THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Advancing systems biology of yeast through machine learning and comparative genomics

LE YUAN

Division of Systems and Synthetic Biology
Department of Life Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023

Advancing systems biology of yeast through machine learning and comparative genomics
LE YUAN
ISBN 978-91-7905-818-0

© Le Yuan, 2023.

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5284
ISSN 0346-718X

Division of Systems and Synthetic Biology
Department of Life Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone + 46 (0)31-772 1000

Cover illustration: Using machine learning techniques and comparative genomics to gain a deeper insight into gene essentiality, enzyme design and genome evolution of yeasts.

Printed by Chalmers digitaltryck
Gothenburg, Sweden 2023

Advancing systems biology of yeast through machine learning and comparative genomics

Le Yuan
Department of Life Sciences
Chalmers University of Technology

Abstract

Synthetic biology has played a pivotal role in accomplishing the production of high value commodities, pharmaceuticals, and bulk chemicals. Fueled by the breakthrough of synthetic biology and metabolic engineering, *Saccharomyces cerevisiae* and various other yeasts (such as *Yarrowia lipolytica*, *Pichia pastoris*) have been proven to be promising microbial cell factories and are frequently used in scientific studies. However, the cellular metabolism and physiological properties for most of the yeast species have not been characterized in detail. To address these knowledge gaps, this thesis aims to leverage the large amounts of data available for yeast species and use state-of-the-art machine learning techniques and comparative genomic analysis to gain a deeper insight into yeast traits and metabolism.

In this thesis, machine learning was applied to various unresolved biological problems on yeasts, i.e., gene essentiality, enzyme turnover number (k_{cat}), and protein production. In the first part of the work, machine learning approaches were employed to predict gene essentiality based on sequence features and evolutionary features. It was demonstrated that the essential gene prediction could be substantially improved by integrating evolution-based features. Secondly, a high-quality deep learning model DLKcat was developed to predict k_{cat} values by combining a graph neural network for substrates and a convolutional neural network for proteins. By predicting k_{cat} profiles for 343 yeast/fungi species, enzyme-constrained models were reconstructed and used to further elucidate the cellular metabolism on a large scale. Lastly, a random forest algorithm was adopted to investigate feature importance analysis on protein production, it was found that post-translational modifications (PTMs) have a relatively higher impact on protein production compared with amino acid composition.

In comparative genomics, a comprehensive toolbox HGTphyloDetect was developed to facilitate the identification of horizontal gene transfer (HGT) events. Case studies on some yeast species demonstrated the ability of HGTphyloDetect to identify horizontally acquired genes with high accuracy. In addition, through systematic evolution analysis (e.g., HGT, gene family expansion) and genome-scale metabolic model simulation, the underlying mechanisms for substrate utilization were further probed across large-scale yeast species.

Keywords: machine learning, deep learning, gene essentiality, enzyme turnover number, horizontal gene transfer, yeast species

List of Publications

This thesis is based on the following publications:

Paper I: Hongzhong Lu[†], Feiran Li[†], **Le Yuan[†]**, Iván Domenzain, Rosemary Yu, Hao Wang, Gang Li, Yu Chen, Boyang Ji, Eduard J Kerkhoven, Jens Nielsen. Yeast metabolic innovations emerged via expanded metabolic network and gene positive selection. *Molecular Systems Biology* 17.10 (2021): e10427.

Paper II: Feiran Li[†], **Le Yuan[†]**, Hongzhong Lu, Gang Li, Yu Chen, Martin K. M. Engqvist, Eduard J Kerkhoven, Jens Nielsen. Deep learning-based k_{cat} prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis* 5.8 (2022): 662-672.

Paper III: Feiran Li, Yu Chen[#], Qi Qi[#], Yanyan Wang[#], **Le Yuan**, Mingtao Huang, Ibrahim E. Elsemman, Amir Feizi, Eduard J Kerkhoven, Jens Nielsen. Improving recombinant protein production by yeast through genome-scale modeling using proteome constraints. *Nature Communications* 13.1 (2022): 2969. ([#]: contributed equally)

Paper IV: Le Yuan, Hongzhong Lu, Feiran Li, Jens Nielsen, Eduard J Kerkhoven. HGTphyloDetect: facilitating the identification and phylogenetic analysis of horizontal gene transfer. *Briefings in Bioinformatics* 24.2 (2023): bbad035.

[†]: co-first author, these authors contributed equally.

Additional papers and manuscripts not included in this thesis:

Paper V: Cheewin Kittikunapong, Feiran Li, **Le Yuan**, Hongzhong Lu, Eduard J Kerkhoven. Exploring the bacterial kinetome across environments and communities through deep learning-based prediction of turnover numbers. *mSphere* (2023) Under review

Paper VI: Kameshwara Peri, Karl Persson, **Le Yuan**, Fábio Faria-Oliveira, Eduard J Kerkhoven, Cecilia Geijer. Gene cluster dynamics enabling lactose metabolism in *Candida intermedia*. (2023) Manuscript

Paper VII: Yu Chen[†], Johan Gustafsson[†], Albert Enrique Tafur Rangel[†], Mihail Anton, Iván Domenzain, Cheewin Kittikunapong, Feiran Li, **Le Yuan**, Jens Nielsen, Eduard J Kerkhoven. Reconstruction, simulation and analysis of enzyme-constrained metabolic models using GECKO toolbox 3.0. *Nature Protocols* (2023) Under revision

Paper VIII: Mengying Han[†], Dachuan Zhang[†], Shaozhen Ding, Yu Tian, Xingxiang Cheng, **Le Yuan**, Dandan Sun, Dongliang Liu, Linlin Gong, Cancan Jia, Pengli Cai, Weizhong Tu, Junni Chen, Qian-Nan Hu. ChemHub: a knowledgebase of functional chemicals for synthetic biology studies. *Bioinformatics* 37.22 (2021): 4275-4276.

Paper IX: Shaozhen Ding, Yu Tian, Pengli Cai, Dachuan Zhang, Xingxiang Cheng, Dandan Sun, **Le Yuan**, Junni Chen, Weizhong Tu, Dong-Qing Wei, Qian-Nan Hu. novoPathFinder: a webserver of designing novel-pathway with integrating GEM-model. *Nucleic Acids Research* 48.W1 (2020): W477-W487.

Contribution summary

Paper I. I co-designed the study, contributed to the machine learning model construction for gene essentiality, conducted part of the evolution analysis, analyzed the data and wrote the paper.

Paper II. I co-designed the study, developed the deep learning model, performed model-related analysis, analyzed the data and wrote the paper.

