Title: Segmentation-Aware Multi-Attention Learning for Enzyme k_{cat} Prediction

Presenting Author Details: Ramesh Chandra, Ph.D

Abstract:

One of the key kinetic parameters describing catalytic efficiency of the enzyme, the enzyme turnover number (k_{cat}), plays an important role in enzyme engineering, metabolic engineering, systems biology, synthetic biology, and pharmaceutical research. Estimation of metabolic flux, identification of pathway bottlenecks, enzyme redesign and rational optimization of these enzymes necessitate these essential k_{cat} values. However, for many years these values are estimated experimentally and experimental determination of k_{cat} is a costly, time-consuming, and technically challenging task. In other side, the kinetic annotation of available protein sequences is still extremely limited despite the exponential growth of available protein sequences in databases since years. In comparison to these traditional methods, the recent deep learning approaches have shown promising predictive capabilities with low cost and utilising less time. Yet their performance is often limited due to some of the reasons like, inadequate representation of long protein sequences, poor capture of local catalytic motifs, and poor biological interpretability. These limitations highlight the need for an interpretable computational framework, capable of simultaneously capturing both local sequence characteristics and long-range contextual dependencies with better performance scores.

To surmount these obstacles, we introduced KcatNeuroCortex, a hybrid deep learning framework that integrates segmentation-based sequence representation, Bidirectional Gated Recurrent Units (Bi-GRU) and multi-attention mechanism for enzyme turnover number prediction. Protein sequences were divided into overlapping segments so as to preserve local functional motifs and to maintain contextual continuity over the whole sequence. We combined protein representations generated with a transformer model with molecular representations of substrates obtained from a pre-trained molecular transformer network. These complementary features were then processed by Bi-GRU layers to learn bidirectional sequential dependencies, followed by a multi-attention module to dynamically weight catalytically informative sequence regions before regression-based prediction of kcat values. We evaluated the model performance on the popular DLKcat benchmark dataset, containing 16,838 experimentally validated enzyme-substrate pairs from distinct Enzyme Commission classes. We did extensive benchmarking against multiple state-of-the-art machine learning and deep learning methods using coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), Pearson correlation coefficient (PCC), and other statistical performance indicators. We

also performed extensive ablation studies, attention visualization, enzyme-class analysis, sequence-length robustness evaluation and a biological case study on the terpenoid indole alkaloid biosynthetic pathway in Catharanthus roseus to validate the predictive ability and biological interpretability of the model.

KcatNeuroCortex consistently outperformed previously published computational methods in all major evaluation metrics. The proposed framework resulted in an R^2 of 0.74 with an RMSE of 0.77, which corresponded to an approximate 57.45% improvement in the predictive accuracy and a nearly 30% reduction in the prediction error relative to the DLKcat baseline. Comparative analysis showed better generalization than several state-of-the-art models. Ablation experiments proved that the sequence segmentation and the multi-attention module were the main reasons for performance improvement of the KcatNeuroCortex. Reduction by 17.6% in R^2 and 27.3% increase in RMSE were observed after removal of individual components, which revealed that sequence segmentation was the most influential module, highlighting its critical role in capturing localized catalytic motifs. Similarly, upon excluding multi-head attention mechanism, a 12.2% decline in R^2 was observed, confirming its importance in identifying biologically relevant sequence regions and improving model interpretability. Furthermore, a substantial deterioration in model performance (R^2 = 0.52, RMSE = 1.12), after replacing ProT5 embeddings with randomly initialized embeddings demonstrates the the value of transfer learning from protein language models. Furthermore, replacing pretrained ProT5 embeddings with randomly initialized embeddings caused model performance to deteriorate substantially (R^2 = 0.52, RMSE = 1.12), demonstrating the value of transfer learning from protein language models. These findings confirm that the synergistic integration of enzyme and substrate representations, segmentation-based learning, pretrained embeddings, and multi-attention collectively contributes to the superior predictive capability of the model. To represent the applicability of the model, we further utilized KcatNeuroCortex for predicting the rate limiting enzyme involved on *Catharanthus roseus*. It also produced biologically meaningful predictions that agree with published experimental evidence, thereby demonstrating the practical applicability of the framework for pathway analysis and metabolic engineering. Evaluation on hydrolase enzymes confirmed that KcatNeuroCortex accurately identified conserved catalytically important sequence regions through its attention mechanism.

In conclusion, the KcatNeuroCortex framework is robust, scalable, and biologically interpretable, with a first-time developed model with segmentation-based and multi-attention learning during feature engineering. Hence, making it a valuable tool for the enzyme

engineering community and for researchers interested in the identification of large-scale kinetic parameter estimation. More than its predictive accuracy, the model represents the importance of feature engineering, showing impressive robustness across a variety of enzyme families and sequence lengths, leading towards more generalizability of the model.
Keywords: *kcat* prediction, Deep learning, Sequence-based features, Enzyme catalysis, Enzyme engineering