nature portfolio

https://doi.org/10.1038/s41929-022-00798-z

Supplementary information

Deep learning-based k_{cat} prediction enables improved enzyme-constrained model reconstruction

In the format provided by the authors and unedited

Supplementary Information

Deep learning based k_{cat} prediction enables improved enzyme constrained model reconstruction

Feiran Li[1, #], Le Yuan[1, 2, #], Hongzhong Lu[1], Gang Li[1], Yu Chen[1], Martin K. M. Engqvist[1], Eduard J Kerkhoven[1, 2, *], Jens Nielsen[1, 3]

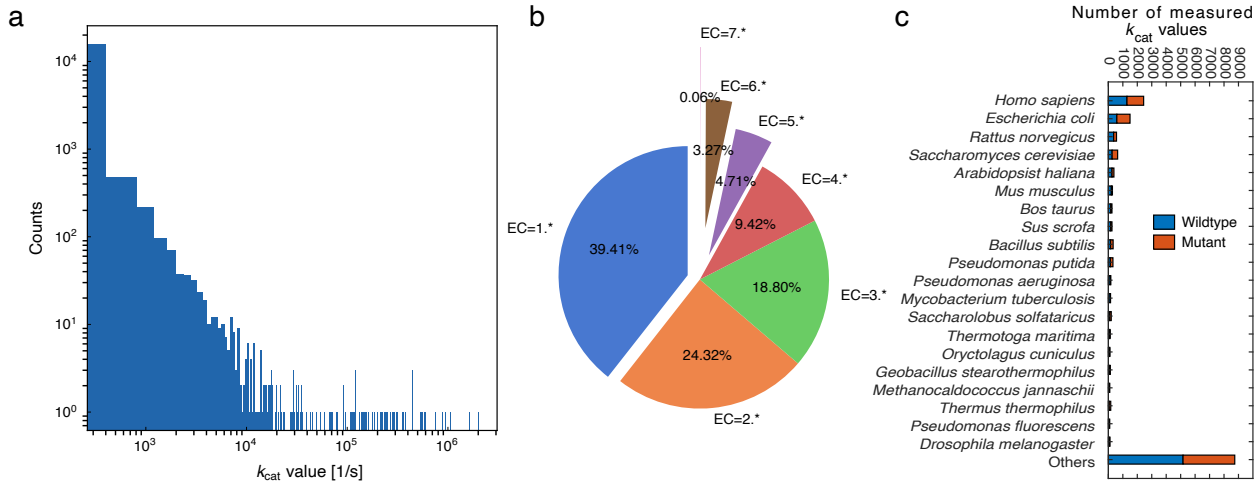
1 Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, SE-412 96 Gothenburg, Sweden

2 Novo Nordisk Foundation Center for Biosustainability, Chalmers University of Technology , Kemivägen 10, SE-412 96, Gothenburg, Sweden

