Machine learning models for the prediction of enzyme properties should be tested on proteins not used for model training

Alexander Kroll[1] & Martin J. Lercher[1*]

[1] Institute for Computer Science and Department of Biology, Heinrich Heine University, D-40225 Düsseldorf, Germany

[*] Correspondence to: Martin.Lercher@hhu.de

Abstract

The turnover number k_{cat} quantifies the catalytic efficiency of enzymes. As experimental k_{cat} estimates are expensive and time consuming, it is desirable to develop computational pipelines able to predict turnover numbers of arbitrary enzymes from easily accessible features. Advances in deep learning have now put such predictions into reach [1]. In a recent publication, Li et al. [2] described DLKcat, a general deep learning model for k_{cat} predictions. The authors state that their approach facilitates "high-throughput k_{cat} prediction for metabolic enzymes from any organism merely from substrate structures and protein sequences." Furthermore, they claim that "DLKcat can capture k_{cat} changes for mutated enzymes". Here, we show that DLKcat predictions are accurate only for enzymes that are highly similar to proteins used for training, and become positively misleading for enzymes without close homologs in the training data. We further show that DLKcat's mutant predictions – all of which were made for enzymes highly similar to training data – are much less accurate than implied by the DLKcat publication, capturing only 3% of the experimentally observed variation across mutants not included in the training data.

Machine learning models can overfit the training data, i.e., the model may provide a very good description of the training dataset but may generalize poorly to unseen data. To reduce the risk of overfitting, it is standard practice to clearly distinguish between (i) validation data, used for choosing among model designs and selecting hyperparameters; (ii) training data, used for fitting optimal model parameters; and (iii) test data, which is set aside before training and validation and which is only used to assess the final model quality [3]. Deep neural networks are particularly prone to overfitting. Their predictions on the training data are almost guaranteed to be good, regardless of whether the trained model is able to make meaningful predictions for data not used for training [4]. Moreover, by choosing hyperparameters that lead to a model with low errors for the validation set, the model will also likely overfit to noise in the validation set. Prediction quality must thus be evaluated exclusively on test data. In violation of this practice, Li et al. include all available data in all main text figures. Accordingly, 90% of the datapoints in the main text figures were also used for hyperparameter choice (10%) and model training (80%), while only 10% of the datapoints – the test data – are independent of the model construction. Accordingly, the main text Figures 2b-d and 3 of Ref. 2 provide no information about the model's ability to make predictions beyond the datasets already used for its training and design: the meaningful test data signal is drowned out by the uninformative training and validation data. For the main text Figures 2b-c and 3a-b, the authors chose to delegate the corresponding test data figures to Supplementary Figures 5-6, while they show no separate test data results for Figures 2d and 3c-f. Presenting the results in this way will likely mislead a large fraction of the readership, in particular those with no experience in machine learning methodology.

A related and central problem in the construction of DLKcat lies in the assignment of the same enzyme sequences to the training and test sets. If the aim is to generate a prediction model that generalizes well to data points that are not highly similar to training data points, this aim should be reflected in the test data: it should be sufficiently different from the training data, such that its analysis indeed provides information on the model's generalization ability [5]. Specifically, the performance of machine learning predictions for enzyme features depends strongly on the sequence similarity between a target enzyme and enzymes in the training set [6], consistent with the widely held notion that enzymes with more similar amino acid sequences are more likely to be functionally similar [7]. Accordingly, it is likely much easier to predict unknown k_{cat} values for enzymes that were used in model training than to make predictions for enzymes that have no close homologs with known kinetic constants. More than two thirds (67.9%) of the enzymes in the DLKcat test set are also included in the training data, and an additional 23.3% have amino acid sequences that are at least 99% identical to sequences in the training data. Thus, 91.2% of the datapoints used for testing are for enzymes at least 99% identical to proteins used for training.

The red line in **Fig. 1** shows how DLKcat's prediction quality depends on the sequence identity between an enzyme in the test dataset and the most similar enzymes used for training. The figure shows sliding window estimates of the coefficient of determination, R^2, which is a widely used standard measure for prediction quality (see equation (2) in Ref. 2). R^2=1 indicates perfect predictions. As seen from **Fig. 1**, DLKcat's coefficients of determination are negative for maximal sequence identities between test and training data below 60%. Thus, when no close homologs have been used for training, DLKcat predictions are worse than simply assuming the same mean k_{cat} value for all reactions, which corresponds to R^2=0.