Paper III. I conducted the feature importance analysis on protein production, co-analyzed the data and co-wrote that part of the paper.

Paper IV. I conceived part of the research, designed the study, constructed the horizontal gene transfer software, analyzed the data and wrote the paper.

Preface

This dissertation serves as partial fulfillment of the requirements to obtain the degree of Doctor of Philosophy at the Department of Life Sciences at Chalmers University of Technology. The PhD studies were carried out between September 2019 and May 2023 at the division of Systems and Synthetic Biology (SysBio) under the supervision of Eduard J Kerkhoven, and co-supervised by Martin Engqvist and Aleksej Zelezniak. The thesis was examined by Ivan Mijakovic. It was mainly funded by the Novo Nordisk Foundation.

Le Yuan
May 2023

Table of Contents

Abstract ..... *iii*

List of Publications ..... *v*

Contribution summary ..... *vi*

Preface ..... *vii*

Abbreviations ..... *x*

Acknowledgements ..... *xi*

1. Background *1*

1.1 Synthetic biology and metabolic engineering **1**

1.2 Systems biology **2**

1.3 Yeast species **3**

1.4 Big data **4**

1.5 Machine learning techniques **6**

1.5.1 Classical machine learning algorithms 8

1.5.2 Deep learning algorithms 8

1.6 Comparative genomics and evolution **10**

1.7 Aims and significance **12**

2. Development and applications of ML and DL approaches *14*

2.1 ML on gene essentiality (Paper I) **14**

2.1.1 Software used in the pipeline 14

2.1.2 Data collection 14

2.1.3 ML workflow for the prediction of essential genes 15

2.1.4 Evaluation of the ML model performance 17

2.2 DL on enzyme turnover number (Paper II) **19**

2.2.1 Data preparation for the DL model 19

2.2.2 Construction of the DL model pipeline 21

2.2.3 DL model performance for k_{cat} prediction 22

2.2.4 Prediction of k_{cat} values for mutated enzymes 25

2.2.5 Interpretation of the DL model 26

2.2.6 Biological insights gained with the aid of the DL model 27

2.3 ML on protein production (Paper III) **30**

2.3.1 Simulation of the production of recombinant proteins in yeast 30

2.3.2 ML for feature importance analysis towards protein production 31

3. Development and applications of comparative genomics tools on yeasts *33*

3.1 HGTphyloDetect - Detection of horizontal gene transfer (Paper IV)..... 33
3.1.1 Detecting horizontal gene transfer from phylogenetically distant organisms..... 34
3.1.2 Detecting horizontal gene transfer from closely related organisms..... 35
3.1.3 Basic usage and applications of the HGTphyloDetect toolbox..... 36
3.1.4 Testing the performance of the HGTphyloDetect toolbox..... 36
3.1.5 Comparison with other existing approaches for HGT detection..... 38
3.1.6 Phylogenetic analysis via HGTphyloDetect..... 39

3.2 Substrate utilization analysis on large-scale yeast species (Paper I)..... 41
3.2.1 Gain of new traits in substrate utilization occurring in yeast species..... 41
3.2.2 Gene family expansion and contraction analysis..... 41
3.2.3 Evolutionary mechanisms underlying the trait diversity in substrate utilization..... 42

***4. Conclusions*..... 45**

***5. Future perspectives*..... 47**

***References*..... 49**

Abbreviations

PHAs	Polyhydroxyalkanoates
GRAS	Generally recognized as safe
NCBI	National Center for Biotechnology Information
PDB	Protein Data Bank
ML	Machine learning
DL	Deep learning
AI	Artificial intelligence
RF	Random forest
SVM	Support vector machine
KNN	K-nearest neighbors
CNN	Convolutional Neural Network
GNN	Graph Neural Network
GAN	Generative Adversarial Network
HGT	Horizontal gene transfer
PTMs	Post-translational modifications
GEM	Genome-scale metabolic model
DNC	Di-Nucleotide Composition
OG	Ortholog group
JSD	Jensen-Shannon divergence
ROC	Receiver operating characteristic
TPR	True positive rate
FPR	False positive rate
AUC	Area under the ROC curve
API	Application programming interfaces
SMILES	Simplified molecular input line entry system
CE	Carbohydrate and energy
AFN	Amino acids, fatty acids, and nucleotides
PNP	Purine nucleoside phosphorylase
ecGEMs	Enzyme-constrained genome-scale metabolic models
RMSE	Root mean square error
ER	Endoplasmic reticulum
SHAP	SHapley Additive exPlanations
nr	Non-redundant
MCC	Matthews correlation coefficient
MSA	Multiple sequence alignment
BYCA	Budding yeast common ancestor

Acknowledgements

PhD is a super important turning point in my life, and it has been an amazing experience studying at SysBio. I believe, as for me, even if 50 years later, I will not regret getting on the plane to Gothenburg. And I will be forever grateful for the decision I made myself in 2019 to be a PhD student at Chalmers.

First of all, an immense thank you must go to my supervisor, Eduard, for guiding me throughout my exciting PhD projects. Thanks for your invaluable support over the years and giving me so much freedom to do various projects. Thanks for your insightful comments and thorough edits in all my manuscripts. Also, I would like to thank Jens for coming to my previous university and providing me with the opportunity to be one part of the SysBio family. Thanks to my two co-supervisors, Martin and Aleksej, for all your help and kindness. Thanks to my examiner Ivan for always giving me signatures with a nice smile. Thanks to Verena and Yun for your constructive input and comments during the MSB subgroup meeting.

The work in my thesis would not have been possible without the help of collaborators. Here, I would like to express my sincere thanks to Feiran, thank you for inviting me to join various interesting projects. It is your kind help and support that have made my life and study at Chalmers a wonderful time. I would also like to extend my special thanks to Hongzhong, your collaborative and diligent spirits will be the most important thing that I should learn forever. Gang, thanks for teaching me a lot of knowledge about machine learning at the beginning of my PhD. Yu, thank you for sharing your critical suggestions during our subgroup meeting. Iván, thanks for our interesting talk every day and your company as an office mate, I enjoy the hotpot time with you very much. Cheewin, it is very nice to be a teaching assistant together with you at the systems biology master course. Thank you, Peishun, Xiang, Angelo, Andrea, Jiwei, Oliver, Marta, Simone, Xin, Lei, Yanyan Wang, Yanyan Chen, Xiaowei, Albert, Fariba, Mihail, Demi, Maximilian, Dany, Jing, Xiaozhi, Yihan, Jerry, Jian Zhang, Juan, Lingyun, Xiaofan, Hao Luo, Hao Wang, for our face-to-face talk every time we met at Chalmers. Thank you, Martina, Gunilla, Erica, Anne-Lise, for your great support and administration over the years. Thanks to the core value team at SysBio for organizing so many colorful activities. Thanks to everyone that I met at SysBio, thank you all for helping me to write an unforgettable story in my life.

