \frac{v_{AT}}{a_{AT}} + \frac{v_{CT}}{a_{CT}} \tag{12}$$

Temperature-dependent $k_{cat}$ values were predicted for enzymes in two bacteria's central carbon metabolism (Tables S1 and S2). The SMILES strings of substrates were queried from PubChem with metabolite names in GSMMs, and protein sequences were queried from UniProt with gene locus tags in genome assemblies of LL, GCF_000009425.1 [45], and ST, GCF_000011825.1 [46]. The predicted $k_{cat}$ for the primary substrate of each reaction was selected as the $k_{cat}$ of the reaction. For isozymes that catalyze the same metabolic reaction, the largest $k_{cat}$ was selected. Both ST and LL are important and widely used lactic acid bacteria, but their enzyme $k_{cat}$ values are quite limited in databases. For example, there are only 11 entries for ST in SABIO-RK, most were contributed by Simon and Hofer [47]. Therefore, this study used DLTKcat to fill the gap and examined DLTKcat's performance in predicting metabolic responses to temperature changes.

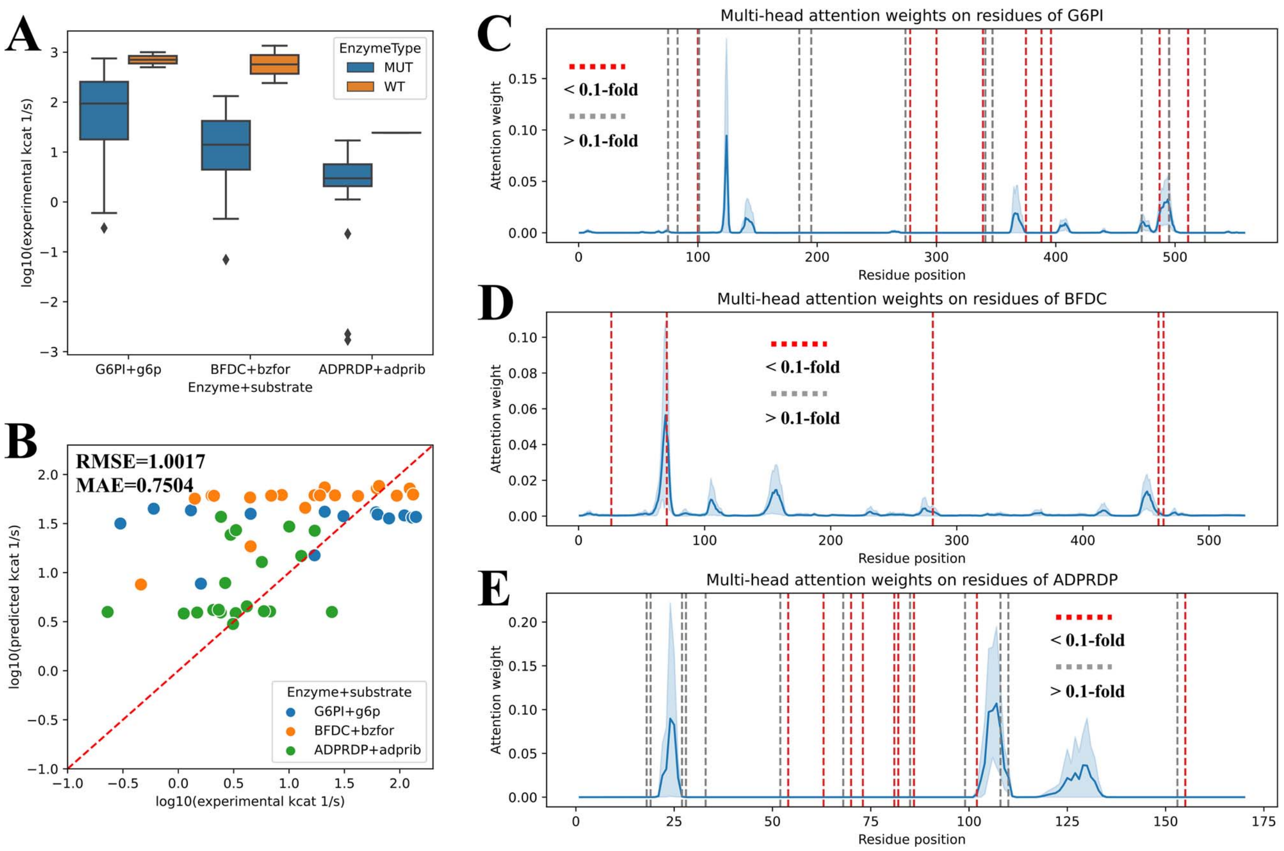


**Figure 3.** DLTKcat for the prediction and interpretation of $k_{cat}$ of mutated enzymes. (**A**) The comparison between experimental $\log_{10}(k_{cat})$ of WT and mutated enzymes for G6PI + g6p, BFDC + bzfor and ADPRDP + adprib. MUT, mutant. (**B**) RMSE and MAE scores of predicted $\log_{10}(k_{cat})$ values for G6PI + g6p, BFDC + bzfor and ADPRDP + adprib. RMSE = 1.0017, MAE = 0.7504. (**C**) Multi-head attention weights on residues of the WT G6PI and mutation sites. (**D**) Multi-head attention weights on residues of the WT BFDC and mutation sites. (**E**) Multi-head attention weights on residues of the WT ADPRDP and mutation sites. Dark dash line: mutation site (<0.1-fold WT $k_{cat}$); pale dash line: mutation site (>0.1-fold WT $k_{cat}$); solid curve: attention weight.