A simple and often used strategy to approximate unknown k_{cat} values is to use values from enzymes with similar sequences. Under what conditions does DLKcat outperform this simple strategy? For each enzyme in the test dataset, we calculated the geometric mean across the k_{cat} values of the three most similar enzymes in the training dataset, independent of the substrates. The cyan line in **Fig. 1** shows the corresponding coefficients of determination. The simple average k_{cat} predictions are marginally better than those provided by DLKcat when calculated across all data points in the test set ($R2$=0.452 vs. $R2$ = 0.445, N = 1687).

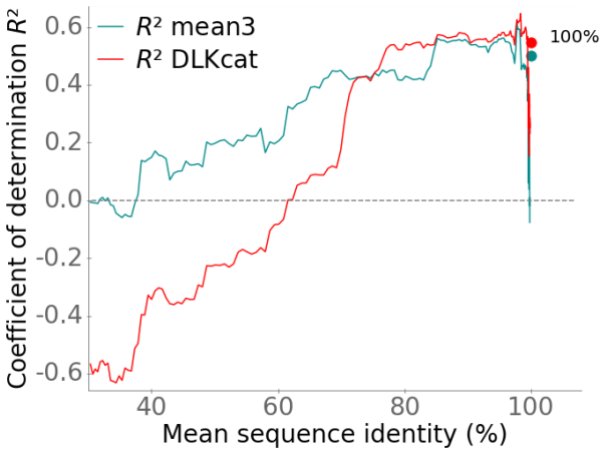

Figure 1. DLKcat predictions become reasonable only when closely related enzymes were used for training (max. sequence identity > 70%), and are barely better than simple k_{cat} averages even when the same enzyme was used for training. The curves are coefficients of variation R^2, calculated in sliding windows of size n=100 across sequences in the test set ordered by the maximal sequence identity between individual test enzymes and all sequences in the training data. Position on the x-axis indicates the mean across the window. Red: DLKcat predictions; cyan: geometric mean of k_{cat} values, calculated over the three most similar enzymes in the training set. The two points at the top right are for test datapoints with enzymes already used for training (100% max. sequence identity, N=1143); these were not included in the sliding windows.

Two thirds of the test data is for enzymes already used for training (100% max. sequence identity, N=1143). For this subset, prediction quality barely differs between the approaches, with R^2=0.547 for DLKcat's deep learning model and R^2=0.534 for the simple averages. For enzymes without close homologs in the training data (max. sequence identity <70%), predictions from simple averages are substantially more accurate than those of DLKcat.

The same problems apply to the subset of mutant enzymes. 385 out of 744 mutant enzymes in the test dataset were also used for training, and a further 354 sequences are at least 99% identical to sequences in the training set. Only 5 mutant enzymes in the test set have a sequence identity below 99% compared to enzymes used for training. From this data, it is impossible to derive any general statements about predicting the effects of mutations on k_{cat} , except for those enzymes already used for training DLKcat.

Arguably the most striking result presented in Ref. 2 is its Fig. 3c, which presents a "comparison between predicted and measured k_{cat} values for several well-studied enzyme-substrate pairs with rich experimental mutagenesis data". The panel is augmented with a Pearson correlation coefficient r=0.94 and corresponding p-value < 10^{-91}, suggesting that DLKcat can accurately predict the effects of mutations on k_{cat}. However, only 15.6% of the data in this panel had been set aside for testing, while 71.9% of the datapoints shown were part of the training data. Moreover, it is not clear why Li et al. calculate a combined Pearson correlation across the measured and predicted k_{cat} values of multiple enzyme-substrate pairs in their Figures 3b and 3c. Due to systematic differences across enzyme-substrate pairs, this design would lead to strong correlations even for a prediction model that outputs the wildtype k_{cat} for all mutants of a given pair.

Fig. 2 shows a more appropriate presentation of the Fig. 3c data, appropriately restricted to datapoints from the test set. To make data across pairs comparable, we scaled the k_{cat} values for each enzyme-substrate pair to z-scores. The z-scaled test data still shows a statistically significant correlation between measured and predicted k_{cat} values (Pearson's r=0.498, p=0.0051, N=30). However, it is important to note that 14 of the 30 mutant enzymes in the test set had also been used for training, and each of the remaining 16 mutant amino acid sequences is at least 99.4% identical to a sequence in the training set. Despite this close relationship between test and training enzymes, DLKcat predicts only a small fraction of the variation in k_{cat} due to mutations: when calculating a coefficient of determination across the z-scores to focus on mutation effects (rather than baseline effects of the different enzyme-substrate pairs), one obtains a low coefficient of determination, R^2=0.162. When restricting this analysis to the 16 mutants that were not in the training data, the coefficient of determination is further reduced to R^2=0.030. Thus, for

previously unseen mutants, DLKcat predicts only 3% of the variance relative to the wildtype k_{cat}. For mutants of enzyme-substrate pairs not included in the training data, the results provide no information on the accuracy of DLKcat predictions.