Thanks of course to my family for their long-term support and encouragement. I am deeply grateful to my mother and father for their unwavering support and belief in me. Thanks to my two elder sisters Ping and Li for our enjoyable and pleasant conversations. Finally, I would like to express my gratitude to my girlfriend, Yanfang. Without her tremendous encouragement and understanding in the past few years, it would be impossible for me to complete my PhD study in Sweden.

1. Background

1.1 Synthetic biology and metabolic engineering

Synthetic biology is an interdisciplinary field that focuses on the design and development of living organisms and living systems. It integrates principles from computer science, biology, chemistry, material science and engineering. The aim is to create biological parts, devices as well as biological systems to empower current biotechnology and drive innovation [1, 2]. Metabolic engineering is a related field that involves the manipulation of cellular regulatory, enzymatic and transport processes to increase the yield of specific products or enable the production of new products [3, 4]. By optimizing industrial fermentation processes, metabolic engineering has become a powerful tool for producing high-value commodities, pharmaceuticals, and bulk chemicals [5, 6]. Together, synthetic biology and metabolic engineering offer promising approaches for developing novel biological technologies and products.

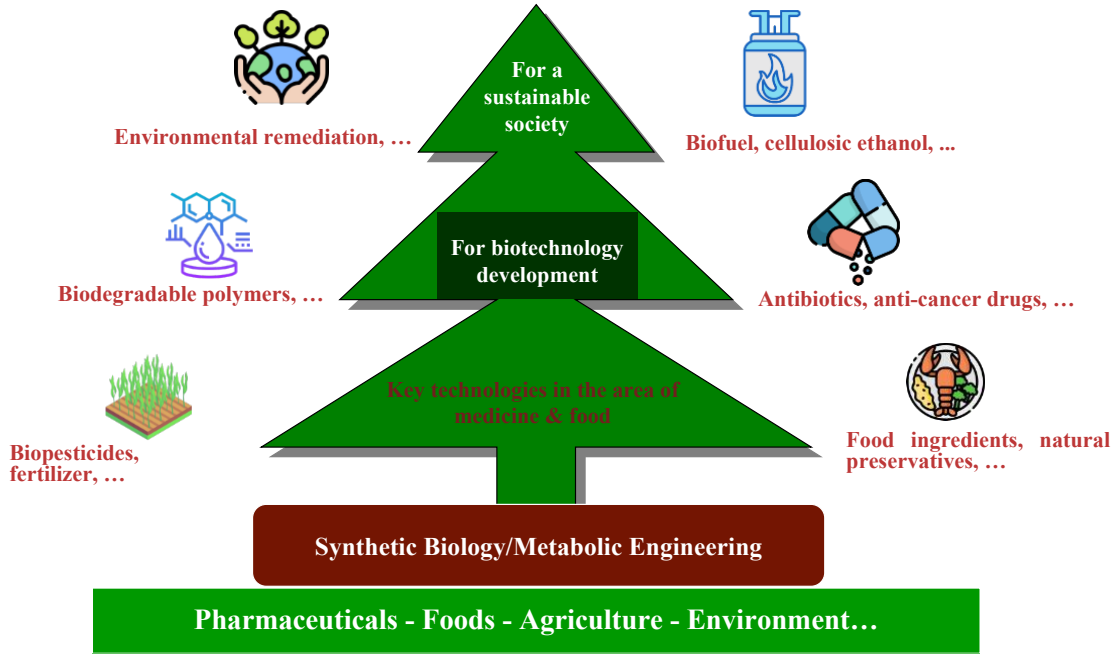

Figure 1 Overview of synthetic biology and metabolic engineering. Synthetic biology and metabolic engineering have been widely applied into various fields, such as pharmaceuticals, foods, agriculture and environment. They provide great potential for biotechnology development and life science research, which could be very useful for the advancement of a sustainable society.

Synthetic biology and metabolic engineering show great promise in addressing global challenges related to sustainable biomanufacturing and renewable energy by developing biotechnological solutions [2, 4]. Synthetic biology and metabolic engineering have been very successful in various domains [7], e.g., food science, drug development, chemical engineering, agriculture, material science and environmental remediation (**Figure 1**). In food science, antioxidant food ingredients and natural preservatives have been able to be produced with the advent of microbial cell factories [8, 9]. This approach based on

microorganisms provides a more viable alternative to previous methods based on chemical synthesis and solvent extraction, which are often time-consuming, expensive, and environmentally harmful. In drug development, synthetic biologists have harnessed microbes to synthesize complex natural products of pharmaceutical interest, e.g., immunosuppressants, antibiotics and anti-cancer drugs [4, 10]. In chemical engineering, synthetic biology and metabolic engineering have been instrumental in the production of biofuels and cellulosic ethanol, which could potentially displace fossil fuels used in heavy vehicles that result in significant carbon emissions [11-14]. In agriculture, plants have been widely used as if they were bioreactors, to produce essential oils and volatile organic compounds with properties as insecticides, fungicides and bactericides [15]. Recent development of genome engineering tools and genome-wide functional genomics can improve the ability to engineer microbes for biofertilization, as well as enhanced crop productivity [16]. In material science, microorganisms have been utilized for industrial scale production of biodegradable polymers, such as polyhydroxyalkanoates (PHAs) [17-19]. In environmental remediation, synthetic biology could contribute by reducing the massive use of harmful industrial chemicals through providing biologically-based alternatives [20].

While synthetic biology and metabolic engineering have proven successful in various domains, their applications could be further enhanced by the holistic perspective that systems biology provides [21]. The quantitative analysis and modeling techniques employed in systems biology can enable the optimization of metabolic pathways and the identification of novel targets for metabolic engineering. Systems biology can thus offer a valuable framework for synthetic biology and metabolic engineering, providing insights into the behavior of biological systems and guiding the design and optimization of synthetic biological systems.