# RESULTS

## DLTKcat has good performance on temperature-dependent $k_{cat}$ prediction

With optimal hyperparameters (Section S1.3 and Figure S4), the model training process reduced RMSE (Equation (S2)) scores of predicted $\log_{10}(k_{cat})$ of the test data set from 1.33 to 0.88, and enhanced $R^2$ (Equation (S1)) scores from 0.25 to 0.66 after 20 epochs (Figure 2A). The $R^2$ scores of previously published deep learning models on temperature-dependent $k_{cat}$ were all reported to be below 0.5 [22], and DLTKcat has outperformed them by reaching a $R^2$ score of 0.66 on the randomly selected test data set (Figure 2B). In addition, DLTKcat showed good prediction accuracy with low RMSE and mean absolute error (MAE; Equation (S3)) scores for sub-data sets with experimental $\log_{10}(k_{cat})$ values at the lower 25%, middle 50% and upper 25% ranges (Figure 2C). In a nutshell, DLTKcat demonstrated superior performance in comparison to previously published deep learning models for temperature-dependent $k_{cat}$, and a robust accuracy for target values [experimental $\log_{10}(k_{cat})$ values] at different ranges.

To explore the predictive power of DLTKcat across different metabolic contexts, the prediction accuracy of $\log_{10}(k_{cat})$ values for enzymes in four different pathways, categorized based on enzyme modules in KEGG database [48], were assessed, and $R^2$, RMSE and MAE scores were all around 0.8, 0.6 and 0.35 (Figure 2D). After the validation of DLTKcat's good accuracy across different metabolic contexts, the model showed its ability to discriminate enzymes in primary metabolism—catabolism/energy (primary-CE) and other pathways, with higher predicted $\log_{10}(k_{cat})$ values in primary-CE (P-value < 0.001) (Figure 2E). In short, DLTKcat could well characterize enzymes from different metabolic contexts.

## Interpretation of $k_{cat}$ prediction of mutated enzymes

First, the accuracy of DLTKcat for both WT and mutated enzymes was examined, and $R^2$, RMSE and MAE scores were around 0.8, 0.6 and 0.4, respectively (Figure S5). After the prediction accuracy was ensured, this study selected three enzyme–substrate pairs with more than 20 mutations in the data set to investigate how DLTKcat captures amino acid substitutions. The three enzyme–substrate pairs were glucose-6-phosphate isomerase and D-glucose 6-phosphate (G6PI + g6p), benzoylformate decarboxylase and benzoylformate (BFDC + bzfor) and ADP-ribose diphosphatase and ADP-ribose (ADPRDP + adprib). The uniprot IDs of three enzyme proteins wereP06744, P20906 and Q5SKW5. Amino acid substitutions on protein sequences of three enzymes all resulted in the decrease of $k_{cat}$ (Figure 3A). The prediction accuracy of the selected three enzyme–substrate pairs was slightly lower than that of all mutated enzymes, but the prediction error was still around one order of magnitude (Figures 3B and S5). Next, the mapping of mutation sites to residue attention weights (section Interpretation of Attention Weights on Protein Residues) shows that most mutation sites (<0.1-fold WT $k_{cat}$) distribute closely to peaks of attention weights (Figure 3C–E). The overlapping between mutation sites (<0.1-fold WT $k_{cat}$) and residues with high attention weights was most noticeable for residue 70, 460 and 464 on BFDC (Figure 3D). Generally speaking, DLTKcat is a good predictor for mutated enzymes, and residue attention weights can reflect the impact of amino acid substitutions on enzyme kinetics.

## The contribution of temperature-related features to $k_{cat}$ prediction

Before feature importance analysis, the prediction accuracy was examined for different temperature ranges (below 20°C, above 40°C and between 20 and 40°C). High $R^2$ and low RMSE scores reflected that DLTKcat could accurately predict $k_{cat}$ for low, middle and high temperatures, with an error far below one order of magnitude (Figure S6). Then, feature shuffling, also known as feature permutation, was performed to show the importance of temperature and inverse temperature values (Section S1.4). The shuffling of temperature features resulted in significantly higher distributions of the prediction error (RMSE and MAE), and lower distributions of $R^2$ than those of predictions with unshuffled temperature features (Figure 4A). The comparison between predicted and experimental values showed that the RMSE and MAE scores increased by around 0.1 and $R^2$ decreased by around 0.1 when temperature-related features were shuffled (Figure 4B). For high ($T > 40$°C) and low ($T < 20$°C) temperature ranges, the increase of RMSE and MAE and decrease of $R^2$, caused by feature shuffling became larger (Figure 4C and D). In short, the decrease in prediction accuracy with shuffled temperature-related features demonstrated the importance of temperature-related features in DLTKcat.

## Use DLTKcat to predict $k_{cat}$ of WT and mutated *Pyrococcus furiosus* Ornithine Carbamoyltransferases

The $k_{cat}$ values of WT and mutated *Pyrococcus furiosus* ornithine carbamoyltransferases at 30 and 55°C were obtained from Roovers *et al.* [49]. The protein sequence of *P. furiosus* ornithine carbamoyltransferase was obtained from Uniprot with the Uniprot ID of Q51742. The prediction achieved high accuracy (RMSE = 0.5, MAE = 0.4338) (Figure 5A). Predicted $k_{cat}$ values at 55°C were higher than those at 30 C (Figure 5A–C), which was both consistent with the experimental data and the nature of *P. furiosus* being a hyperthermophile favoring high temperature [50].