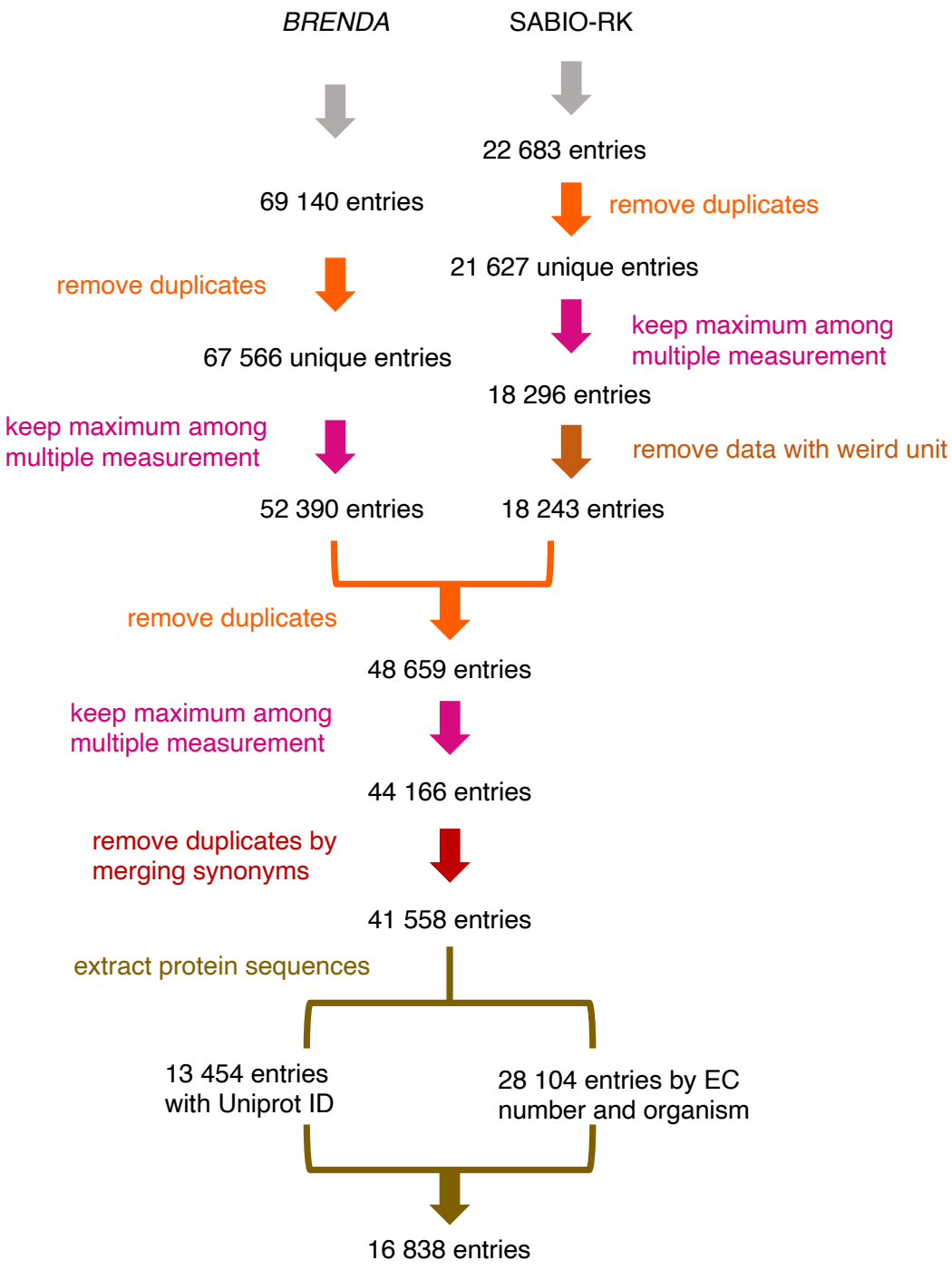
3 BioInnovation Institute, Ole Måløes Vej 3, DK2200 Copenhagen N, Denmark

# These authors contributed equally to this work: Feiran Li, Le Yuan.

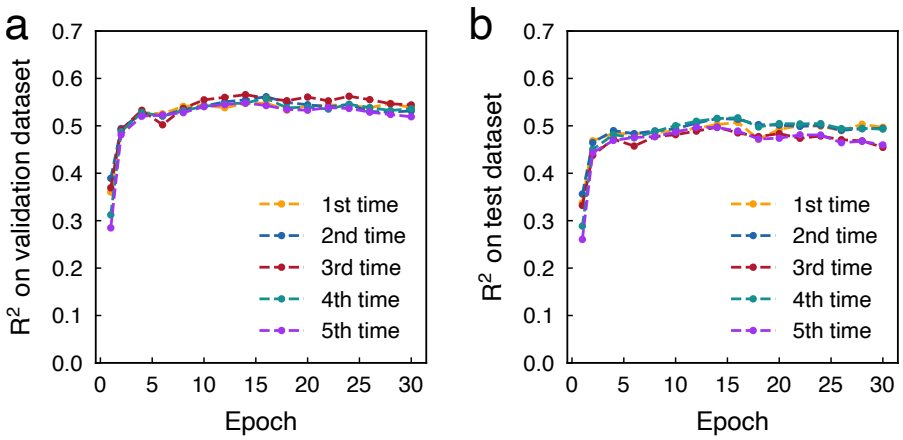
* Corresponding author:
Eduard J Kerkhoven, Email: eduardk@chalmers.se