1.2 Systems biology

What is systems biology? Systems biology is another interdisciplinary field that seeks to understand biological systems as a whole, rather than only investigating individual components [22]. In systems biology, there mainly exist two different approaches: 1) top-down systems biology generally relies on omics data and integration of these datasets with mathematical models to elucidate the cellular functions; and 2) bottom-up systems biology involves the formulation of mathematical models based on accumulated knowledge [23]. In most cases, the goal of systems biology is to create a quantitative description for the biological systems, which is usually achieved through the development of mathematical models. The ultimate aim of systems biology is to gain a deeper understanding of the biological systems and use the models to make predictions on how the system will behave under different conditions. Besides, systems biology can be particular useful for investigating cellular metabolic networks, which can help to explore how the phenotype is generated from the genotype and how evolution has crafted the phenotype [24].

In recent years, advances in high-throughput analytical methodologies have enabled the comprehensive analysis of cellular processes in organisms, providing a wealth of data that can be analyzed using systems biology approaches [25]. These advances are very valuable for the top-down systems biology, as they allow for the study of complex biological systems in detail that was previously impossible. The high-throughput techniques can generate large amounts of data, including genomics, transcriptomics, proteomics, and metabolomics data. These data can be further analyzed using computational and mathematical models, which is particularly supportive of the top-down systems biology. In the bottom-up systems biology, the system is reconstructed based on existing biological knowledge, with the aim of combining individual models into a holistic model describing the biological systems as a whole [21]. By taking the bottom-up approach, researchers can gain a deeper understanding of emergent properties, which are behaviors of the system that cannot be understood by examining the individual parts in isolation, but only become apparent when the whole system is functioning together.

1.3 Yeast species

For the above-mentioned synthetic biology and metabolic engineering (section 1.1) and systems biology (section 1.2), yeasts are widely used as model organisms in these fields. Yeasts are eukaryotic, unicellular microorganisms that are classified as members of the fungus kingdom [26]. One of the typical and well-known yeast species is the bakers' yeast, *Saccharomyces cerevisiae*. It has become an indispensable model system for understanding eukaryotic biology at the cellular, molecular and genomic levels [27]. *S. cerevisiae* is a model organism for eukaryotic cells, which is particularly useful for studying biological process in eukaryotes. The insights gained from studying *S. cerevisiae* can be applied to other eukaryotes, including human. For instance, *S. cerevisiae* has been used as a model organism to investigate the cell cycle and human diseases such as Parkinson's and Alzheimer's [28, 29].

Moreover, *S. cerevisiae* has been widely used in industry because it is a generally recognized as safe (GRAS) organism, which makes this species very suitable for large-scale production of specific products [30]. As a model organism, the yeast *S. cerevisiae* has been widely applied in traditionally wine, bread and beer making. More recently, this yeast has also served as a cell factory for producing various bulk chemicals, fuels and pharmaceuticals through metabolic engineering [31]. Meanwhile, several non-conventional yeast species, including *Yarrowia lipolytica*, *Pichia pastoris* (*Komagataella phaffii*) and *Hansenula polymorpha*, have recently gained more interest as microbial hosts to produce recombinant proteins and various value-added natural products due to their specific physiological properties [32, 33].

With the rapid development of high-throughput sequencing technologies, researchers have been able to deeply sequence 1,011 natural *S. cerevisiae* isolates from a broad array of human-associated biotopes [34]. Furthermore, the whole genomes of 332 different yeast

species have already been sequenced and are now publicly available, covering nearly one third of all known budding yeasts [35]. Through comparative genomic analysis of different yeast strains and species, researchers can delve deeper into the genetic factors that contribute to specific traits or phenotypes, e.g., stress tolerance, pathogenicity and fermentation efficiency [36, 37]. Additionally, the availability of genomes from a wide range of yeast species can aid in understanding yeast biodiversity, evolution and adaptation [38]. These valuable resources provide a unique opportunity to gain novel insights into genotype-phenotype relationships in eukaryotic biology.

1.4 Big data

In the current era of data explosion, big data has played a significant role in both systems and synthetic biology [39]. Not only does experimental data continue to accumulate, but the rate of data generation is even rapidly increasing. In an attempt to leverage the knowledge contained in this data for synthetic biology, a plethora of databases have been developed that are being applied to various levels of biosynthesis research. These levels include molecules, reactions, pathways, and enzymes. The development and application of these open-source databases have significantly accelerated synthetic biology and is expected to continue to do so as more experimental data are generated and integrated into these resources.

At the molecular level, several integrated databases are available to provide valuable information on chemical structures, molecular properties, and biological activities. PubChem is one of the most widely used databases maintained by the National Center for Biotechnology Information (NCBI) [40]. It is a comprehensive resource for chemical information and has amassed data from over 750 sources, containing more than 111 million chemical structures, more than 303 million biological activity data points and over 37 million scientific publications [40]. As a result of its extensive coverage, PubChem has served as an indispensable chemical information platform in many fields, such as cheminformatics, synthetic biology and chemical biology. ChEBI, the Chemical Entities of Biological Interest, is a freely available database and ontology that includes over 46,000 manually curated entries mainly focusing on small chemical compounds [41]. Each entry is classified within the ontology and annotated with multiple relevant information, such as chemical structures, chemical synonyms, database cross-references and literature citations. DrugBank is a unique web resource that provides comprehensive information about drugs approved by the Food and Drug Administration (FDA), as well as experimental drugs undergoing FDA certification [42]. This database contains detailed information about drugs, including their mechanisms of action, drug-drug interactions, and drug-target interactions. In addition, DrugBank provides valuable information on drug metabolism, pharmacology, and pharmaceutical formulation.