With respect to the effect of mutations, DLTKcat suggested that amino acid substitutions at 227th, 240th and 277th amino acids could increase the $k_{cat}$ value, consistent with the experimental data, despite that the numerical difference between predicted $k_{cat}$ values of mutants and WT was small (Figure 5B and C; note the difference in scale between upper and lower y-axes). Furthermore, DLTKcat also captured that the combination of two amino acid substitutions, Y227C/E277G and A240D/E277G, could result in greater improvement on the $k_{cat}$ value than the substitution at each single site, though it failed to predict that the $k_{cat}$ of A240D/E277G was higher than that of Y227C/E277G (Figure 5B and C). The mapping of mutation sites to residue attention weights showed that E277G, as the mutation with a higher enhancement of $k_{cat}$ than other two mutations, was also closer to the high peak of attention weights (Figure 5D). In addition, residue attention weights indicated other potential mutation sites on *P. furiosus* ornithine carbamoyltransferase that might have substantial effects on $k_{cat}$ (Figure 5D).

## Temperature sensitive metabolic modeling with predicted $k_{cat}$

DLTKcat predicted $k_{cat}$ values for enzymes of LL at 30, 32, 34, 36 and 38°C, and of ST at 25, 32, 37, 42, 46 and 49°C, which were temperatures where LL and ST's growth rates were measured in experimental data [41, 42]. DLTKcat predicted that $k_{cat}$ of most catabolic enzymes in LL would decrease when temperature increased from 30 to 38°C, especially for G6PI (PGI), phosphofructokinase (PFK), phosphoglycerate kinase (PGK), pyruvate kinase (PYK), pyruvate formate lyase (PFL) and phosphotransacetylase (PTAr) (Figure 6A). The predicted decrease of the activity of catabolism in LL in response to temperature increase is consistent with the experimental observation that LL stopped growing after temperature became larger than 38°C [41]. For catabolic enzymes in ST, DLTKcat predicted that most enzymes' $k_{cat}$ would increase when temperature increased from 25 to 42°C, especially for fructose-bisphosphate aldolase (FBA, not the abbreviation of flux balance analysis), Glyceraldehyde-3-phosphate dehydrogenase (GAPD), phosphoglycerate mutase (PGM), enolase (ENO) and pyruvate kinase (PYK) (Figure 6B). The predicted increase of catabolic activity in ST when temperature increases to 42°C is consistent with both the experimental data [42] and the nature of ST being a thermophile [51]. These results showed that, in general, DLTKcat could qualitatively predict metabolic responses of bacteria to certain temperature changes.

However, the quantitative accuracy of growth rates computed by proteome constrained flux balance analysis was low. In proteome constrained flux balance analysis for LL, the $k_{cat}$ of fructose-bisphosphate aldolase (FBA) in LL was fixed at 13.9 $\frac{1}{s}$ [52] in sacrifice of temperature sensitivity, because predicted $k_{cat}$ values at different temperatures were unrealistically low (0.04 ∼ 0.065 $\frac{1}{s}$), compared with experiment $k_{cat}$ values in other bacteria [52, 53]. The predicted growth rates of LL by proteome constrained flux balance analysis captured the decreasing trend in response to the increase of temperature, but the predicted values were deviant from experimental values (Figure 6C). The proteome constrained flux balance analysis predicted the increase of ST's growth rate from 25 to 42°C, but it failed to predict the drop of growth rate from 42 to 49°C (Figure 6D). Also, the predicted increase of $k_{cat}$ values from 42 to 49°C by DLTKcat (Figure S7) contradicted the experimental finding that 49°C is close to the theoretical maximum temperature for ST to survive, 47–50°C [51]. To conclude, the log10-scale RMSE score within 1 of DLTKcat is not low enough to enable temperature sensitive proteome constrained flux balance analysis to predict bacterial growth and metabolism with good quantitative accuracy.

# DISCUSSION

The expensive cost of obtaining enzyme $k_{cat}$ values in wet lab stimulates the need of developing computational models to predict $k_{cat}$. Nevertheless, predicting temperature-dependent $k_{cat}$ is a challenging task, as temperature is not only a variable in the exponential factor of the Arrhenius equation, it also affects the activation energy of the enzyme catalyzed reaction, which is governed by the CPI [20]. To tackle the challenging task, this study constructed a CPI deep learning model called DLTKcat. DLTKcat used the bidirectional attention mechanism [27] to represent the interactions between compounds and proteins, and attention weights could capture important regions on protein sequences (section Interpretation of $k_{cat}$ Prediction of Mutated Enzymes). The use of both temperature and inverse temperature values facilitated the learning process of the neural network by representing features in the most biophysical relevant form to $k_{cat}$ [20]. Also, oversampling on entries at low and high temperature ranges compensated for the imbalanced distribution of temperature values in the data set (Figure S3). As a result,