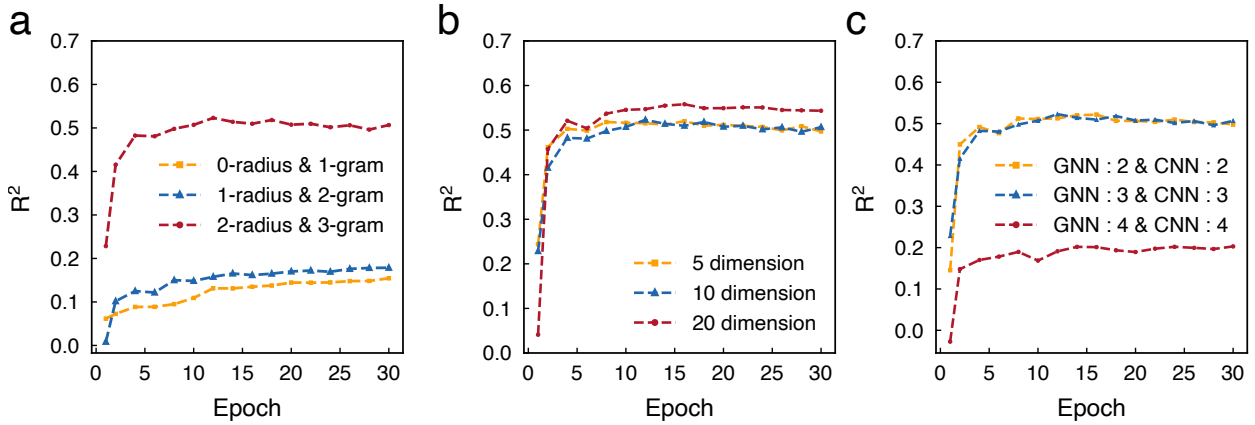
Supplementary Figure 1 Data analysis for *in vitro* k_{cat} values collected from the BRENDA and the SABIO-RK database after several rounds of data preprocessing, cleaning and combination. (a) Data distribution of *in vitro* k_{cat} values. (b) Classification of *in vitro* k_{cat} values based on the first digit of EC number. (c) Classification of *in vitro* k_{cat} values based on species.



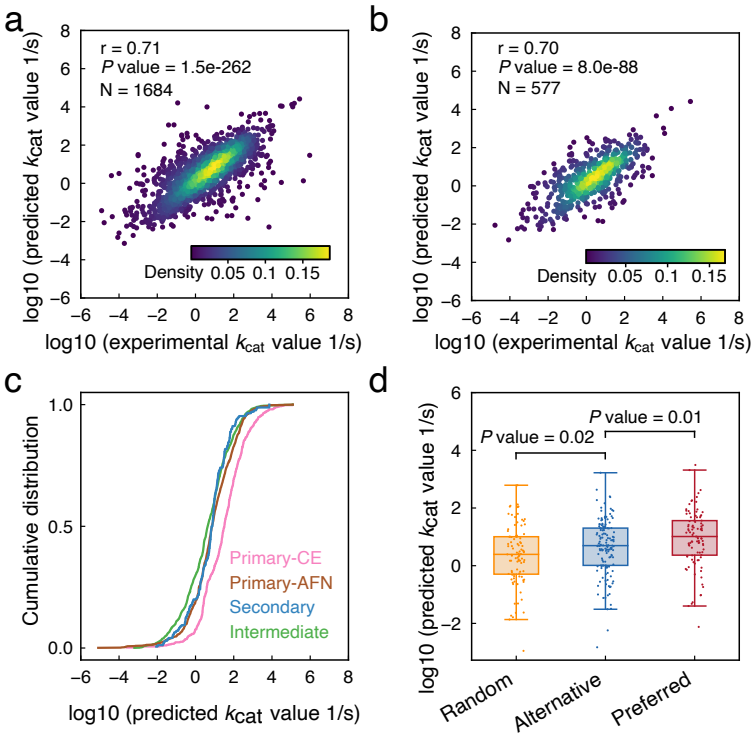
Supplementary Figure 2 Data cleaning process for deep learning model input.