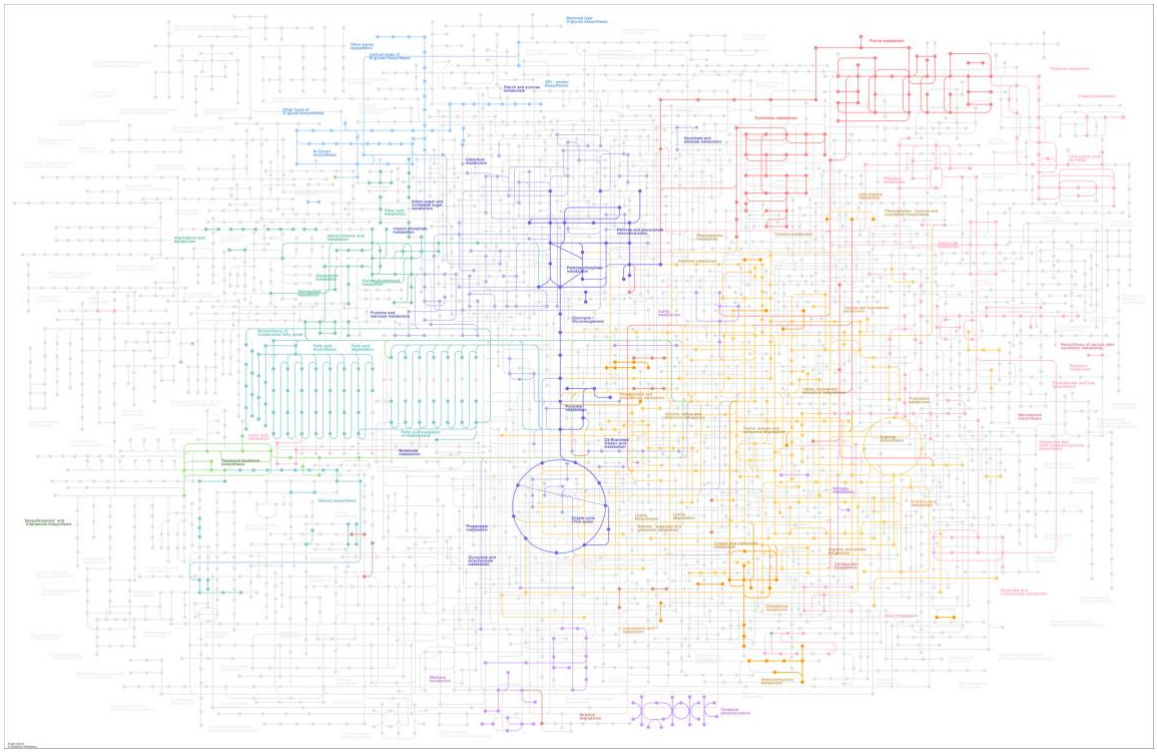

Figure 2 A global map of KEGG metabolic pathways for *S. cerevisiae*. KEGG is a database that contains a collection of manually drawn graphical diagrams for various pathways. The source of this comprehensive map for *S. cerevisiae*: https://www.genome.jp/pathway/sce01100.

At the reaction and pathway level, there are also various open-access databases that can be used to obtain information on biochemical reactions and pathways. One such database is KEGG, which contains manually curated pathway maps, representing molecular interaction and reaction networks for various organisms [43]. This comprehensive database provides a global view of different biological processes, e.g., metabolism, genetic information processing, signaling, cellular processes and environmental information processing. A specific example for *S. cerevisiae* is shown in **Figure 2**. Thus, KEGG is a valuable resource in many fields, including bioinformatics, systems biology and synthetic biology. Another knowledgebase of biosynthetic reactions is Rhea, where the reaction data is carefully curated from the scientific literature by expert biochemists, with support from natural language processing tools [44]. Additionally, Rhea provides programmatic access to all data, queries and tools available through the Rhea website via RESTful URLs. The current release of Rhea contains 15,572 reactions with 13,038 unique chemical compounds from 17,313 unique references, making it a valuable resource for researchers interested in studying biosynthetic reactions and pathways. MetaCyc is another highly curated database of metabolism that provides a comprehensive and integrated view of metabolic pathways from all domains of life [45]. The database contains information about chemical compounds, reactions, enzymes and metabolic pathways that have been experimentally validated and reported in the scientific literature. MetaCyc is constantly being updated and currently contains 3,105 pathways, 18,566 reactions and 18,973 metabolites. The data in MetaCyc is manually curated by a team of experts to ensure its accuracy and completeness.

At the enzyme level, Uniprot, Protein Data Bank (PDB), BRENDA and Sabio-RK are some of the valuable resources that provide information on enzyme function, structure and kinetics. Uniprot is dedicated to providing the scientific community with a freely available, high-quality dataset of protein sequences annotated with detailed functional information [46]. The Uniprot database, UniprotKB, includes both reviewed entries and unreviewed entries. The reviewed UniProtKB/Swiss-Prot entries contain data added by their expert biocuration team, while the unreviewed UniProtKB/TrEMBL entries are annotated by automated systems. PDB is a comprehensive repository established in 1971 that stores a vast amount of structure data of proteins, nucleic acids, and complex assemblies [47]. The extensive collection of protein structures in PDB allows users to analyze structures of interest, identify structural motifs, and design experiments to explore protein function. It has become an essential resource enabling the development of education and research in biomedicine, biochemistry, and fundamental biology. BRENDA is an exceptional database that offers a comprehensive view of enzymes and their ligand interactions [48]. The information in BRENDA encompasses details on enzyme names, structures, stability, localization, specific activity, and kinetics parameters, among other features. Through ongoing curation and updates of classified enzymes, BRENDA currently holds over 5 million data entries for approximately 90,000 enzymes from around 13,000 organisms, which is manually extracted from roughly 157,000 primary literature references, with the aid of text mining. Sabio-RK is a manually curated database that focuses primarily on enzyme kinetics [49]. It contains approximately 57,000 data entries, including around 42,000 entries for wildtype enzymes and 13,000 entries for mutant enzymes. Each data entry in Sabio-RK provides enzyme kinetics data for a single enzymatic reaction in one organism under specific environmental conditions.

1.5 Machine learning techniques

With the emergence of big data at various levels (as shown in the above section), coupled with the rapid advancements in cloud computing, machine learning (ML), deep learning (DL) and artificial intelligence (AI), there is immense potential for these technologies to be adopted to better understand various yeast species and significantly promote the development of yeast systems biology. ML is one such technique that has gained popularity in recent years, in part due to the rise of big data. In face of big data, ML has found a broad range of applications, including but not limited to e-commerce and banking, transportation, medicine discovery, and beyond (**Figure 3**).

ML, DL and AI have become some of the most popular and widely discussed technologies in the world today. In particular, ChatGPT, an AI chatbot developed by OpenAI, has recently captured considerable public attention across various domains [50]. In terms of their relationship, DL is a subfield of ML, and ML is a subfield of AI. DL involves advanced algorithms that mainly develop artificial neural networks (similar to neurons present in the human brain) to automatically learn and extract features from data, as well

as create models and predictions accordingly. ML is a broader field that encompasses a range of algorithms that enable computers to learn from data without being explicitly programmed. AI refers to the broad field of computer science that includes not only ML and DL, but also robotics, natural language processing, expert systems and computer vision, among others.