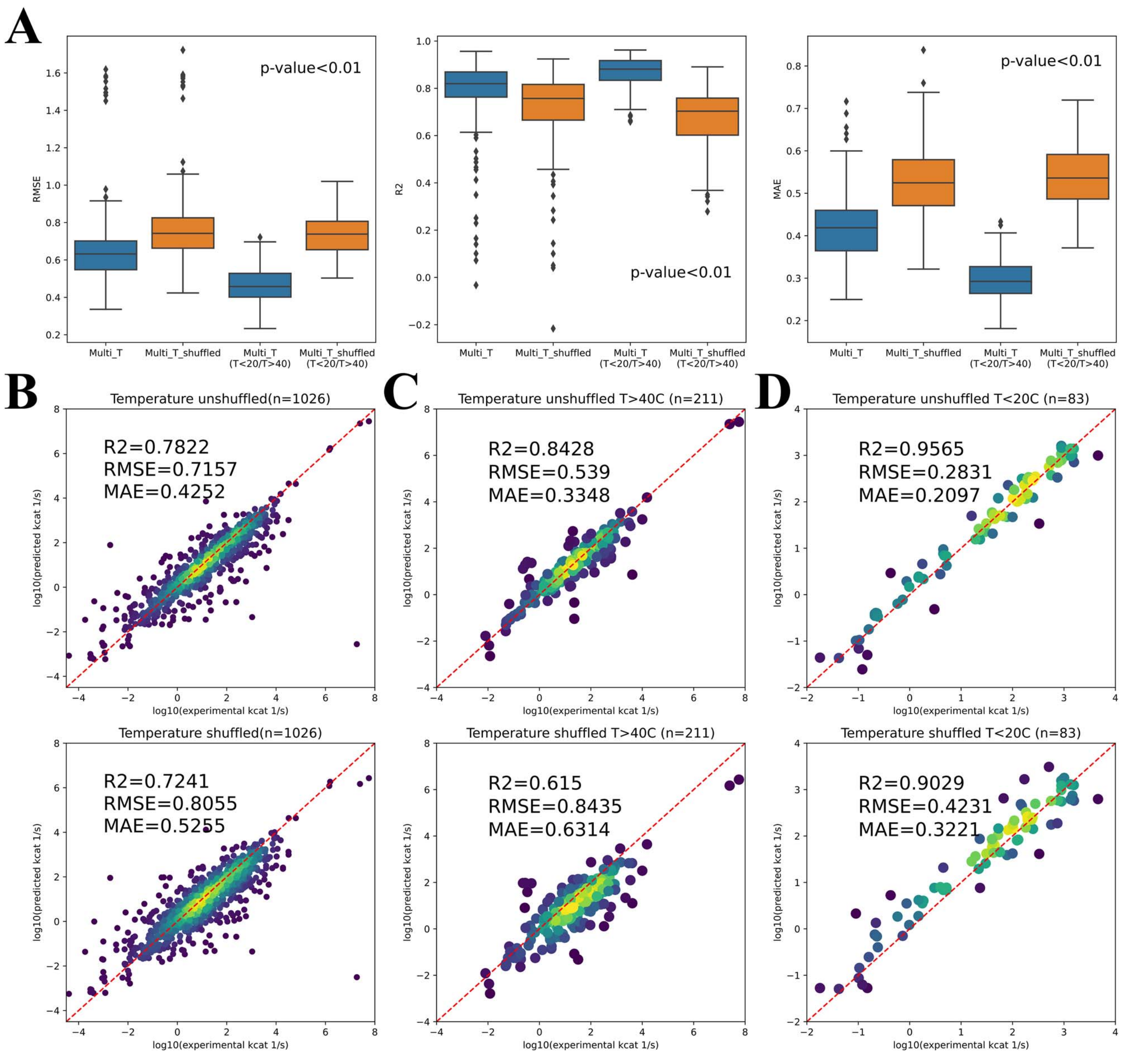


**Figure 4.** The importance of temperature-related features in DLTKcat. (**A**) The distributions of RMSE, $R^2$, MAE scores of predicted $\log_{10}(k_{cat})$ values with unshuffled and shuffled temperature-related features for the selected data set with 1026 entries (Multi_T) and for entries of low ($T < 20°C$) and high ($T > 40°C$) temperature. (**B**) $R^2$, RMSE and MAE scores of predicted $\log_{10}(k_{cat})$ values with unshuffled and shuffled temperature-related features for the selected data set with 1026 entries. (**C**) $R^2$, RMSE and MAE scores of predicted $\log_{10}(k_{cat})$ values with unshuffled and shuffled temperature-related features for entries of high temperature. (**D**) $R^2$, RMSE and MAE scores of predicted $\log_{10}(k_{cat})$ values with unshuffled and shuffled temperature-related features for entries of low temperature.

DLTKcat showed superior performance (log10-scale RMSE = 0.88, $R^2$ = 0.66) than previously published models (e.g. EF-UniKP) and robust accuracy for $k_{cat}$ predictions for different conditions (e.g. metabolic contexts). In addition, feature shuffling demonstrated the contribution of temperature-related features to this deep learning model.

By accurately predicting the effect of protein sequence mutations on the $k_{cat}$ value of *P. furiosus* ornithine carbamoyltransferase at different temperatures (section Use DLTKcat to Predict $k_{cat}$ of WT and Mutated *P. furiosus* Ornithine Carbamoyltransferases), DLTKcat exhibited its function in scoring the efficiency of in silico designed enzyme protein sequences. Imaginably, the combination of DLTKcat and optimization algorithms (e.g. genetic programming) can become a computational tool to design site-specific mutagenesis to optimize enzyme catalysis, which will be more efficient than directed evolution that relies on random mutagenesis.