In sum, while Li et al. claim that DLKcat can predict k_{cat} for metabolic enzymes from any organism and can capture k_{cat} changes for mutated enzymes, these claims were not substantiated by their results. For enzymes without close homologs in the training data (<60% sequence identity), DLKcat predictions are positively misleading (R^2<0, **Fig. 1**) and perform worse than simple averages across the k_{cat} of similar enzymes. Overall, DLKcat provides no clear advantage over this much simpler averaging approach. In particular, DLKcat provides little benefit when predicting turnover numbers for enzyme families and mutants not already characterized kinetically, which arguably constitute the most important use cases in most applications, including the parameterization of enzyme-constrained genome-scale metabolic models (ecGEMs). This disappointing performance is likely rooted in the construction of the DLKcat datasets, which did not challenge the model to learn to predict k_{cat} values for enzymes dissimilar to those used for training.

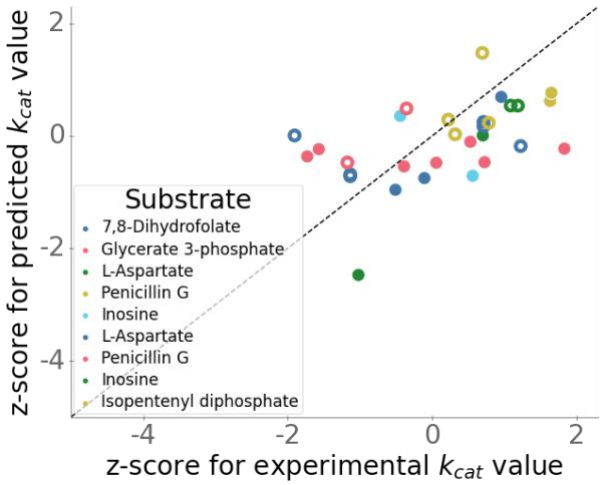

Figure 2. DLKcat predicts only a small fraction of k_{cat} variation due to mutations (R^2=0.163, r^2=0.248), even though the test enzyme sequences or very close relatives had been used for training. Colors indicate different enzyme-substrate pairs. Data from Fig. 3c of Ref. 2; only datapoints from the test dataset are shown. Open circles are mutants identical to enzymes in the training data, while solid dots are mutants with between 99.4% and 99.8% sequence identity to training data. We scaled the log10-transformed k_{cat} values to z-scores as follows. For each enzyme-substrate pair, we calculated mean and standard deviation across all measured k_{cat} for the different mutants in the complete dataset; from each measured and each predicted k_{cat} for this pair, we then subtracted the mean and divided the result by the standard deviation.

Methods

To reproduce the DLKcat model and to make predictions for the corresponding test set, we downloaded the code provided on GitHub by Li et al. [2]. To calculate the maximal amino acid sequence identity between each enzyme in the test set and all enzymes in the training set, we used the Needleman-Wunsch algorithm implemented in the EMBOSS software package [8].

Code and data availability

All software was coded in Python 3. The code used to generate the results of this paper, in the form of Jupyter notebooks, as well as all datasets, are available from https://github.com/AlexanderKroll/DLkcat_Matters_Arising .

Acknowledgements

We are grateful to the authors of Ref. 2 for comments on an earlier draft. Computational support and infrastructure was provided by the "Centre for Information and Media Technology" (ZIM) at Heinrich Heine University Düsseldorf (Germany). This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC 1310, and, under Germany's Excellence Strategy, through grant EXC 2048/1 (Project ID: 390686111).

Conflict of interest

The authors declare that they have no conflicts of interest.

Author contributions

AK & MJL conceived of the study, performed the analyses, interpreted the results, and wrote the manuscript.

References

1. Heckmann, D. *et al.* Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nat Commun* **9**, 5252 (2018).
2. Li, F. *et al.* Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nat Catal* **5**, 662–672 (2022).
3. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning*, page 118. (MIT Press, 2016).
4. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* **64**, 107–115 (2021).
5. Søgaard, A., Ebert, S., Bastings, J. & Filippova, K. We Need to Talk About Random Splits. Preprint at http://arxiv.org/abs/2005.00636 (2021).
6. Kroll, A., Ranjan, S., Engqvist, M. K. M. & Lercher, M. J. The substrate scopes of enzymes: a general prediction model based on machine and deep learning. Preprint at https://doi.org/10.1101/2022.05.24.493213 (2022).
7. Tian, W. & Skolnick, J. How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* **333**, 863–882 (2003).
8. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276–277 (2000).