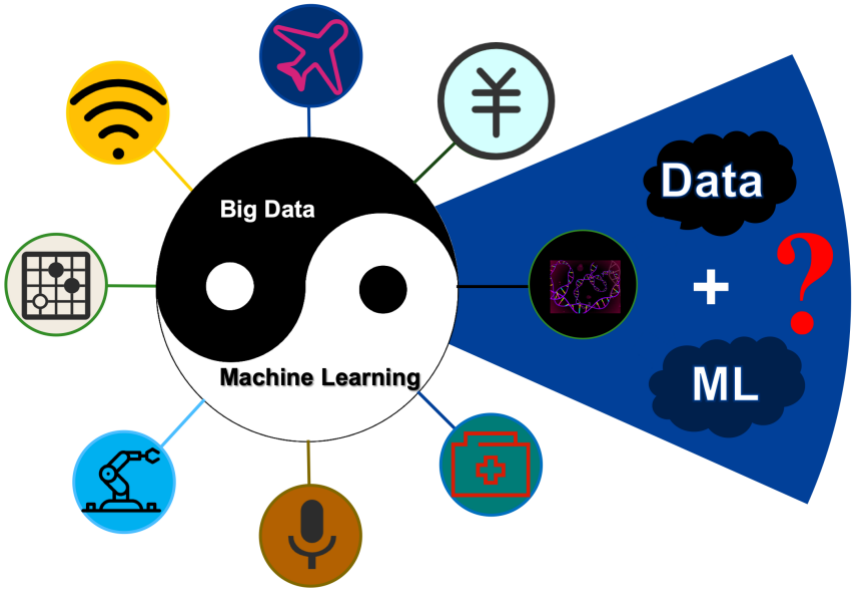

Figure 3 The combination and applications of big data and machine learning. In the area of big data, machine learning has been widely applied to various fields, e.g., e-commerce and banking, airlines, communication systems, manufacturing, medical science, and more.

ML is primarily categorized into three types: supervised learning, unsupervised learning, and reinforcement learning [51]. Supervised learning is one of the fundamental types of machine learning, in which the algorithms are trained based on labelled data [52]. There are two basic types of supervised learning: classification and regression algorithms. The main difference between both is that classification algorithms are used to classify different classes or labels, such as high or low, true or false, while regression algorithms are used to predict continuous numerical values. In contrast, unsupervised learning involves training models on data that has not been labeled. This type of learning is particularly useful for exploratory data analysis, as it helps to explore the underlying trends and patterns from raw data or cluster similar data into a specific number of groups [53]. Reinforcement learning is another type of ML, where an agent interacts with an environment by sensing its state and learns to take actions that maximize long-term reward [54]. It is goal-oriented, and the agent aims to learn sequences of actions by exploitation in an uncertain environment to maximize future rewards. Unlike supervised learning, reinforcement learning does not require labeled data because it learns by interacting with the environment. Reinforcement learning has numerous practical applications, such as self-driving cars, robotics, and adaptive controls.

1.5.1 Classical machine learning algorithms

Random forest (RF) is a versatile and powerful ML algorithm that is commonly utilized for both classification and regression problems [55]. It is an ensemble learning approach that works by creating multiple decision trees and combining their outputs to generate a final prediction. Each decision tree in the RF model is built using a random subset of the training dataset. This kind of randomness can help to prevent overfitting and enhance the model's precision. For classification tasks, the RF algorithm selects the class that is predicted by the majority of the decision trees. For regression tasks, the algorithm generates the mean prediction of all the trees as the output. RF has numerous advantages over other classical ML algorithms. It is user-friendly and does not require extensive hyper-parameter tuning. Moreover, the RF algorithm can provide estimates of feature importance, allowing for easier interpretation and understanding of the ML model's predictions.

Support vector machine (SVM) is another prevalent supervised learning algorithm, which can also be used for classification as well as regression problems. SVM was originally developed for classification, and its basic idea is to find the best hyperplane or boundary that separates data points based on predefined classes or labels [56]. SVM is not only effective for linear datasets, but it can also work with non-linear datasets using kernel functions. Compared to other ML algorithms, SVM is particularly effective in high-dimensional spaces, and it can handle complex datasets with many features. SVM has been validated its usefulness in a wide variety of biological applications, such as the prediction of chronic kidney disease [57], protein fold recognition [58], and the identification of anticancer peptides [59].

K-nearest neighbors (KNN) is a frequently employed and straightforward ML algorithm that can be used for both classification and regression problems, but it is more commonly used for classification [60]. The KNN classifier operates by identifying the k nearest neighbors to a given data point and then using the majority vote of their labels to classify the data point. In the case of regression, the algorithm calculates the average of the labels and returns it as the output value. Since KNN does not make any assumptions about the underlying data, it is considered as a non-parametric learning algorithm. For the KNN algorithm, selecting an appropriate value of k is crucial for preventing both underfitting and overfitting of the ML model. Cross-validation can be used to determine the optimal value of k for the KNN algorithm, which can improve its performance.

1.5.2 Deep learning algorithms

DL is a type of advanced ML that utilizes neural networks to perform intricate computations and predictions based on large amounts of data [61]. Similar to the human brain, neural networks consist of artificial neurons or nodes that are arranged in three layers, i.e., input layer, one or more hidden layers and output layer. Here is how a neural network operates: (i) Data feeds input information to each node. (ii) After multiplying the inputs

with random weights and adding bias, the node computes the output. (iii) Nonlinear functions, also as activation functions, are mathematical functions that are applied to the output of each node or neuron in a neural network. These functions introduce nonlinearity to the network, allowing it to model and learn complex relationships between inputs and outputs. Without these nonlinear functions, the neural network would be limited to only modeling linear relationships. Consequently, DL can automatically extract features or representations from data without relying on data pre-processing. DL encompasses various architectures that are widely used in different domains, such as Convolutional Neural Network (CNN), Graph Neural Network (GNN), Generative Adversarial Network (GAN), and more.

CNN is a typical DL algorithm that consists of three essential building blocks: convolutional layers, pooling layers, and fully connected layers [62]. The convolutional layer is the core building block of a CNN that extracts basic features from the input dataset. This layer uses a set of filters or kernels to perform convolution operations on the input and produce a new output feature map. The size and number of filters can be adjusted depending on the complexity of the input data. The pooling layer, also known as a downsampling method, helps to reduce the number of parameters and increase the efficiency of the network. Similar to the convolutional layer, the pooling operation sweeps a filter across the entire input. There are two main types of pooling, i.e., max pooling and average pooling, which help to reduce complexity and improve the network's ability to generalize to new data. Finally, the output of the pooling layer is fed into one or more fully connected layers, performing the specific task based on the features extracted from previous layers and their different filters. Each neuron in the fully connected layers is connected to every neuron in the previous layer, allowing the network to learn complex non-linear relationships between the input data and the target output.