Nonetheless, the second case study (section Temperature Sensitive Metabolic Modeling with Predicted $k_{cat}$) of generating temperature-dependent proteome constraints for metabolic modeling revealed the limitation of DLTKcat that its prediction error was not low enough to accurately model the response of cellular metabolism to temperature changes. Because all $k_{cat}$ values of catabolic enzymes in ST and LL were predicted by DLTKcat, the propagation of error led to the inaccuracy of proteome constrained flux balance analysis. In short, deep learning can gap fill a few missing $k_{cat}$ values in the metabolic network, as done in Li *et al*. [16], but the accuracy of proteome constrained flux balance analysis will not be high if most proteome constraints are based on predicted $k_{cat}$ values.

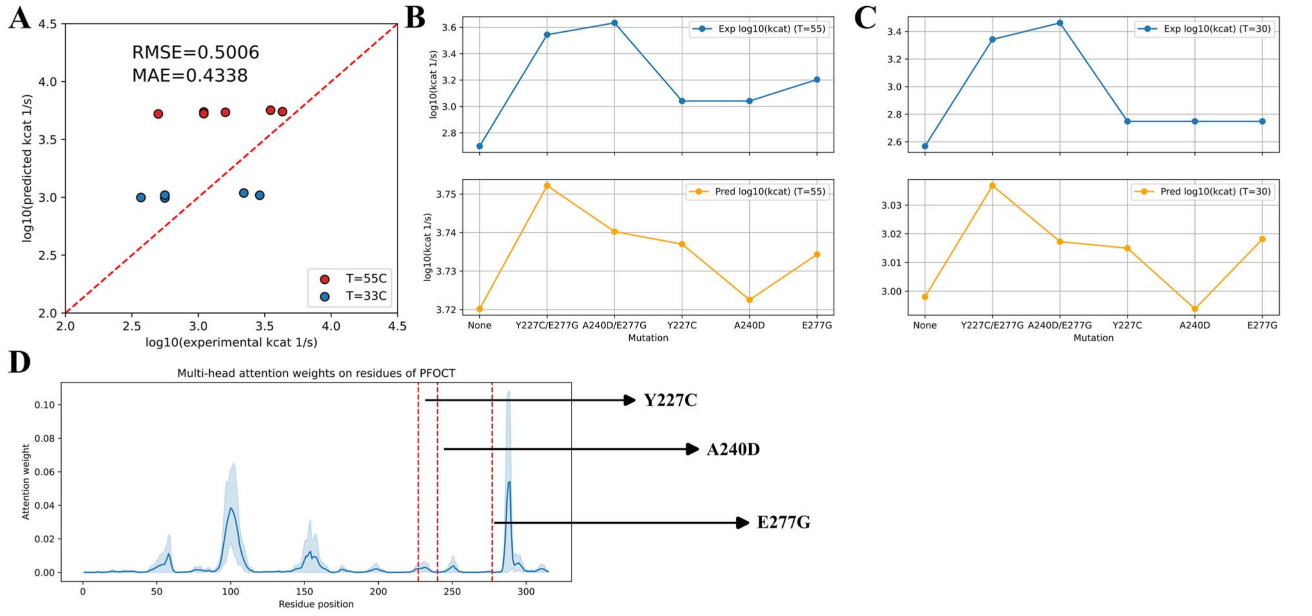

**Figure 5.** Prediction of the effect of amino acid substitutions on $k_{cat}$ values. (**A**) Comparison between experimental and predicted $\log_{10}(k_{cat})$ of *P. furiosus* ornithine carbamoyltransferase, RMSE = 0.5006, MAE = 0.4338. (**B**) Experimental (Exp) and predicted (Pred) $\log_{10}(k_{cat})$ values of WT and mutants at 55°C. (**C**) Experimental (Exp) and predicted (Pred) $\log_{10}(k_{cat})$ values of WT and mutants at 30°C. Exp, experimental value; Pred, predicted value. (**D**) Multi-head attention weights on residues of the WT *P. furiosus* ornithine carbamoyltransferase protein sequence. Dash-line: mutation site.

