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

Authorship

Sizhe Qiu[1], Simiao Zhao[2], Aidong Yang[1]*

[1]Department of Engineering Science, University of Oxford, OX1 3PJ, United Kingdom

[2]Radcliffe Department of Medicine, University of Oxford, OX3 9DU, United Kingdom

*Corresponding author: aidong.yang@eng.ox.ac.uk (A. Yang)

Abbreviation

API: application programming interface

CNN: convolutional neural network

CPI: compound protein interaction

ECFP: Extended Connectivity Fingerprint

GAT: graph attention network

gDW: gram dry weight

GNN: graph neural network

LL: *Lactococcus lactis MG1363*

Leaky ReLU: leaky Rectified Linear Unit

MSE: mean squared error

R2: r-squared, the coefficient of determination

RMSE: root mean squared error

RNN: recurrent neural network

SAV: single amino acid variation

SMILES: simplified molecular-input line-entry system

ST: *Streptococcus thermophilus LMG18311*

WT: wild type

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 due to 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 RMSE = 0.88, R2 = 0.66) than previously published models. Through two case studies, DLTKcat showed its ability to predict the effect 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;

1. Introduction

In the age of synthetic biology, more and more chemical processes are being catalyzed by enzymes [(Stephanopoulos, 2012)][(Madhavan *et al*, 2021)], 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 [(Davidi *et al*, 2016)]. In spite of its importance, there currently exists a huge gap of measured k_{cat} for most organisms in commonly used enzyme databases [(Nilsson *et al*, 2017)], i.e. BRENDA [(Schomburg *et al*, 2017)] and SABIO-RK [(Wittig *et al*, 2018)]. Also, measuring k_{cat} values via enzyme assays is expensive and labor intensive [(Nilsson *et al*, 2017)], which means that it's 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 [(Adadi *et al*, 2012)], fuels 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; 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 [(Heckmann *et al*, 2020)][(Bulović *et al*, 2019)]. Although this method has been proved successful in resource allocation models of various microorganisms [(Goelzer *et al*, 2015)][(Jahn *et al*, 2021)][(Coppens *et al*, 2023)][(Wendering *et al*, 2023)], 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) [(Li *et al*, 2020)], michaelis-menten constants (K_m) [(Kroll *et al*, 2021)] and enzyme turnover rates (k_{cat}) [(Li *et al*, 2022a)]. The inputs are usually SMILES

(simplified molecular-input line-entry system) strings of compounds and subsequences of proteins. Compound and protein features are extracted by graph neural network (GNN), recurrent neural network (RNN) or convolutional neural network (CNN), and then concatenated for the regression of the target value, such as k_{cat} or K_m [(Lim *et al*, 2021)]. For better performance, attention layers are added to capture the interaction between compound and protein features [(Shin *et al*, 09--10 Aug 2019)][(Karimi *et al*, 2019)]. DLKcat [(Li *et al*, 2022a)], the first CPI deep learning model for k_{cat} prediction, can predict $log_{10}(k_{cat})$ with the RMSE (root mean squared error) score below 1 and Pearson's r = 0.71 for the test dataset. 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 [(Arroyo *et al*, 2022)] and temperature is widely available in databases, developing a deep learning model that takes compound, protein and temperature features together as inputs is both necessary and approachable.

In 2023, EF-PreKcat, Revised PreKcat [(Yu *et al*, 2023)] and TurNuP [(Kroll *et al*, 2023)] were developed to predict temperature dependent k_{cat} values. They all considered k_{cat} values at different temperatures and include the temperature value as a feature. But none of them assessed the feature importance of temperature or demonstrated any case studies to show the model's ability to predict the effect of temperature on k_{cat} values. Also, the R2 (r-squared) scores of predictions by those models were all 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 bi-directional 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, R2 = 0.66) than previously published models (e.g., TurNuP), 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.

2. Methods

2.1 Dataset preparation

The dataset used to construct the deep learning model was extracted from the BRENDA and SABIO-RK databases. 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 was fetched using BRENDApyrser [(Estévez, 2022)]. The canonical SMILES string [(Weininger, 1988)] of the substrate, that describes the molecular structure of chemical species, was obtained by querying the PubChem compound database [(Kim *et al*, 2023)] via API. The amino acid sequence of each enzyme protein was queried from the UniProt database [(UniProt Consortium, 2023)] 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 dataset 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., 2022 [(Li *et*

al, 2022a)]. Finally, 4383 entries from SABIO-RK and 11866 entries from BRENDA remained. 10556 entries' enzymes were WTs and 5693 entries' enzymes were mutants (**see SI, Figure S1**). k_{cat} values of 87 enzyme classes (EC numbers) were found to have significant correlations with temperature, which covered 2430 entries (**see SI, Figure S2**). Considering the uneven distribution of temperature values in the dataset, oversampling was performed to append two times of entries at low (T<20□) and high (T>40□) temperature ranges by randomly duplicating existing entries at those temperature ranges (**see SI, 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 [(Kroll *et al*, 2021)][(Kroll *et al*, 2023)], the finalized dataset of this study only contained SMILES strings of substrates, amino acid sequences of enzyme proteins and temperature values.

2.2 Construction of the deep learning model

Similar to other CPI deep learning models, DLTKcat uses Graph Attention Network (GAT) and Convolutional Neural Network (CNN) to extract features from the substrate molecular graph and enzyme protein sequence, respectively (**Figure 1**). The use of bi-directional attention, adopted from BACPI by Li et al., 2022 [(Li *et al*, 2022b)], 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.