Supplementary Figure 3 Performance of the deep learning model on the (a) validation dataset and (b) test dataset by running five times of random splitting.

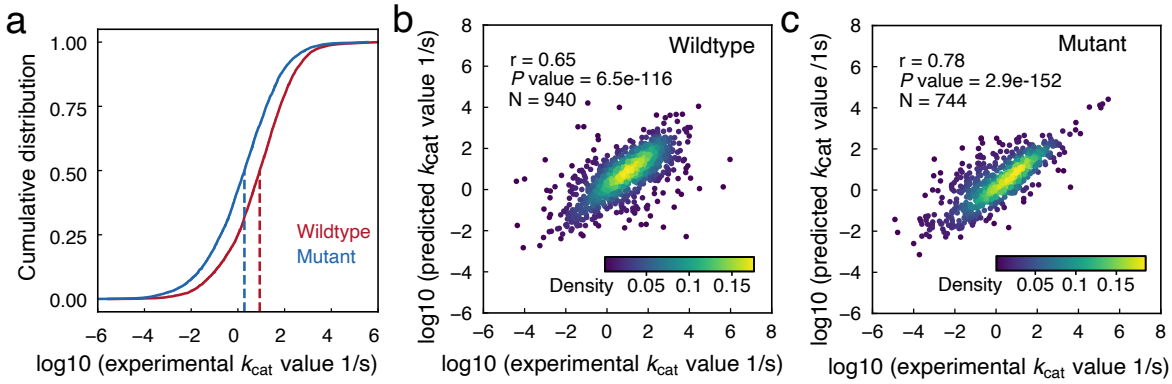


Supplementary Figure 4 Learning curves with various hyperparameters on the validation dataset, including (a) various r-radius subgraphs and n-gram amino acids, (b) various dimension vectors and (c) various numbers of layers in GNN and CNN.

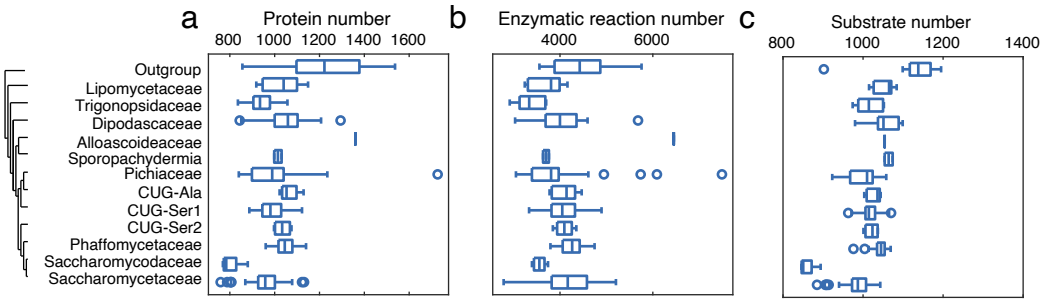


Supplementary Figure 5 Deep learning model performance for k_{cat} prediction. (a) Performance of the final deep learning model on test dataset. The correlation between predicted k_{cat} values and those present in the test dataset was evaluated. The brightness of colour represents the density of data points. (b) Performance of the final deep learning model on subset of test dataset where either the proteins sequence or the substrates were not involved in the training data set. In a-b, Student's t test was used to calculate *P* value for Pearson's correlation. (c) Cumulative distribution of deep learning-based k_{cat} values for enzyme-substrate pairs belonging to different metabolic contexts. Abbreviations: CE, carbohydrate and energy; AFN, amino acids, fatty acids, and nucleotides. (d) Enzyme promiscuity analysis on test dataset. For enzymes with multiple substrates, we divided the substrates as preferred and alternative by their experimental measured k_{cat}, then used the predicted k_{cat} values for this boxplot. Random substrates were randomly chosen from the compound dataset in our training data except the documented substrates and products for the tested enzyme. A two-sided Wilcoxon rank sum test was used to calculate *P* value. In the boxplot, the