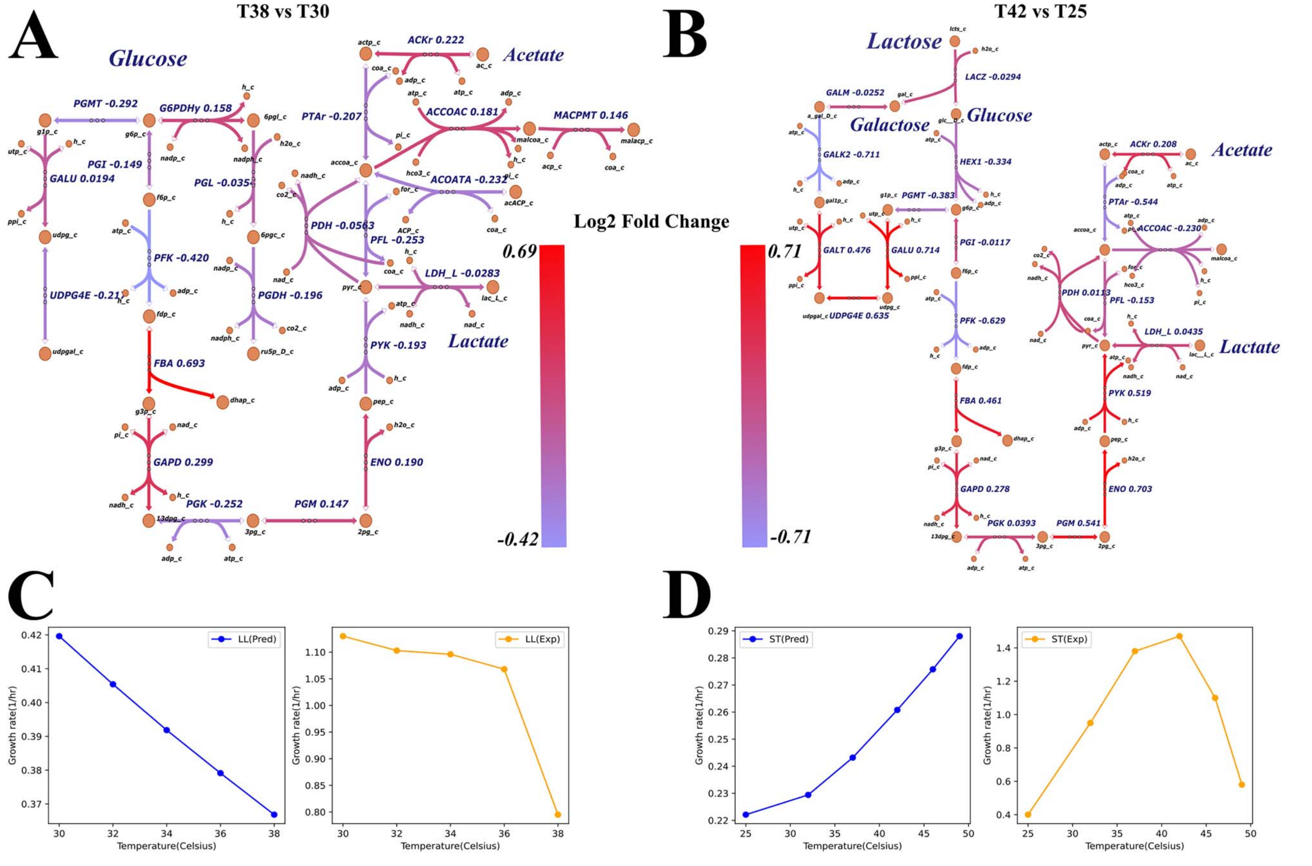

**Figure 6.** Prediction of bacteria metabolism at different temperatures. (**A**) Log2-fold change of predicted $k_{cat}$ values for LL at 38 and 30°C (38 versus 30°C). (**B**) Log2-fold change of predicted $k_{cat}$ values for ST at 42 and 25°C (42 versus 25°C). (**C**) Comparison of predicted (Pred) and experimental (Exp) growth rates of LL at 30, 32, 34, 36 and 38 °C. (**D**) Comparison of predicted (Pred) and experimental (Exp) growth rates of ST at 25, 32, 37, 42, 46 and 49°C. Exp, experimental value; Pred, predicted value. Reaction information can be found in Tables S1 and S2.

To further improve the performance and utility of DLTKcat, including additional experimental conditions like pH, metal ion concentrations might be an approach, but the lack of data restricted existing models from accounting for those factors [22]. Including the optimal enzyme temperature either from databases or predictions [54] might be able to enhance the temperature sensitivity of DLTKcat. The difference between the experimental temperature and optimal temperature could inform the model whether the temperature feature has a negative or positive effect on the $k_{cat}$ value. However, the success of this approach depends on the accuracy of enzyme optimal temperature prediction, which was reported to have a RMSE around 2 [54, 55].

Overall, DLTKcat can provide accurate predictions of $k_{cat}$ and account for the effect of temperature changes. Two case studies (3.4 Use DLTKcat to predict $k_{cat}$ of wild-type and mutated Pyrococcus furiosus Ornithine Carbamoyltransferases and 3.5 Temperature sensitive metabolic modeling with predicted $k_{cat}$) have revealed potential applications of DLTKcat on protein engineering, bacterial phenotype prediction, etc. Additionally, DLTKcat can be easily modified to predict other temperature-dependent CPIs, such as $K_m$ [56, 57]. With future improvements of the model framework, DLTKcat, as we envisage, will become a computational tool to quantitatively model the temperature dependence of biological systems, and contribute to the development of bioprocess digital twins.

**Key Points**

- This study constructed a deep learning model, DLTKcat, to predict temperature-dependent enzyme $k_{cat}$ with superior accuracy.
- The feature importance of temperature in predicting enzyme $k_{cat}$ was validated.
- DLTKcat can predict enzyme $k_{cat}$ for WT and mutated enzymes under different temperatures with good accuracy.
- With predicted enzyme $k_{cat}$ under different temperatures, proteome constrained flux balance analysis has the potential to model temperature-dependent cellular metabolism.

## SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

## ACKNOWLEDGEMENTS

The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility (http://dx.doi.org/10.5281/zenodo.22558) in carrying out this work. The authors would like to thank Yichi Zhang for offering technical guidance.

## FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## AUTHOR CONTRIBUTIONS

Sizhe Qiu constructed the deep learning model and performed case studies. Simiao Zhao assisted in the construction of the deep learning model. Aidong Yang supervised this research project and critically reviewed the manuscript.

## DATA AVAILABILITY

The code and data are openly available at https://github.com/SizheQiu/DLTKcat.

## ABBREVIATION

ADPRDP: ADP-ribose diphosphatase
API: application programming interface.
BFDC: benzoylformate decarboxylase.
CNN: convolutional neural network.
CPI: compound–protein interaction.
ECFP: Extended Connectivity Fingerprint.
GAT: graph attention network.
G6PI: glucose-6-phosphate isomerase.
gDW: gram dry weight.
GNN: graph neural network.
LL: *Lactococcus lactis MG1363*.
Leaky ReLU: leaky Rectified Linear Unit.
MAE: mean absolute error.
MSE: mean squared error.
MUT: mutant.
$R^2$: R-squared, the coefficient of determination.
RMSE: root mean squared error.
RNN: recurrent neural network.
SMILES: simplified molecular-input line-entry system.
ST: *Streptococcus thermophilus LMG18311*.
WT: wild type.

## REFERENCES


1. Stephanopoulos G. Synthetic biology and metabolic engineering. *ACS Synth Biol* 2012;**1**:514–25.
2. Madhavan A, Arun KB, Binod P, *et al.* Design of novel enzyme biocatalysts for industrial bioprocess: harnessing the power of protein engineering, high throughput screening and synthetic biology. *Bioresour Technol* 2021;**325**:124617.
3. Davidi D, Noor E, Liebermeister W, *et al.* Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro $k_{cat}$ measurements. *Proc Natl Acad Sci U S A* 2016;**113**:3401–6.
4. Nilsson A, Nielsen J, Palsson BO. Metabolic models of protein allocation call for the Kinetome. *Cell Syst* 2017;**5**:538–41.
5. Schomburg I, Jeske L, Ulbrich M, *et al.* The BRENDA enzyme information system–from a database to an expert system. *J Biotechnol* 2017;**261**:194–206.
6. 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.
7. 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.
8. Heckmann D, Campeau A, Lloyd CJ, *et al.* Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. *Proc Natl Acad Sci U S A* 2020;**117**:23182–90.
9. Bulović A, Fischer S, Dinh M, *et al.* Automated generation of bacterial resource allocation models. *Metab Eng* 2019;**55**:12–22.
10. Goelzer A, Muntel J, Chubukov V, *et al.* Quantitative prediction of genome-wide resource allocation in bacteria. *Metab Eng* 2015;**32**:232–43.
11. Jahn M, Crang N, Janasch M, *et al.* Protein allocation and utilization in the versatile chemolithoautotroph *Cupriavidus necator*. *Elife* 2021;**10**:10.
12. Coppens L, Tschirhart T, Leary DH, *et al.* *Vibrio natriegens* genome-scale modeling reveals insights into halophilic adaptations and resource allocation. *Mol Syst Biol* 2023;**19**:e10523.
13. Wendering P, Arend M, Razaghi-Moghadam Z, Nikoloski Z. Data integration across conditions improves turnover

number estimates and metabolic predictions. *Nat Commun* 2023;**14**:1485.