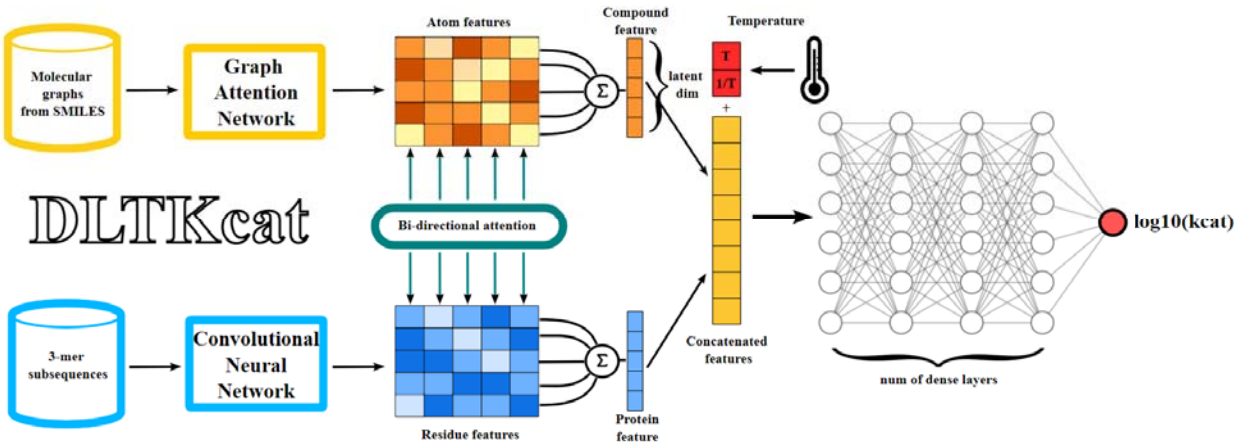

Figure 1. The overview of DLTKcat. With a pair of substrate and enzyme as the input, GAT and CNN learn the representations of the atom and residue from the compound molecular graph and protein sequence. Next, atom and residue representations are fed into the bi-directional attention neural network to integrate the representations and capture the important regions of compounds and proteins. Then, temperature ( ) and inverse temperature (–) are integrated into the concatenated features. Finally, the concatenated features are used to predict the value.

2.2.1 Compound representation

RDKit [(Landrum, 2006)] 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 into higher-level features of the compound. The multi-head attention mechanism in GAT concatenates output features from several 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 (**Figure 1**). Extended Connectivity Fingerprints (ECFPs)

[(Rogers & Hahn, 2010)] of length 1024, computed by RDKit, are also used to represent the compound. A multi-layer neural network transforms ECFPs into the compound space.

2.2.2 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. Through several convolutional layers with leaky ReLU [(Maas *et al*, 2013)] 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 (**Figure 1**).

2.2.3 Bi-directional attention and integration of temperature

The bi-directional 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, and a soft alignment matrix indicates the interaction strengths. The weighted information is extracted from the soft alignment matrix, and attention is computed in both atom-to-residue and residue-to-atom directions. The final outputs are compound and protein features (**Figure 1**). To improve learning stability and representation capacity, a multi-head attention model is used to capture diverse aspects of compound-protein interactions.

Inspired by the Arrhenius equation ($k_{cat} = Ae^{-\frac{E_a}{RT}}$) [(Arroyo *et al*, 2022)], 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 bi-directional attention process. The inverse of temperature best represents the linear relationship between $\frac{1}{T}$ and $log_{10}(k_{cat})$. The

concatenated features are then fed into several dense layers, with leaky ReLU as the activation function, for the regression of the $log_{10}(k_{cat})$ value.

2.2.4 Model training

Due to the large size of the dataset, batch training was used with a batch size of 32. Adam optimization algorithm [(Kingma & Ba, 2014)] 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 the section 1 of supplementary information.

2.3 Deep learning model evaluation

The original dataset curated in section 2.1 was randomly split into the test and the train dataset with a ratio of 1:9. The test dataset was held out to examine the accuracy of the model. Before the training, 10% of the train dataset was randomly split as the validation dataset (also called dev set). During the training process, R2 (Eq. 1) and RMSE (Eq. 2) scores of k_{cat} predictions were computed at each epoch for the test and validation datasets.

$$R^2 = \frac{\sum_{i=1}^{n}(y_{ie} - y_{ip})^2}{\sum_{i=1}^{n}(y_{ie} - \bar{y})^2} \quad (Eq.\,1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ie} - y_{ip})^2} \quad (Eq.\,2)$$

To find the optimal combinations of the latent dimension, the dimension of protein and compound features in bi-directional attention process, and the number of dense layers after feature concatenation (**Figure 1**), training processes were performed for latent dimension = 40, 64 and number of dense layers = 3, 4, 5, 6. The latent dimension and number of dense layers that achieved the lowest RMSE score on the dev set were used to generate the trained deep

learning model (**see SI, Figure S4**). The other hyperparameters were the same as default values used in BACPI [(Li *et al*, 2022b)].

1026 entries in the curated dataset had the same protein sequence and substrate as other entries but different temperatures. To assess the feature importance of temperature, they were extracted to obtain R2 and RMSE scores of k_{cat} predictions with shuffled and unshuffled temperature related features. The feature shuffling was performed by randomly sampling from the original list of feature values, so that the shuffled and unshuffled features still had the same distribution.

2.4 Proteome constrained flux balance analysis with predicted k_{cat}

Flux balance analysis (FBA) has been used to estimate metabolic fluxes and cellular growth rates for decades [(Orth *et al*, 2010)]. The basic required inputs are the stoichiometric matrix (S) from the genome-scale metabolic model (GSMM) [(Orth *et al*, 2010)] and growth medium parameters that set upper bounds for nutrient uptake rates. FBA computes metabolic fluxes (v_i) by maximizing an objective function (Eq. 3), 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 (Eq. 4) and lower/upper bounds (v_{lb}, v_{ub}) of reaction fluxes (Eq. 5). FBA was conducted using COBRApy [(Ebrahim *et al*, 2013)] in this study.

$$Max\ v_{growth}\ (Eq.3)$$

$$S * v\ =\ 0\ (Eq.4)$$

$$v_{lb} \leq v_i \leq v_{ub}\ (Eq.5)$$

Proteome constrained FBA tightens the solution space by integrating proteome constraints of reactions into conventional FBA [(Mori *et al*, 2016)]. The reaction flux (v_i, $\frac{mmol}{hr * gDW}$) is constrained by the enzyme capacity ($k_i[E_i]$ or $a_i(MW_i * [E_i])$) (Eq. 6). 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 (Eq. 7) [(Zeng & Yang, 2020)][(Regueira *et al*, 2021)][(Qiu *et al*, 2023)].

