ENZYMES

ML helps predict enzyme turnover rates

Constraining metabolic models by enzyme capacities greatly improves genotype–phenotype predictions. Now, a method for estimating enzyme turnovers based on deep learning has been developed and used to reconstruct enzyme-constrained genome-scale metabolic models for more than 300 yeast species.

Veda Sheersh Boorla, Vikas Upadhyay and Costas D. Maranas

Genome-scale metabolic (GSM) models encode the complete metabolism of an organism through a list of gene-protein-reaction relations. Ultimately, these relations help establish constraints on the magnitude, direction and catalytic resource needed for the network of metabolic reactions. Flux balance analysis of GSM models (Fig. 1a) aims to quantify the flow of metabolites through these networks under different environmental and genetic constraints[1]. However, such analysis does not account for the fact that the reaction flux of an enzyme inside a cell (in vivo) is ultimately constrained by its catalytic capacity, defined as the product of the intracellular concentration and the catalytic turnover rates (k_{cat}) of the enzyme. Imposing enzyme constraints in GSM models (Fig. 1b) can be very informative for recapitulating important metabolic attributes such as maximum growth characteristics, metabolic shifts and proteome reallocations[2] using flux balance analysis calculations. In the absence of enzyme capacity constraints, all reactions are allowed to reach their stoichiometrically allowable limits, without accounting for the trade-offs organisms need to establish under different physical and energy constraints.

However, to date, genome-wide availability of in vivo k_{cat} values is lacking, even for well-studied organisms. Now, writing in *Nature Catalysis*, Eduard J. Kerkhoven and colleagues established a deep learning framework (DLKcat, Fig. 1c) to estimate the k_{cat} values of metabolic enzymes for any organism of choice[2]. The ability to predict k_{cat} values (even with a certain level of approximation) can facilitate the automated construction of enzyme-constrained GSMs, especially for non-model organisms.

Earlier efforts for constructing enzyme-constrained GSM models[3] relied on tabulated values from databases such as BRENDA and SABIO-RK. In most cases, organism-specific values were absent and thus were adopted from those of other organisms[4]. Alternatively, Heckmann et al[5]. trained regression models to predict enzyme turnover numbers from several hand-picked features spanning an enzyme's network context and biochemical and structural properties. However, owing to the requirement of several detailed features, this method was only trained on a few hundred *Escherichia coli* enzymes and hence had limited adaptability to other under-studied organisms. In DLKcat, Kerkhoven and colleagues used a representation-based learning approach to automatically extract underlying features in the input data relevant for k_{cat} prediction. In this approach, the enzymes' amino acid sequences and the structures of their corresponding substrates are discretized into sets of sequence words and substrate substructures (Fig. 1c), respectively, similar to a previously described procedure[2].

Each discrete word and substructure hence created is assigned a unique vector of real numbers, called an embedding[6]. The mathematical representations for enzyme and substrate features are then obtained by combining the embeddings of the comprising words and sub-structures using a convolutional neural network and a graph neural network for each enzyme and substrate, respectively. Obtained representations are then input into a fully connected neural network layer to predict k_{cat} values. The resulting neural network architecture, DLKcat, was trained iteratively using the training data curated by collating all available k_{cat} values in the BRENDA and SABIO-RK databases summing up to 16,838 data points. During training, both the weights of the neural networks and the embeddings are updated to minimize the root mean squared error (r.m.s.e.) between the predicted and true k_{cat} values for all enzyme–substrate pairs in the training data. The resulting model trained on this data accurately predicted k_{cat} values with a r.m.s.e. of 1.06 in $\log_{10}$ scale when evaluated on an unseen test dataset (the evaluation gave a Pearson's correlation coefficient of $r = 0.71$). This is quite an achievement given that k_{cat} values span a range of up to 10 to 14 orders of magnitude.

However, it is important to stress that the predicted k_{cat} values are only approximate in nature. Given that the correlation agreement was achieved for a log-based scale, the predicted values could be off by an order of magnitude or more when establishing upper limits of reaction fluxes. The advantage, however, is that using this tool, one can readily impose constraints on almost all enzymes of GSM models. Upon estimating a value for k_{cat}, the stoichiometric matrix of the model is updated to yield constraints of the form $\nu \leq k_{cat}$ [E] for each enzyme-catalysed reaction (Fig. 1b).