14. Li S, Wan F, Shu H, *et al.* MONN: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Syst* 2020;**10**:308–22.e11.
15. Kroll A, Engqvist MKM, Heckmann D, Lercher MJ. Deep learning allows genome-scale prediction of Michaelis constants from structural features. *PLoS Biol* 2021;**19**:e3001402.
16. 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.
17. Lim S, Lu Y, Cho CY, *et al.* A review on compound-protein interaction prediction methods: data, format, representation and model. *Comput Struct Biotechnol J* 2021;**19**:1541–56.
18. Shin B, Park S, Kang K, *et al.* Self-attention based molecule representation for predicting drug-target interaction. PMLR, 2019;**106**:230–48.
19. Karimi M, Wu D, Wang Z, Shen Y. DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* 2019;**35**:3329–38.
20. Arroyo JI, Díez B, Kempes CP, *et al.* A general theory for temperature dependence in biology. *Proc Natl Acad Sci U S A* 2022;**119**:e2119872119.
21. Kroll A, Rousset Y, Hu X-P, *et al.* Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. *Nat Commun* 2023;**14**:4139.
22. Yu H, Deng H, He J, *et al.* UniKP: a unified framework for the prediction of enzyme kinetic parameters. *Nat Commun* 2023; **14**:8211.
23. Estévez SR. *BRENDApyrser: a Python package to parse and manipulate the BRENDA database*. Zenodo, 2022.
24. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;**28**:31–6.
25. Kim S, Chen J, Cheng T, *et al.* PubChem 2023 update. *Nucleic Acids Res* 2023;**51**:D1373–80.
26. UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res* 2023;**51**:D523–31.
27. Li M, Lu Z, Wu Y, Li YH. BACPI: a bi-directional attention neural network for compound–protein interaction and binding affinity prediction. *Bioinformatics* 2022;**38**:1995–2002.
28. Landrum G, Tosco P, Kelley B, *et al. rdkit/rdkit: 2023_09_3 (Q3 2023) Release (Release_2023_09_3)*. Zenodo, 2023.
29. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;**50**:742–54.
30. Maas AL, Hannun AY, Ng AY. *Rectifier Nonlinearities Improve Neural Network Acoustic Models*. ICML, 2013.
31. Kingma DP, Ba J. Adam: a method for stochastic optimization. *ICLR* 2015.
32. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nat Biotechnol* 2010;**28**:245–8.
33. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-based reconstruction and analysis for python. *BMC Syst Biol* 2013;**7**:74.
34. Mori M, Hwa T, Martin OC, *et al.* Constrained allocation flux balance analysis. *PLoS Comput Biol* 2016;**12**:e1004913.
35. Zeng H, Yang A. Bridging substrate intake kinetics and bacterial growth phenotypes with flux balance analysis incorporating proteome allocation. *Sci Rep* 2020;**10**:4283.
36. Regueira A, Rombouts JL, Aljoscha Wahl S, *et al.* Resource allocation explains lactic acid production in mixed-culture anaerobic fermentations. *Biotechnol Bioeng* 2021;**118**:745–58.
37. Qiu S, Zeng H, Yang Z, *et al.* Dynamic metagenome-scale metabolic modeling of a yogurt bacterial community. *Biotechnol Bioeng* 2023;**120**:2186–98.
38. Schumacher R. *Metabolic Trade-Offs Arising from Increased Free Energy Conservation in Saccharomyces cerevisiae*. Delft University of Technology, 2018.
39. Flahaut NAL, Wiersma A, van de Bunt B, *et al.* Genome-scale metabolic model for *Lactococcus lactis* MG1363 and its application to the analysis of flavor formation. *Appl Microbiol Biotechnol* 2013;**97**:8729–39.
40. Pastink MI, Teusink B, Hols P, *et al.* Genome-scale model of *Streptococcus thermophilus* LMG18311 for metabolic comparison of lactic acid bacteria. *Appl Environ Microbiol* 2009;**75**: 3627–33.
41. Chen J, Shen J, Ingvar Hellgren L, *et al.* Adaptation of *Lactococcus lactis* to high growth temperature leads to a dramatic increase in acidification rate. *Sci Rep* 2015;**5**:14199.
42. Vaningelgem F, Zamfir M, Adriany T, de Vuyst L. Fermentation conditions affecting the bacterial growth and exopolysaccharide production by *Streptococcus thermophilus* ST 111 in milk-based medium. *J Appl Microbiol* 2004;**97**: 1257–73.
43. Christiansen I, Hengstenberg W. Staphylococcal phosphoenolpyruvate-dependent phosphotransferase system–two highly similar glucose permeases in *Staphylococcus carnosus* with different glucoside specificity: protein engineering in vivo? *Microbiology* 1999;**145**(Pt 10):2881–9.
44. Geertsma ER, Duurkens RH, Poolman B. The activity of the lactose transporter from *Streptococcus thermophilus* is increased by phosphorylated IIA and the action of beta-galactosidase. *Biochemistry* 2005;**44**:15889–97.
45. Wegmann U, O'Connell-Motherway M, Zomer A, *et al.* Complete genome sequence of the prototype lactic acid bacterium *Lactococcus lactis* subsp. cremoris MG1363. *J Bacteriol* 2007;**189**: 3256–70.
46. Bolotin A, Quinquis B, Renault P, *et al.* Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*. *Nat Biotechnol* 2004;**22**:1554–8.
47. Simon WA, Hofer HW. Phosphofructokinases from Lactobacteriaceae. II. Purification and properties of phosphofructokinase from *Streptococcus thermophilus*. *Biochim Biophys Acta* 1981;**661**: 158–63.
48. Kanehisa M, Furumichi M, Tanabe M, *et al.* KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**:D353–61.
49. Roovers M, Sanchez R, Legrain C, Glansdorff N. Experimental evolution of enzyme temperature activity profile: selection in vivo and characterization of low-temperature-adapted mutants of *Pyrococcus furiosus* ornithine carbamoyltransferase. *J Bacteriol* 2001;**183**:1101–5.
50. Fiala G, Stetter KO. *Pyrococcus furiosus* sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. *Arch Microbiol* 1986;**145**:56–61.
51. Harnett J, Davey G, Patrick A, *et al.* Lactic acid bacteria | *Streptococcus thermophilus*. Encyclopedia of Dairy Sciences (Second Edition), 2011, 143–8.
52. Callens M, Kuntz DA, Opperdoes FR. Kinetic properties of fructose bisphosphate aldolase from *Trypanosoma brucei* compared to aldolase from rabbit muscle and *Staphylococcus aureus*. *Mol Biochem Parasitol* 1991;**47**:1–9.
53. Plater AR, Zgiby SM, Thomson GJ, *et al.* Conserved residues in the mechanism of the *E. coli* class II FBP-aldolase. *J Mol Biol* 1999;**285**: 843–55.

54. 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.
55. Li G, Rabe KS, Nielsen J, Engqvist MKM. Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. *ACS Synth Biol* 2019;**8**:1411–20.
56. Quinlan AV. The thermal sensitivity of Michaelis-Menten kinetics as a function of substrate concentration. *J Franklin Inst* 1980;**310**:325–42.
57. Maggi FM, Tang FH, Riley WJ. The thermodynamic links between substrate, enzyme, and microbial dynamics in Michaelis-Menten-Monod kinetics. *Int J Chem Kinet* 2018;**50**:343–56.