$$v_i \leq k_i[E_i]\ \ or\ v_i \leq a_i(MW_i * [E_i])\ (Eq.6)$$

$$\phi_Q(50\%) + \phi_C + \phi_A + \phi_T \leq 100\%\ (Eq.7)$$

$$\phi_A * P_{TOT} = MW_{ribosome} * [E_{ribosome}] = \frac{v_{growth}}{a_{ribosome}}\ (Eq.8)$$

$$\phi_C * P_{TOT} = \sum MW_i * [E_i] = \sum \frac{v_i * MW_i}{k_i}\ (Eq.9)$$

ϕ_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 $107.4\ \frac{mmol}{hr * g\ E}$ (Eq. 8) [(Schumacher, 2018)][(Regueira *et al*, 2021)]. k_{cat} values were predicted for the catabolic sector (sector C), by DTLKcat (Eq. 9).

In this study, proteome constrained FBA was performed for *Lactococcus lactis MG1363* (**LL**) and *Streptococcus thermophilus LMG18311* (**ST**). The GSMMs used were obtained from the work of Flahaut et al., 2013 [(Flahaut *et al*, 2013)] and Pastink et al., 2009 [(Pastink *et al*, 2009)]. Experimental data of LL and ST's growth rates at different temperatures were obtained from Chen et al., 2015 [(Chen *et al*, 2015)] and Vaningelgem et al., 2004 [(Vaningelgem *et al*, 2004)]. 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 $361.14\ \frac{mmol}{hr * gE}$ [(Christiansen & Hengstenberg, 1999)], and $540\ \frac{mmol}{hr * gE}$ [(Geertsma *et al*, 2005)]

(Eq. 10). Because both lactose and glucose were sufficient in the growth medium [(Vaningelgem *et al*, 2004)][(Chen *et al*, 2015)], 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 $6360\ \frac{mmol}{hr * gE}$ [(Schumacher, 2018)][(Regueira *et al*, 2021)] (Eq. 10).

$$\phi_T * P_{TOT} = MW_{AT} * [E_{AT}] + MW_{CT} * [E_{CT}] = \frac{v_{AT}}{a_{AT}} + \frac{v_{CT}}{a_{CT}}\ (Eq. 10)$$

Temperature dependent k_{cat} values were predicted for enzymes in two bacteria's central carbon metabolism (**see SI, Table 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 [(Wegmann *et al*, 2007)], and ST, GCF_000011825.1 [(Bolotin *et al*, 2004)]. 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, 1981 [(Simon & Hofer, 1981)]. Therefore, this study used DLTKcat to fill the gap and examined DLTKcat's performance in predicting metabolic responses to temperature changes.

3. Results

3.1 DLTKcat outperforms previous models on temperature dependent k_{cat} prediction

With optimal hyperparameters (**see SI, Figure S4**), the model training process reduced RMSE scores of predicted $log_{10}(k_{cat})$ of the test dataset from 1.33 to 0.88, and enhanced R2 scores from 0.25 to 0.66 after 20 epochs (**Figure 2A**). The R2 scores of previously published

deep learning models on temperature dependent were all reported to be below 0.5 [(Yu *et al*, 2023)][(Kroll *et al*, 2023)], and DLTKcat has outperformed them by reaching a R2 score of 0.66 (**Figure 2B**). In addition, DLTKcat showed good prediction accuracy with low RMSE scores for datasets of experimental 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, and a robust accuracy for target values (experimental values) at different ranges.

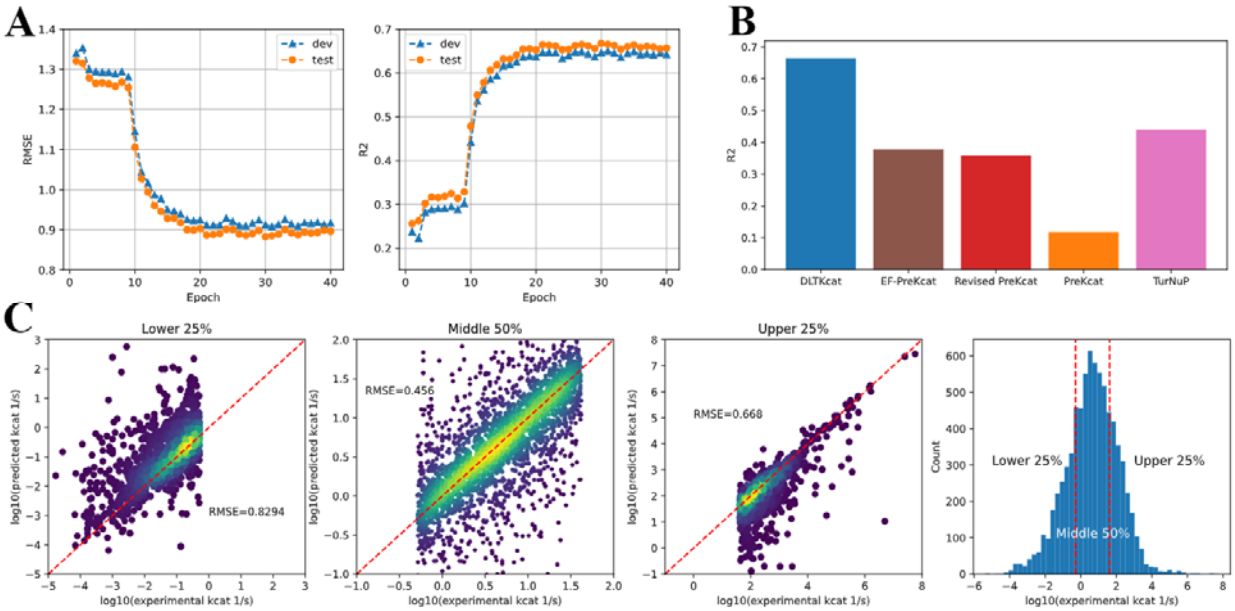

Figure 2. Assessment of the model performance. (A) The RMSE and R2 scores of prediction during the training process. test: the test set; dev: the validation set. (B) Comparison of R2 scores of DLTKcat, EF-PreKcat, Revised PreKcat, PreKcat [(Yu *et al*, 2023)] and TurNuP [(Kroll *et al*, 2023)] on prediction with temperature values. (C) RMSE scores of prediction for datasets with experimental values at lower 25%, middle 50% and upper 25%.