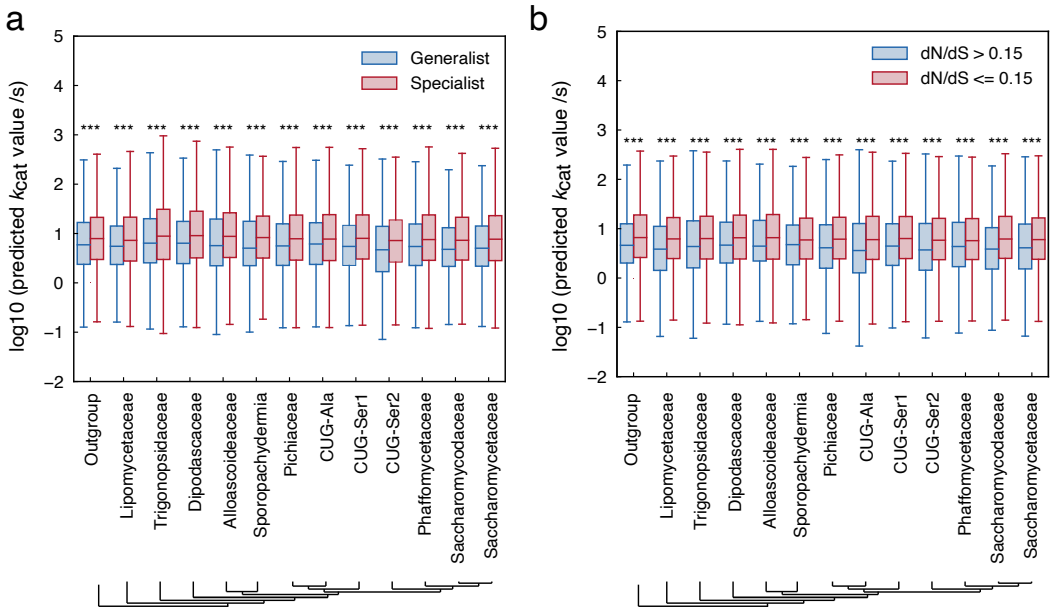
central band represents median value, box represents the upper and lower quartiles, and the whiskers extend up to 1.5 times the interquartile range beyond the box range.



Supplementary Figure 6 Comparison of k_{cat} values for wildtype and mutated enzymes and prediction performance for both types of enzymes on test dataset. (a) Cumulative distribution of experimentally measured k_{cat} values for wildtype and mutated enzymes. (b) Prediction performance of k_{cat} values for all of the wildtype enzymes via deep learning model. The brightness of colour represents the density of data points. (c) Prediction performance of k_{cat} values for all of the mutated enzymes via deep learning model. The brightness of colour represents the density of data points. Student's t test was used to calculate *P* value for Pearson's correlation for b-c.



Supplementary Figure 7 Comparison of (a) protein number, (b) enzymatic reaction number and (c) substrate number in deep learning predicted k_{cat} profiles of 343 yeast/fungi GEMs. The x-axis represents the outgroup (11 species) together with 12 major clades divided by the genus-level phylogeny for 332 yeast species. In each boxplot (a-c), the central band represents median value, box represents the upper and lower quartiles, and the whiskers extend up to 1.5 times the interquartile range beyond the box range.