As training data for the model are based on in vitro experiments, the researchers further developed a Bayesian framework[7] that can estimate in vivo-like k_{cat} values starting from the DLKcat predicted k_{cat} values by minimizing the distance between the predicted and experimentally observed growth data. The researchers put forth DLKcat and the Bayesian framework, together, as an automated and convenient enzyme-constraint-based GSM (ecGSM) reconstruction pipeline, DL-ecGEM. Using DL-ecGEM, the researchers demonstrated 80% coverage of enzymes for k_{cat} values to define constraints for 90% of the enzymatic reactions for a set of 343 yeast/fungi species. In comparison, previous ecGSM models[4] only covered k_{cat} values for 40% of enzymes and generated constraints for 60% of annotated enzymatic reactions for the same set of species. In summary, DL-ecGEM is a convenient pipeline for reconstructing enzyme-constrained GSM models with near-complete coverage for any genome-sequenced organisms.

In addition to predicting k_{cat} values, the current work showcases the advantage of attention weights derived from the neural network to identify regions of enzyme sequence that have the highest impact on catalytic activity. This capability can be further extended to identify specific

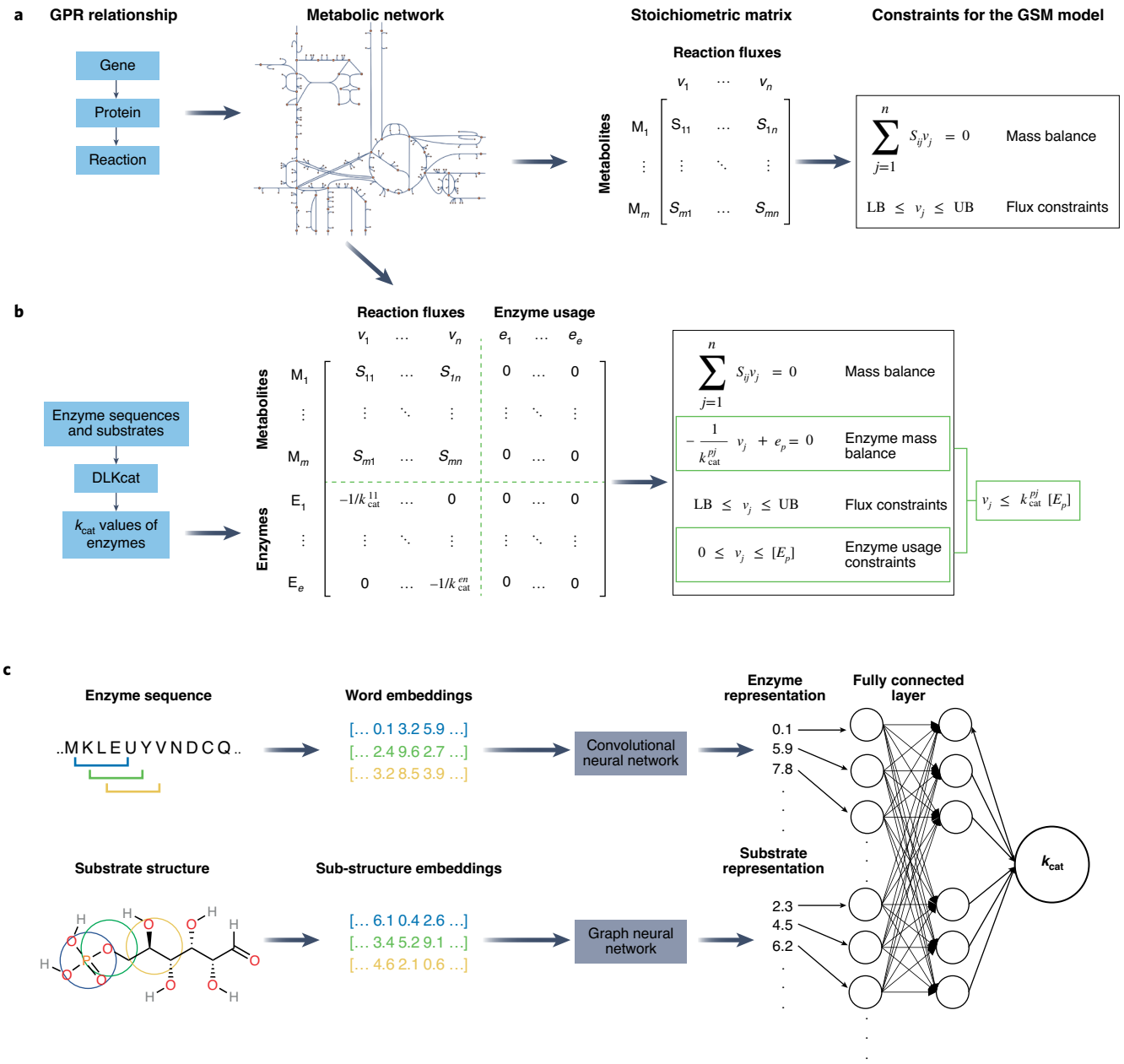

Fig. 1 | Machine learning aided enzyme constraint-based genome-scale modelling framework. a, Genome-scale metabolic (GSM) model formulated with mass and flux balance constraints. S_{ij} is the stoichiometry of the *i*th metabolite in the *j*th reaction, and v_j is the metabolic flux of the *j*th reaction. LB and UB are the lower and upper bounds, respectively, for the reaction fluxes. **b**, Enzyme-constrained GSM model formed by adding additional enzyme entries to the stoichiometric matrix using k_{cat} values obtained from the DLKcat neural network model. For each enzyme, *p*, e_p denotes its usage, and E_p denotes its maximum possible usage. **c**, DLKcat uses the enzyme sequence and substrate structure as inputs and encodes them into mathematical representations using convolutional and graph neural networks, respectively. The final representations are combined using a fully connected layer to output k_{cat} values. GPR, gene–protein–reaction.