Subsequently, the accuracy of DLTKcat for WT/mutant enzymes and enzymes from different metabolic pathways was examined. For both WTs and mutants, the R2 and RMSE

scores were all around 0.8 and 0.6 respectively (**Figure 3A**). When predicting values for different metabolic pathways, DLTKcat showed its ability to discriminate enzymes in primary metabolism - catabolism/energy (primary-CE) and other pathways, with a significant higher distribution (p-value < 0.05) of predicted values in primary-CE (**Figure 3B**). Also, DLTKcat could predict values for enzymes from different metabolic pathways with R2 scores around 0.8 and RMSE scores around 0.6. In short, DLTKcat could well characterize enzymes from different contexts.

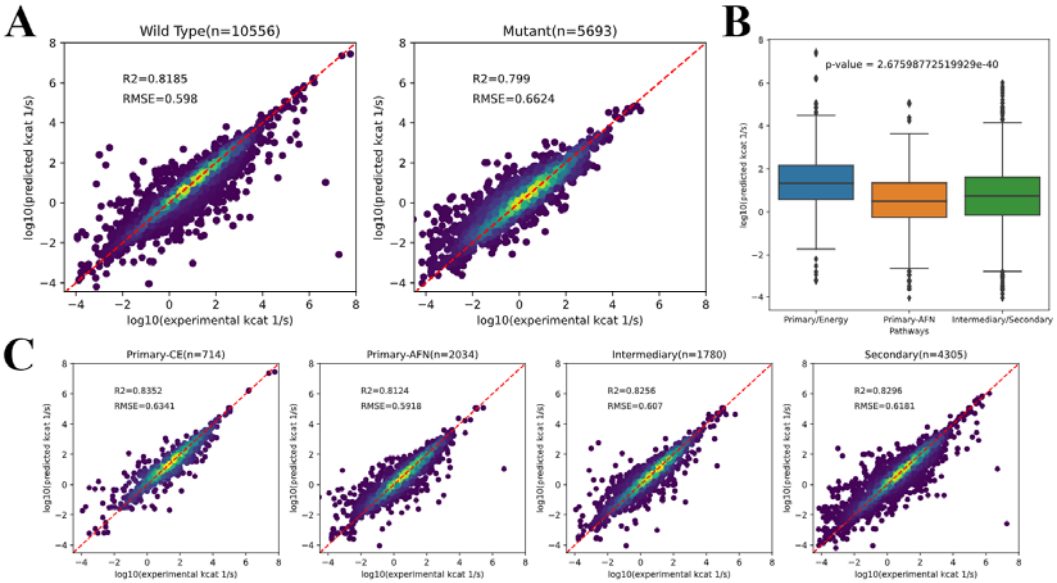

Figure 3. The accuracy of prediction for enzymes from different contexts. (A) R2 and RMSE scores of prediction for WT and mutant enzymes. (B) The comparison of distributions of predicted values in different metabolic pathways (p-value < 0.05). (C) R2 and RMSE scores of prediction for enzymes in primary metabolism - catabolism/energy (primary-CE), primary metabolism - amino acid/fatty acid/nucleotide (primary-AFN), intermediary and secondary metabolism.

3.2 The contribution of temperature related features to prediction

Before feature importance analysis, the prediction accuracy was examined for different temperature ranges ($T < 20°\mathrm{C}$, $20°\mathrm{C} \leq T \leq 40°\mathrm{C}$ and $T > 40°\mathrm{C}$). High R2 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 4A**). Then, feature shuffling, also known as feature permutation, was performed to show the importance of temperature and inverse temperature values. It was observed that the RMSE score increased by around 0.1 and R2 score decreased by around 0.1 when temperature related features were shuffled (**Figure 4B**). For high ($T > 40°\mathrm{C}$) and low ($T < 20°\mathrm{C}$) temperature ranges, the increase of RMSE and decrease of R2, caused by feature shuffling, became larger (**Figure 4CD**). In short, the decrease in prediction accuracy with shuffled temperature related features demonstrated the importance of temperature related features in DLTKcat.