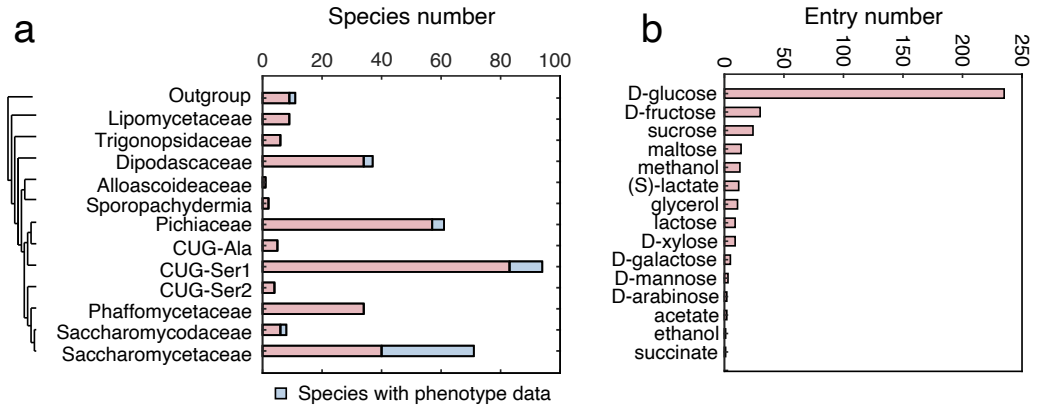


71

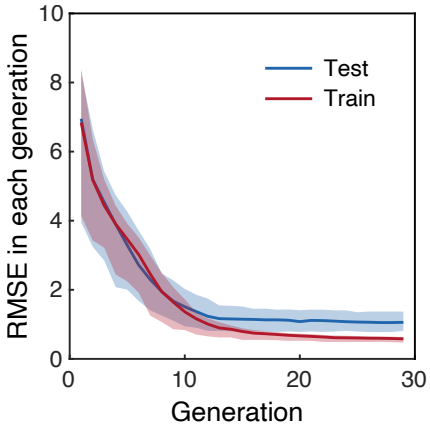
72 **Supplementary Figure 8** Evolution analysis of predicted k_{cat} values for 343 yeast/fungi species.
73 (a) Enzyme k_{cat} values linked with generalist and specialist in the GEMs for all 343 yeast/fungi
74 species. (b) Enzyme k_{cat} values linked with the ratio of non-synonymous to synonymous
75 substitution (dN/dS) analysis for all of 343 yeast/fungi species. The x-axis represents the outgroup
76 (11 species) together with 12 major clades divided by the genus-level phylogeny for 332 yeast
77 species. The cut-off of 0.15 was set according to the distribution of dN/dS values in these species.
78 In each boxplot (a-b), the central band represents median value, box represents the upper and lower
79 quartiles, and the whiskers extend up to 1.5 times the interquartile range beyond the box range. A
80 two-sided Wilcoxon rank sum test was used to calculate *P* value. *** means *P* value < 0.001 in
81 the correlation test analysis.