GNN is a powerful algorithm for processing and analyzing structured data that can be represented in the form of graphs [63]. In a GNN, a graph is a data structure consisting of two main components: nodes and edges. The nodes and edges of a graph are transitioned to vectors or matrices, which are used as input to a neural network. The main idea of GNN is based on message passing, where each node in the graph sends messages to its neighboring nodes and updates its own state according to the received messages. One of the advantages of this DL architecture is that it allows researchers to work directly on natural input representations of compounds or materials, which are chemical graphs with nodes described as atoms and edges described as chemical bonds. Thus, GNN can learn complex representations of chemicals that are very useful for specific tasks, such as predicting molecular properties [64], de novo drug design [65], and protein-protein interaction prediction [66].

GAN was introduced in 2014 by lan Goodfellow and his colleges as a framework for unsupervised learning tasks in DL [67]. GAN has become an increasingly popular framework in recent years due to its ability to generate high-quality and realistic data in

various areas. The GAN architecture consists of two neural network models - a generator and a discriminator. The generator model is typically designed to learn how to generate synthetic data that closely resembles real data, while the discriminator model is tasked with distinguishing between the generator's synthetic data and real data. This architecture is adversarial because the generator and the discriminator work against each other with opposing objectives – the generator tries to mimic reality while the discriminator tries to identify fake data. These two models are trained simultaneously, and they improve their capabilities over time through an iterative process.

1.6 Comparative genomics and evolution

As described in the above section, ML has shown great potential in various areas. However, one of the challenges of ML is its weakness in interpretation, especially in investigating complex biological problems, such as evolution. ML models are often referred to as black-box models, meaning that their decision-making process is difficult to interpret and understand. This is where comparative genomics can be especially valuable. Comparative genomics is a field of biological study that aims to compare the genomic characteristics of different species or individuals to identify similarities and differences [68]. By analyzing these similarities and differences, comparative genomics and its related evolutionary analyses contribute to our understanding of how new species or new traits emerge and shed light on various biological mechanisms [68]. One essential aspect of study in comparative genomics is the analysis of evolutionary changes that occur in the genome over time. Three common areas of study within this field that are also topics addressed in this thesis include dN/dS analysis, horizontal gene transfer (HGT), and gene family expansion and contraction. These approaches allow us to identify functional changes in genes, as well as to explore the underlying mechanisms that contribute to genetic diversity and evolution.

The dN/dS ratio is a common metric used in comparative genomics to quantify selection pressures acting on protein-coding regions. It represents the ratio of non-synonymous (dN) to synonymous (dS) substitutions. Non-synonymous mutations result in changes to the amino acid sequence of a protein, while synonymous mutations do not [69]. Therefore, the dN/dS ratio can provide insight into whether a gene has undergone positive selection, neutral selection, or purifying selection. A dN/dS ratio greater than one suggests positive selection, meaning that non-synonymous mutations that change the protein sequence are being favored. Conversely, a ratio value less than one indicates purifying selection, where synonymous mutations are being favored to maintain the protein's function [70].

HGT is the process by which genetic material moves between different species across the tree of life, beyond the transmission of DNA from parent to offspring [71]. HGT is of great interest because it can drive functional innovation by introducing novel genes or pathways [72]. Moreover, it has been recognized as a significant contributor to niche specification, disease emergence and the shift in metabolic capabilities [35, 73, 74]. Transformation is a

crucial mechanism for HGT, which involves the active uptake and integration of extracellular naked DNA that can be inherited [75]. Prokaryotes, in particular, exhibit a high frequency of HGT events, representing one of the primary mechanisms that drive genetic variation and microbial evolution in these organisms [76]. Although HGT occurs less frequently in eukaryotes than in prokaryotes, it still plays a vital role in the evolution of eukaryotic genomes by enabling the acquisition of adaptive functions [72].

Gene family expansion and contraction are dynamic processes that involve changes in the number of genes within a family over time. A gene family is a group of homologous genes that are likely to have highly similar functions [77]. Gene family expansion can occur through gene duplication events, which result in the creation of new genes that are similar in sequence and function to existing ones [78]. In contrast, gene family contraction may occur due to gene loss or deletion events, resulting in a reduction in the number of genes within the family [77]. The expansion or contraction of gene families can have significant implications for species differentiation, phenotypic diversification, and adaptation to environmental changes [79]. For instance, gene family expansion can lead to the emergence of new functions or adaptations, while gene family contraction may result in the loss of important traits.

1.7 Aims and significance

Yeasts, including *Saccharomyces cerevisiae*, *Yarrowia lipolytica*, *Pichia pastoris*, etc., have emerged as promising microbial cell factories due to the advanced synthetic biology and metabolic engineering technologies. However, the cellular metabolism and physiological properties of most yeast species remain poorly understood. This thesis aims to address the knowledge gap by utilizing large amounts of data and applying state-of-the-art ML techniques and comparative genomic analysis to gain a deeper understanding of yeast traits and metabolism.

ML is a powerful tool with wide applicability in the prediction of various biology-related problems, e.g., gene expression, EC numbers and enzyme catalytic temperature optima [80-83]. In this thesis, I harnessed the power of advanced ML techniques to drive the development of yeast systems biology. In **Paper I**, I utilized two ML algorithms that leveraged both sequence features and evolution-based features to predict essential genes. Notably, the inclusion of evolutionary features led to a marked improvement in the accuracy of gene essentiality prediction. Then I used an SVM pipeline to annotate essential genes for large-scale yeast/fungi species, providing a valuable resource for the yeast community. In **Paper II**, I developed a high-quality deep learning model called DLKcat, which predicted k_{cat} values by combining a graph neural network for substrates and a convolutional neural network for proteins. Through the prediction of k_{cat} profiles for large-scale yeast/fungi species, enzyme-constrained models were reconstructed, allowing for a more comprehensive exploration of cellular metabolism on a large scale. In **Paper III**, an investigation was conducted on the impact of different features on protein production using a RF algorithm, the findings revealed that post-translational modifications (PTMs) have a higher influence on protein production in comparison to amino acid composition.

Comparative genomics plays an important role in understanding evolutionary relationships and identifying genetic changes that occur between different species. In **Paper IV**, I developed the HGTphyloDetect toolbox, a comprehensive tool for the identification of HGT events, regardless of whether the acquired genes are from distantly related species or closely related species, highlighting its versatility. Using case studies on several yeast species, HGTphyloDetect was shown to accurately identify horizontally acquired genes. More importantly, the HGTphyloDetect toolbox facilitates the generation of high-quality phylogenetic trees, which can aid in the navigation of potential donors and elucidate feasible paths of gene transmission in detail. In addition, through systematic evolution analysis (e.g., HGT, gene family expansion and contraction) and genome-scale metabolic model (GEM) simulation in **Paper I**, I probed the underlying mechanisms for substrate utilization across large-scale yeast species. revealing that gene family expansion and enzyme promiscuity are prominent mechanisms for metabolic trait gains.