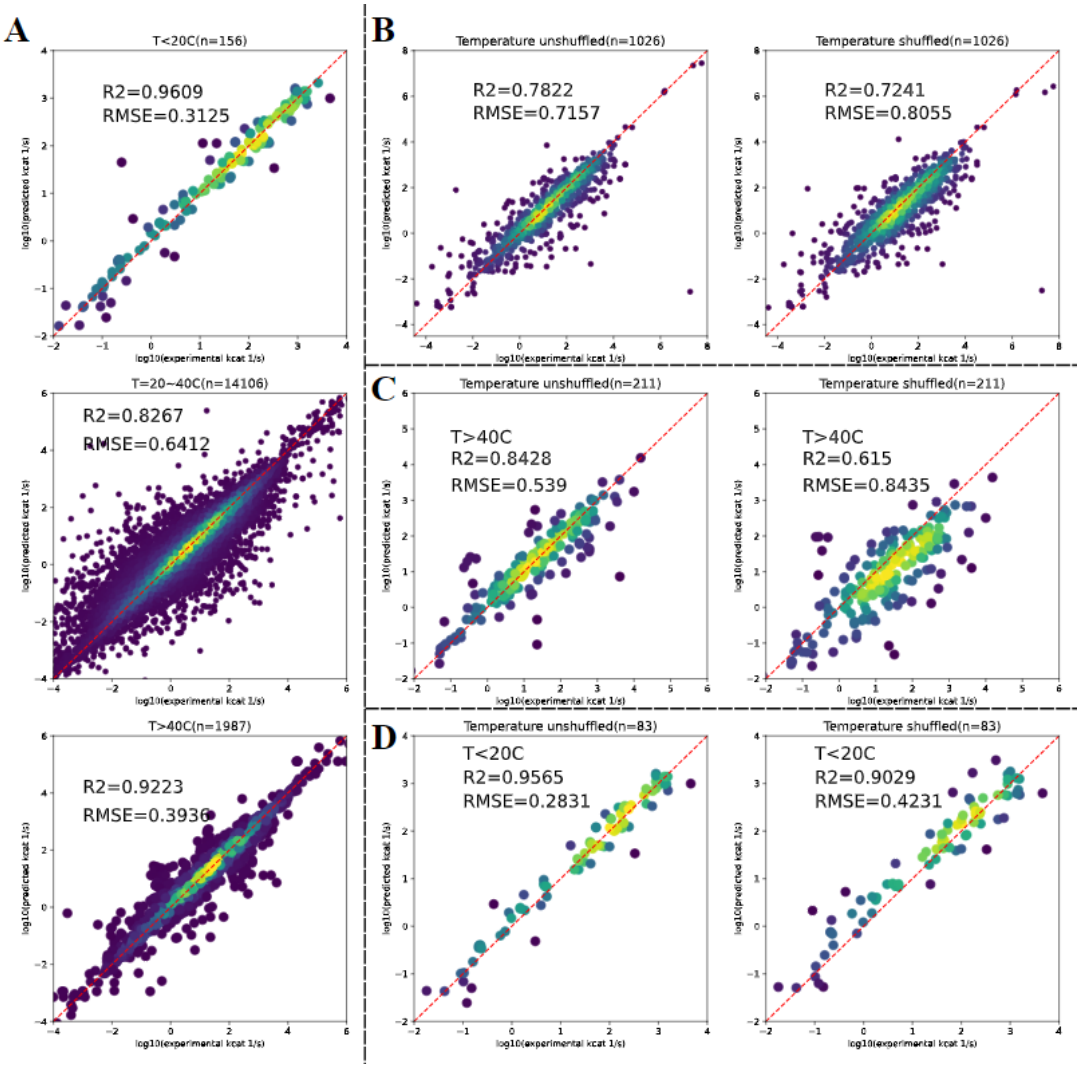

Figure 4. The importance of temperature related features in DLTKcat. (A) R2 and RMSE scores of predicted values at low (<20□), middle (20~40□) and high (>40□) temperature ranges. (B) R2 and RMSE scores of predicted values with unshuffled and shuffled temperature related features for the selected dataset with 1026 entries. (C) R2 and RMSE scores of predicted values with unshuffled and shuffled temperature related features for entries of high temperature. (D) R2 and RMSE scores of predicted values with unshuffled and shuffled temperature related features for entries of low temperature.

3.3 DLTKcat can predict the effect of protein sequence mutations on

Pyrococcus furiosus Ornithine Carbamoyltransferase values of WT and mutants, generated via directed evolution, at 30□ and 55□ were obtained from Roovers et al., 2001 [(Roovers *et al*, 2001)]. The protein sequence of *Pyrococcus furiosus* Ornithine Carbamoyltransferase was obtained from Uniprot with the Uniprot ID of Q51742. The prediction achieved high accuracy (RMSE = 0.5) (**Figure 5A**). Predicted values at 55 □ were higher than those at 30□ (**Figure 5ABC**), which was both consistent with the experimental data and the nature of *Pyrococcus furiosus* being a hyperthermophile favoring high temperature [(Fiala & Stetter, 1986)].

With respect to the effect of mutations, DLTKcat suggested that amino acid substitutions at 227th, 240th and 277th amino acids could increase the value, consistent with the experimental data, despite that the numerical difference between predicted values of mutants and WT was small (**Figure 5BC**; 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 value than the substitution at each single site, though it failed to predict that the of A240D/E277G was higher than that of Y227C/E277G (**Figure 5BC**).

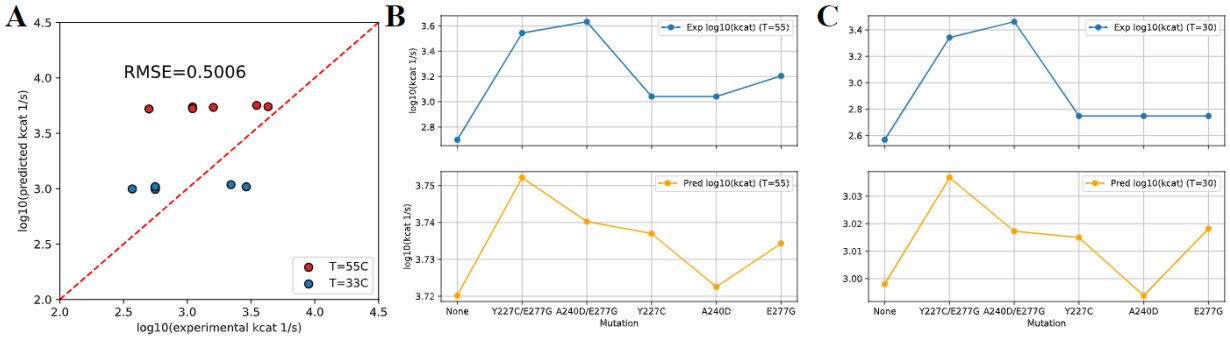

Figure 5. Prediction of the effect of SAVs on values. (A) Comparison between experimental and predicted of Pyrococcus furiosus Ornithine Carbamoyltransferase, RMSE = 0.5006. (B) Experimental (blue line) and predicted (orange line) values of WT and

mutants at 55□. (C) Experimental (blue line) and predicted (orange line) $log_{10}(k_{cat})$ values of WT and mutants at 30□. Exp: experimental value; Pred: predicted value.