82

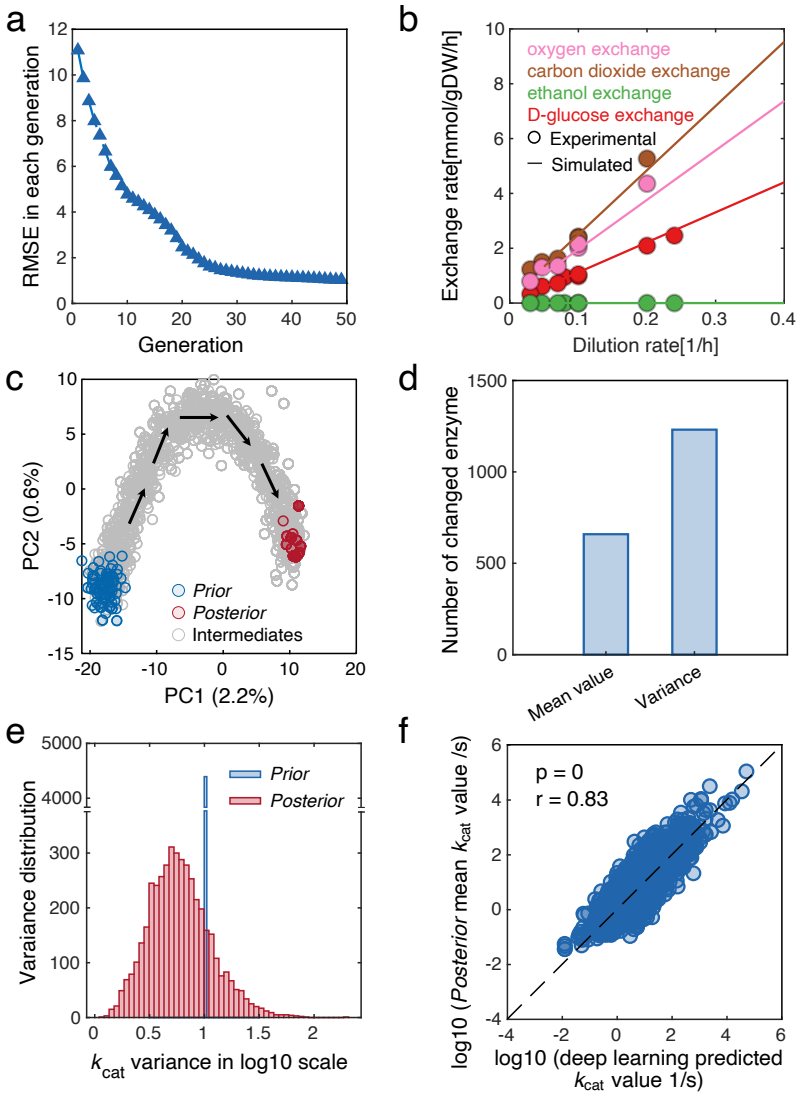
83



Supplementary Figure 9 Bayesian approach related information. (a) Species analysed in this study. (b) Collected growth phenotype data based on carbon source.

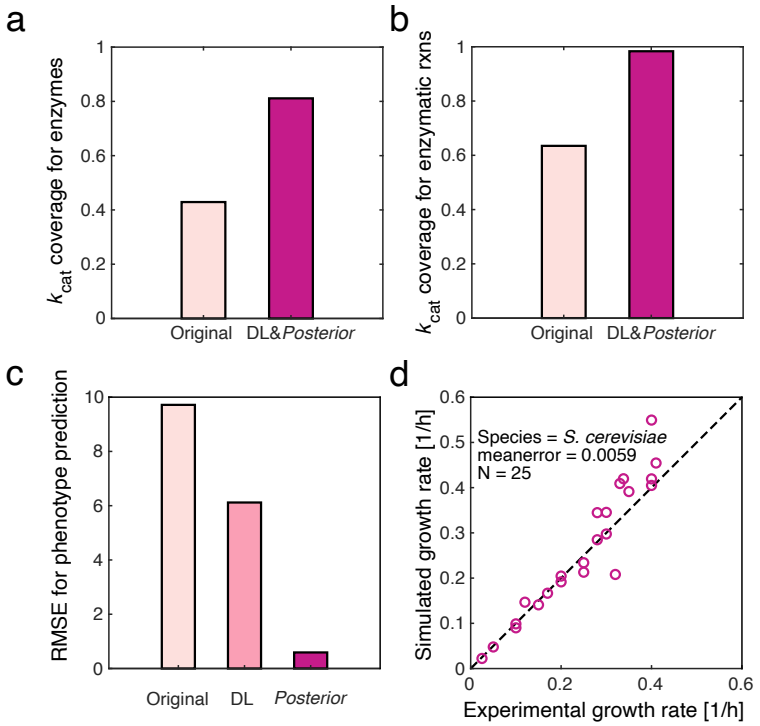


Supplementary Figure 10 Validation of SMC-ABC approach. In this validation approach, data of 50% of experimental points were used to update the *Prior* and then tested on the remaining 50%. The split was done by first sorting all the data based on growth rates, then choosing the ones with even index for training and others for test. Root mean square error (RMSE) was calculated as described in Methods. Lines indicate median values and shaded areas indicate regions between the 5-th and 95-th percentiles (n=100).