amino acid residues that could potentially cause drastic changes in the k_{cat} value of any given enzyme. Amino acid residues identified as such can be modified to aid enzyme engineering efforts. Recent applications in deep learning for structure generation, such as the AlphaFold2 (ref. [8]), promise genome-wide prediction of protein structures at angstrom-level accuracy.

Although embeddings of the amino acid sequence words alone have proven to calculate accurate enzyme representations for k_{cat} prediction, using structural features could potentially boost the model's accuracy and interpretability, as demonstrated for similar applications[9].

One limitation of DLKcat is that it only incorporates the features of one primary substrate in the model, leaving cofactor/ion-dependent and/or non-unimolecular reactions approximately described. Also, enzyme turnover numbers are known to vary by orders of magnitude owing to variations in the experimental conditions such as pH and temperature, which is beyond the scope of DLKcat. As in all such studies, there is always the potential danger of

introducing systematic biases due to the over-representation of well-studied organisms in the training datasets. Hence, any extrapolation of results for other organisms and enzymes needs to be cautiously executed. ❒

Veda Sheersh Boorla, Vikas Upadhyay and Costas D. Maranas ✉
Department of Chemical Engineering, The Pennsylvania State University, University Park, PA, USA.
✉*e-mail: costas@psu.edu*

Published online: 19 August 2022
https://doi.org/10.1038/s41929-022-00827-x

References

1. Wang, H. et al. *Proc. Natl Acad. Sci. USA* **118**, 30 (2021).
2. Li, F. et al. *Nat. Catal.* https://doi.org/10.1038/s41929-022-00798-z (2022).
3. Chen, Y. & Nielsen, J. *Curr. Opin. Syst. Biol.* **25**, 50–56 (2021).
4. Domenzain, I. et al. *Nat. Commun.* **13**, 3766 (2022).
5. Heckmann, D. et al. *Nat. Commun.* **9**, 5252 (2018).
6. Tsubaki, M., Tomii, K. & Sese, J. *Bioinformatics* **35**, 309–318 (2019).
7. Li, G. et al. *Nat. Commun.* **12**, 190 (2021).
8. Jumper, J. et al. *Nature* **596**, 583–489 (2021).
9. Gligorijević, V. et al. *Nat. Commun.* **12**, 3168 (2021).

Competing interests

The authors declare no competing interests.