3.4 Temperature sensitive metabolic modeling with predicted k_{cat}

DLTKcat predicted k_{cat} values for enzymes of *Lactococcus lactis MG1363* (LL) at 30, 32, 34, 36 and 38 □, and of *Streptococcus thermophilus LMG18311* (ST) at 25, 32, 37, 42, 46 and 49 □, which were temperatures where LL and ST's growth rates were measured in experimental data [(Vaningelgem *et al*, 2004)], [(Chen *et al*, 2015)]. DLTKcat predicted that k_{cat} of most catabolic enzymes in LL would decrease when temperature increased from 30 □ to 38 □, especially for glucose-6-phosphate isomerase (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 □ [(Chen *et al*, 2015)]. For catabolic enzymes in ST, DLTKcat predicted that most enzymes' k_{cat} would increase when temperature increased from 25 □ to 42 □, especially for fructose-bisphosphate aldolase (FBA), 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 □ is consistent with both the experimental data [(Vaningelgem *et al*, 2004)] and the nature of ST being a thermophile [(Harnett *et al*, 2011)]. 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 FBA was low. In proteome constrained FBA for LL, the k_{cat} of fructose-bisphosphate aldolase

(FBA) in LL was fixed at [(Callens *et al*, 1991)] in sacrifice of temperature sensitivity, because predicted values at different temperatures were unrealistically low ( ), compared with experiment values in other bacteria [(Callens *et al*, 1991)][(Plater *et al*, 1999)]. The predicted growth rates of LL by proteome constrained FBA 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 FBA predicted the increase of ST's growth rate from 25 to 42 , but it failed to predict the drop of growth rate from 42 to 49 (**Figure 6D**). Also, the predicted increase of values from 42 to 49 by DLTKcat (**see SI, Figure S5**) contradicted the experimental finding that 49 is close to the theoretical maximum temperature for ST to survive, 47~50 [(Harnett *et al*, 2011)]. To conclude, the log10-scale RMSE score within 1 of DLTKcat is not low enough to enable temperature sensitive proteome constrained FBA to predict bacterial growth and metabolism with good quantitative accuracy.

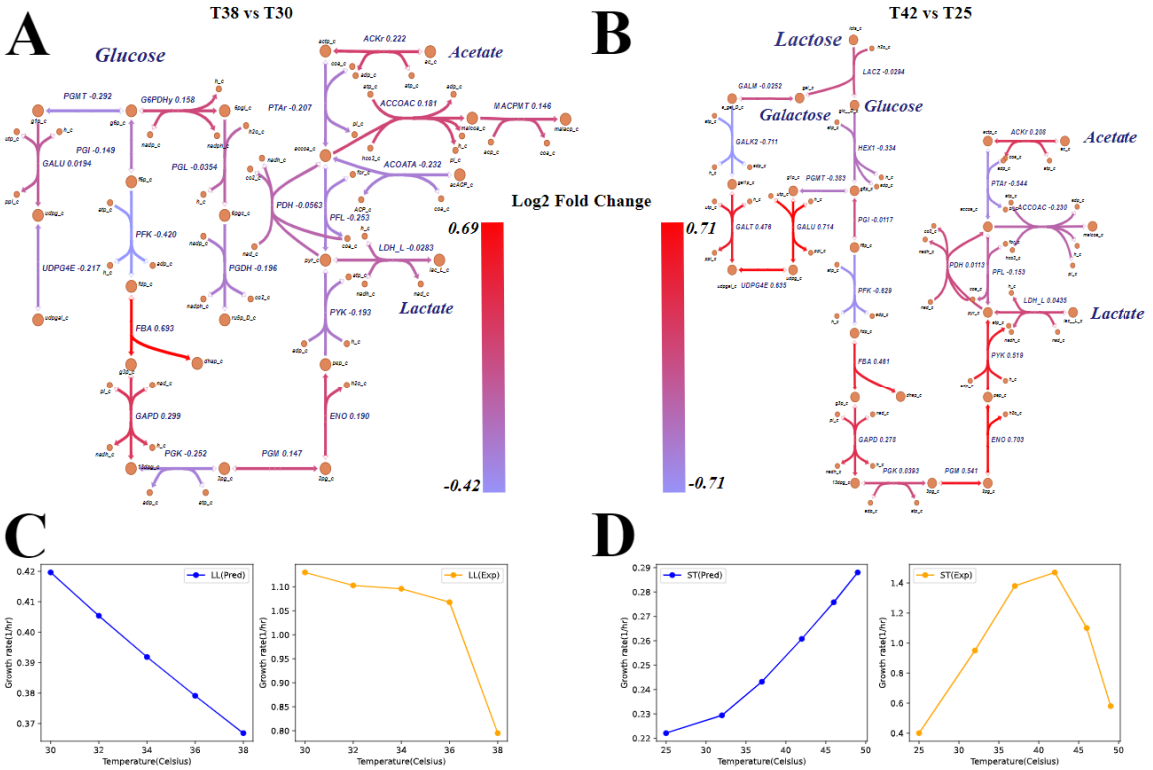

Figure 6. Prediction of bacteria metabolism at different temperatures. (A) Log2 fold change of predicted k_{cat} values for LL at 38 □ and 30 □ (38 □ vs 30 □). (B) Log2 fold change of predicted k_{cat} values for ST at 42 □ and 25 □ (42 □ vs 25 □). (C) Comparison of predicted (blue) and experimental (orange) growth rates of LL at 30, 32, 34, 36 and 38 □. (D) Comparison of predicted (blue) and experimental (orange) growth rates of ST at 25, 32, 37, 42, 46 and 49 □. LL: *Lactococcus lactis MG1363*; ST:*Streptococcus thermophilus LMG18311*; Exp: experimental value; Pred: predicted value. Reaction information can be found in **SI, Table S1&2**.

4. 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 compound protein interaction [(Arroyo *et al*, 2022)]. To tackle the challenging task, this study constructed a CPI deep learning model called DLTKcat. DLTKcat used bi-directional attention mechanism, whose advantage in capturing important regions in compounds and proteins was already shown by BACPI [(Li *et al*, 2022b)]. 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} [(Arroyo *et al*, 2022)]. Also, oversampling on entries at low and high temperature ranges compensated for the imbalanced distribution of temperature in the dataset (**see SI, Figure S3**). As a result, DLTKcat showed superior performance (log10-scale RMSE = 0.88, R2 = 0.66) than previously published models (e.g., Revised PreKcat, TurNuP) and robust accuracy for k_{cat} predictions in different 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 *Pyrococcus furiosus* Ornithine Carbamoyltransferase at different temperatures (**section 3.3**), DLKcat 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 3.4**) 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 FBA. In short, deep learning can gap fill a few missing k_{cat} values in the metabolic network, as done in Li et al., 2022 [(Li *et al*, 2022a)], but the accuracy of proteome constrained FBA will not be high if most proteome constraints are based on predicted k_{cat} values.