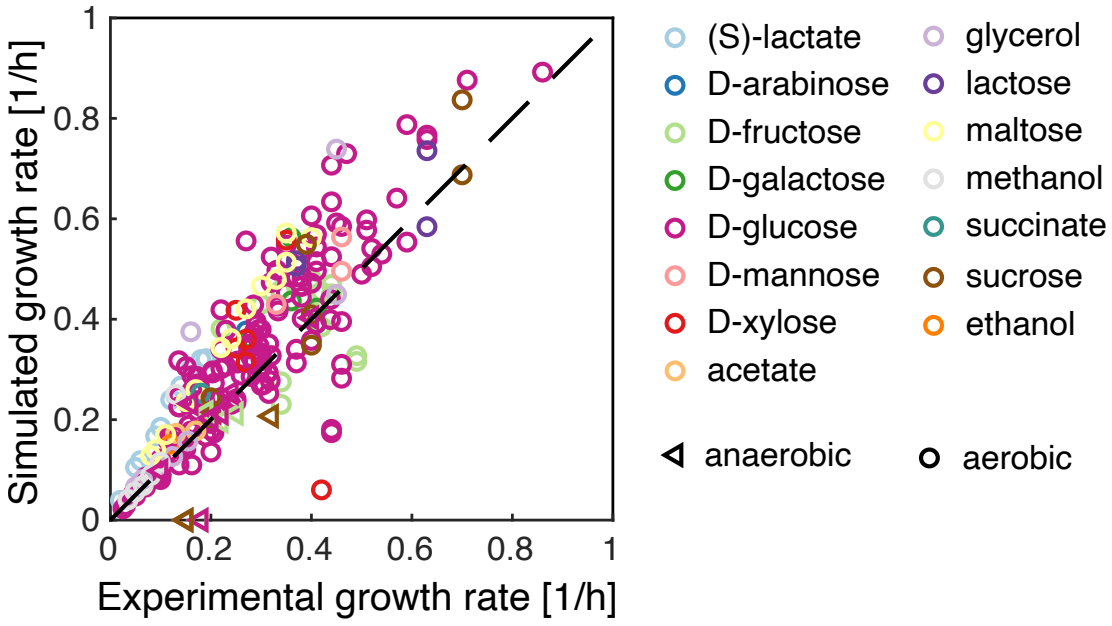

a
RMSE in each generation
Generation
b
Exchange rate[mmol/gDW/h]
Dilution rate[1/h]
oxygen exchange
carbon dioxide exchange
ethanol exchange
D-glucose exchange
Experimental
Simulated
c
PC2 (0.6%)
PC1 (2.2%)
Prior
Posterior
Intermediates
d
Number of changed enzyme
Mean value
Variance
e
Varaiance distribution
Prior
Posterior
kcat variance in log10 scale
f
p = 0
r = 0.83
log10 (Posterior mean kcat value /s)
log10 (deep learning predicted kcat value 1/s)

96
97

Supplementary Figure 11 Bayesian modelling training performance for *Y. lipolytica* ecGEM. (a) Root mean square error (RMSE) for phenotype measurement and prediction during Bayesian training process. (b) Simulated exchange rates by *Posterior*-mean-ecGEM (line) compared with experimental data (dot). k_{cat} values in the *Posterior*-mean-ecGEMs are mean values from 100 sampled *Posterior* datasets obtained from the Bayesian training process. (c) Principal component analysis (PCA) for k_{cat} datasets sampled during the Bayesian training approach, showing the progression from *Prior* to *Posterior* dataset. Each parameter in the set was standardized by subtracting the mean and then divided by the standard deviation before PCA. In blue are 100 sampled *Prior* datasets; in red are 100 sampled *Posterior* datasets; in grey are all other intermediate datasets. (d) The number of enzymes with a significantly changed mean (Šidák adjusted Welch's t test $P < 0.01$, two-sided) and variance (Šidák adjusted one-tailed F-test $P < 0.01$) between sampled *Prior* and *Posterior* k_{cat} datasets. (e) Variance distribution comparison for *Prior* and *Posterior* distribution. (f) Correlation between deep learning predicted k_{cat} and *Posterior* mean k_{cat}. Student's t test was used to calculate *P* value for Pearson's correlation



Supplementary Figure 12 Evaluation of three ecGEM modelling approaches including original-ecGEM, DL-ecGEMs and *Posterior*-mean-ecGEMs for *S. cerevisiae*. Enzymatic constraint coverage comparison for (a) enzymes and (b) enzymatic reactions. (c) Root mean square error (RMSE) for the phenotype prediction. (d) Growth prediction of *Posterior*-mean-ecGEM for *S. cerevisiae*.



Supplementary Figure 13 Growth prediction of *Posterior*-mean-ecGEMs on different carbon sources. This figure is detailed version of Fig. 5d.