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 [(Yu *et al*, 2023)]. Including the optimal enzyme temperature either from databases or predictions [(Gado *et al*, 2020)] 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 [(Li *et al*, 2019)][(Gado *et al*, 2020)].

Overall, DLTKcat can provide accurate predictions of k_{cat} and account for the effect of temperature changes. Two case studies (3.3 and 3.4) 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 [(Quinlan, 1980)][(Maggi *et al*, 2018)]. 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.

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.

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.

Conflict of Interest Statement

The authors declare that there is no conflict of interests.

Data availability statement

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

References

Adadi R, Volkmer B, Milo R, Heinemann M & Shlomi T (2012) Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. *PLoS Comput*

Biol 8: e1002575

Arroyo JI, Díez B, Kempes CP, West GB & Marquet PA (2022) A general theory for temperature dependence in biology. *Proc Natl Acad Sci U S A* 119: e2119872119

Bolotin A, Quinquis B, Renault P, Sorokin A, Ehrlich SD, Kulakauskas S, Lapidus A, Goltsman E, Mazur M, Pusch GD, *et al* (2004) Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus. *Nat Biotechnol* 22: 1554–1558

Bulović A, Fischer S, Dinh M, Golib F, Liebermeister W, Poirier C, Tournier L, Klipp E, Fromion V & Goelzer A (2019) Automated generation of bacterial resource allocation models. *Metab Eng* 55: 12–22

Callens M, Kuntz DA & Opperdoes FR (1991) Kinetic properties of fructose bisphosphate aldolase from Trypanosoma brucei compared to aldolase from rabbit muscle and Staphylococcus aureus. *Mol Biochem Parasitol* 47: 1–9

Chen J, Shen J, Ingvar Hellgren L, Ruhdal Jensen P & Solem C (2015) Adaptation of Lactococcus lactis to high growth temperature leads to a dramatic increase in acidification rate. *Sci Rep* 5: 14199

Christiansen I & Hengstenberg W (1999) Staphylococcal phosphoenolpyruvate-dependent phosphotransferase system--two highly similar glucose permeases in Staphylococcus carnosus with different glucoside specificity: protein engineering in vivo? *Microbiology* 145 (Pt 10): 2881–2889

Coppens L, Tschirhart T, Leary DH, Colston SM, Compton JR, Hervey WJ IV, Dana KL, Vora GJ, Bordel S & Ledesma-Amaro R (2023) Vibrio natriegens genome-scale modeling reveals insights into halophilic adaptations and resource allocation. *Mol Syst Biol* 19: e10523

Davidi D, Noor E, Liebermeister W, Bar-Even A, Flamholz A, Tummler K, Barenholz U, Goldenfeld M, Shlomi T & Milo R (2016) Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements. *Proc Natl Acad Sci U S A* 113: 3401–3406

Ebrahim A, Lerman JA, Palsson BO & Hyduke DR (2013) COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst Biol* 7: 74

Estévez SR (2022) Robaina/BRENDApyrser: zenodo

Fiala G & Stetter KO (1986) Pyrococcus furiosus sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100 C. *Arch Microbiol* 145: 56–61

Flahaut NAL, Wiersma A, van de Bunt B, Martens DE, Schaap PJ, Sijtsma L, Dos Santos VAM & de Vos WM (2013) Genome-scale metabolic model for Lactococcus lactis MG1363 and its application to the analysis of flavor formation. *Appl Microbiol Biotechnol* 97: 8729–8739

Gado JE, Beckham GT & Payne CM (2020) Improving Enzyme Optimum Temperature Prediction with Resampling Strategies and Ensemble Learning. *J Chem Inf Model* 60: 4098–4107

Geertsma ER, Duurkens RH & Poolman B (2005) The activity of the lactose transporter from Streptococcus thermophilus is increased by phosphorylated IIA and the action of beta-galactosidase. *Biochemistry* 44: 15889–15897

Goelzer A, Muntel J, Chubukov V, Jules M, Prestel E, Nölker R, Mariadassou M, Aymerich S, Hecker M, Noirot P, *et al* (2015) Quantitative prediction of genome-wide resource allocation in bacteria. *Metab Eng* 32: 232–243

Harnett J, Davey G, Patrick A, Caddick C & Pearce L (2011) Lactic Acid Bacteria | Streptococcus thermophilus. In *Encyclopedia of Dairy Sciences (Second Edition)*, Fuquay JW (ed) pp 143–148. San Diego: Academic Press

Heckmann D, Campeau A, Lloyd CJ, Phaneuf PV, Hefner Y, Carrillo-Terrazas M, Feist AM, Gonzalez DJ & Palsson BO (2020) Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. *Proc Natl Acad Sci U S A* 117: 23182–23190

Jahn M, Crang N, Janasch M, Hober A, Forsström B, Kimler K, Mattausch A, Chen Q, Asplund-Samuelsson J & Hudson EP (2021) Protein allocation and utilization in the versatile chemolithoautotroph Cupriavidus necator. *Elife* 10

Karimi M, Wu D, Wang Z & Shen Y (2019) DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* 35: 3329–3338

Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, *et al* (2023) PubChem 2023 update. *Nucleic Acids Res* 51: D1373–D1380

Kingma DP & Ba J (2014) Adam: A Method for Stochastic Optimization. *arXiv [csLG]*

Kroll A, Engqvist MKM, Heckmann D & Lercher MJ (2021) Deep learning allows genome-scale prediction of Michaelis constants from structural features. *PLoS Biol* 19: e3001402

Kroll A, Rousset Y, Hu X-P, Liebrand NA & Lercher MJ (2023) Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. *Nat Commun* 14: 4139

Landrum G (2006) RDKit: Open-source cheminformatics. 2006. *Google Scholar*

Li F, Yuan L, Lu H, Li G, Chen Y, Engqvist MKM, Kerkhoven EJ & Nielsen J (2022a) Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis*: 1–11

Li G, Rabe KS, Nielsen J & Engqvist MKM (2019) Machine Learning Applied to Predicting Microorganism Growth Temperatures and Enzyme Catalytic Optima. *ACS Synth Biol* 8: 1411–1420

Li M, Lu Z, Wu Y & Li Y (2022b) BACPI: a bi-directional attention neural network for compound–protein interaction and binding affinity prediction. *Bioinformatics* 38: 1995–2002

Lim S, Lu Y, Cho CY, Sung I, Kim J, Kim Y, Park S & Kim S (2021) A review on compound-

protein interaction prediction methods: Data, format, representation and model. *Comput Struct Biotechnol J* 19: 1541–1556

Li S, Wan F, Shu H, Jiang T, Zhao D & Zeng J (2020) MONN: A multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Syst* 10: 308–322.e11

Maas AL, Hannun AY & Ng AY (2013) Rectifier nonlinearities improve neural network acoustic models.

Madhavan A, Arun KB, Binod P, Sirohi R, Tarafdar A, Reshmy R, Kumar Awasthi M & Sindhu R (2021) Design of novel enzyme biocatalysts for industrial bioprocess: Harnessing the power of protein engineering, high throughput screening and synthetic biology. *Bioresour Technol* 325: 124617

Maggi F, M. Tang FH & Riley WJ (2018) The thermodynamic links between substrate, enzyme, and microbial dynamics in Michaelis-Menten-Monod kinetics. *Int J Chem Kinet* 50: 343–356

Mori M, Hwa T, Martin OC, De Martino A & Marinari E (2016) Constrained Allocation Flux Balance Analysis. *PLoS Comput Biol* 12: e1004913

Nilsson A, Nielsen J & Palsson BO (2017) Metabolic Models of Protein Allocation Call for the Kinetome. *Cell Syst* 5: 538–541

Orth JD, Thiele I & Palsson BØ (2010) What is flux balance analysis? *Nat Biotechnol* 28: 245–248

Pastink MI, Teusink B, Hols P, Visser S, de Vos WM & Hugenholtz J (2009) Genome-scale model of Streptococcus thermophilus LMG18311 for metabolic comparison of lactic acid bacteria. *Appl Environ Microbiol* 75: 3627–3633

Plater AR, Zgiby SM, Thomson GJ, Qamar S, Wharton CW & Berry A (1999) Conserved residues in the mechanism of the E. coli Class II FBP-aldolase. *J Mol Biol* 285: 843–855

Qiu S, Zeng H, Yang Z, Hung W-L, Wang B & Yang A (2023) Dynamic metagenome-scale metabolic modeling of a yogurt bacterial community. *Biotechnol Bioeng*

Quinlan AV (1980) The Thermal Sensitivity of Michaelis-Menten Kinetics as a Function of Substrate Concentration. *J Franklin Inst* 310: 325–342

Regueira A, Rombouts JL, Aljoscha Wahl S, Mauricio□Iglesias M, Lema JM & Kleerebezem R (2021) Resource allocation explains lactic acid production in mixed□culture anaerobic fermentations. *Biotechnology and Bioengineering* 118: 745–758 doi:10.1002/bit.27605 [PREPRINT]

Rogers D & Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50: 742–754

Roovers M, Sanchez R, Legrain C & Glansdorff N (2001) Experimental evolution of enzyme temperature activity profile: selection in vivo and characterization of low-temperature-adapted mutants of Pyrococcus furiosus ornithine carbamoyltransferase. *J Bacteriol* 183: 1101–1105

Schomburg I, Jeske L, Ulbrich M, Placzek S, Chang A & Schomburg D (2017) The BRENDA enzyme information system–From a database to an expert system. *J Biotechnol* 261: 194–206

Schumacher R (2018) Metabolic trade-offs arising from increased free energy conservation in Saccharomyces cerevisiae.

Shin B, Park S, Kang K & Ho JC (09--10 Aug 2019) Self-Attention Based Molecule Representation for Predicting Drug-Target Interaction. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, Doshi-Velez F Fackler J Jung K Kale D Ranganath R Wallace B & Wiens J (eds) pp 230–248. PMLR

Simon WA & Hofer HW (1981) Phosphofructokinases from Lactobacteriaceae. II. Purification and properties of phosphofructokinase from Streptococcus thermophilus. *Biochim Biophys Acta* 661: 158–163

Stephanopoulos G (2012) Synthetic biology and metabolic engineering. *ACS Synth Biol* 1: 514–525

UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* 51: D523–D531

Vaningelgem F, Zamfir M, Adriany T & De Vuyst L (2004) Fermentation conditions affecting the bacterial growth and exopolysaccharide production by Streptococcus thermophilus ST 111 in milk-based medium. *J Appl Microbiol* 97: 1257–1273

Wegmann U, O'Connell-Motherway M, Zomer A, Buist G, Shearman C, Canchaya C, Ventura M, Goesmann A, Gasson MJ, Kuipers OP, *et al* (2007) Complete genome sequence of the prototype lactic acid bacterium Lactococcus lactis subsp. cremoris MG1363. *J Bacteriol* 189: 3256–3270

Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 28: 31–36

Wendering P, Arend M, Razaghi-Moghadam Z & Nikoloski Z (2023) Data integration across conditions improves turnover number estimates and metabolic predictions. *Nat Commun* 14: 1485

Wittig U, Rey M, Weidemann A, Kania R & Müller W (2018) SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. *Nucleic Acids Res* 46: D656–D660

Yu H, Deng H, He J, Keasling J & Luo X (2023) Highly accurate enzyme turnover number prediction and enzyme engineering with PreKcat.

Zeng H & Yang A (2020) Bridging substrate intake kinetics and bacterial growth phenotypes with flux balance analysis incorporating proteome allocation. *Sci Rep* 10: 4283