Overall, my thesis leveraged advanced ML and DL techniques to shed light on various aspects of yeast systems biology, spanning gene essentiality, enzyme kinetics, and protein

production. Furthermore, the development of the comprehensive HGTphyloDetect toolbox enabled the identification of HGT events and the construction of high-quality phylogenetic trees. Moreover, the investigation of substrate utilization mechanisms through systematic evolution analysis and GEM simulation provided insights into the genetic and biochemical factors underlying metabolic trait gains in large-scale yeast species.

These findings of this thesis have significant implications for the broader fields of synthetic biology and evolutionary biology, providing valuable resources and knowledge for the yeast community. The insights gained through these analyses can further lead to the development of novel therapeutic and biotechnological applications. The successful integration of advanced ML and DL techniques with comparative genomics and metabolic modeling approaches has opened up new avenues for exploring the intricate complexities of biological systems and has the potential to transform our understanding of various biological processes.

2. Development and applications of ML and DL approaches

As outlined in the background section, ML and DL techniques have enormous potential to enhance our understanding of a wide range of biological problems when combined with big data. In this chapter, I present three studies that demonstrate the applications of ML and DL methods in advancing systems biology of yeast. The first study (**Paper I**) investigates the role of ML in the prediction of gene essentiality. The second study (**Paper II**) utilizes deep neural networks to predict enzyme turnover number and applies the DL model to large-scale yeast species. Finally, the third study (**Paper III**) explores the impact of different features on protein production, both positively and negatively.

2.1 ML on gene essentiality (Paper I)

Essential genes are those genes that are necessary for the survival of an organism [84]. Identifying essential genes is important for discovering new drug targets, exploring disease genes, and understanding the minimal requirements of an organism [85]. Although it has been shown that high-throughput experimental methods can be applied to identify gene essentiality, particularly in organisms like *S. cerevisiae* [86], they can be prohibitively expensive, time-consuming, and labor-intensive. Given the availability of large-scale annotation of gene essentiality data for certain yeast organisms, I developed ML approaches to predict essential genes by incorporating sequence features and evolution-based features.

2.1.1 Software used in the pipeline

Various ML methods were employed to predict essential genes in a computational way that integrated several open-source software packages. The NumPy version 1.17.2 (https://numpy.org/) and SciPy version 1.3.1 (https://www.scipy.org/) packages were utilized for handling data arrays. The data visualizations were generated using the matplotlib version 3.1.2 (https://matplotlib.org/) and seaborn version 0.9.0 (https://seaborn.pydata.org/) packages. The ML algorithms were implemented using the scikit-learn version 0.22.1 (https://scikit-learn.org/stable/) library, which is based on the Python programming language.

2.1.2 Data collection

To develop the ML models for predicting essential genes, I collected datasets of reported essential genes for five yeast/fungi species (**Table 1**), namely *S. cerevisiae*, *Y. lipolytica*, *P. pastoris*, *Schizosaccharomyces pombe*, and *Candida albicans*. Gene and protein sequence FASTA files for *S. cerevisiae*, *C. albicans*, and *S. pombe* were sourced from the SGD database [87], the CGD database [88], and the PomBase database [89], respectively. Furthermore, the gene and protein sequence data for *P. pastoris* and *Y. lipolytica* were retrieved from the NCBI RefSeq database [90].

Table 1. Essential gene data collected from literature reports.

Organism	Essential genes	Non-essential genes	Reference
S. cerevisiae	1037	4543	Chen, et al. 2012 [91]
S. pombe	1346	3689	Chen, et al. 2012 [91]
C. albicans	633	1714	O'Meara, et al. 2015 [92]
P. pastoris	144	465	Cankorur-Cetinkaya, et al. 2017 [93]
Y. lipolytica	108	534	Wei, et al. 2017 [94]

2.1.3 ML workflow for the prediction of essential genes

Gene essentiality can be predicted by ML based on sequence-derived properties [95]. Upon obtaining a high-quality dataset for gene essentiality, the dataset was randomly divided into a training dataset and a test dataset at a ratio of 80:20, respectively. The gene sequences were then utilized to calculate sequence features such as Di-Nucleotide Composition (DNC) and codon frequency represented by Kmer. They can be calculated by the following mathematical formulas:

$$DNC(r,s) = \frac{N_{rs}}{N-1} \; r,s \; \in \{A,C,G,T\} \; (1)$$

$$Kmer(t) = \frac{N\,(t)}{N} \; t \; \in \{AAA, AAC, AAG, \dots, TTT\} \; (2)$$

In these calculations, N_{rs} represents the number of combinations of any two nucleic acid r and s, N(t) is the number of type t, and N is the length of the nucleotide sequence.

In addition to sequence features, evolutionary information has been hypothesized to improve gene essentiality prediction by leveraging the possible fact that essential genes are more conserved than non-essential genes. This is because mutations in essential genes can be detrimental to the organism's survival, leading to stronger purifying selection pressure [96]. To incorporate evolutionary information, several features were calculated for each gene based on its ortholog information, including protein conservation score, dN/dS, number of gene occurrence across species, and average paralog number (**Figure 4A**). Since each gene could be mapped with an ortholog group (OG) across multiple species, it was possible to calculate these evolution-based features for each gene.

In a previous research, genome sequences of 343 yeast/fungi species along with their comprehensive genome annotations were made publicly available [35]. As part of the annotation process, each gene in the dataset was assigned to an ortholog group (OG) spanning multiple species. To compute the conservation score, this study used the Jensen-Shannon divergence (JSD) [97], where js_divergence was used as an estimation method

for estimation. The JSD software was utilized to compare the distribution of amino acids at each position in a protein alignment to a background distribution of amino acids. The resulting JSD value for each position was then transformed into a conservation score, with higher scores indicating greater conservation. The dN/dS ratios at the gene level were determined for pairs of orthologous genes across a set of 343 species using yn00 from PAML v4.7 on their respective gene sequences [98], in which yn00 is a program specifically designed to estimate the synonymous (dS) and nonsynonymous (dN) substitution rates between pairs of protein-coding sequences. This computational framework takes single-copy OGs as input and extracts gene-level dN/dS values from PAML output files as output. Moreover, the number of gene occurrence can be directly obtained from the annotation data, while the average paralog number was calculated by dividing the number of sequences in one OG by the total number of unique species in that OG.

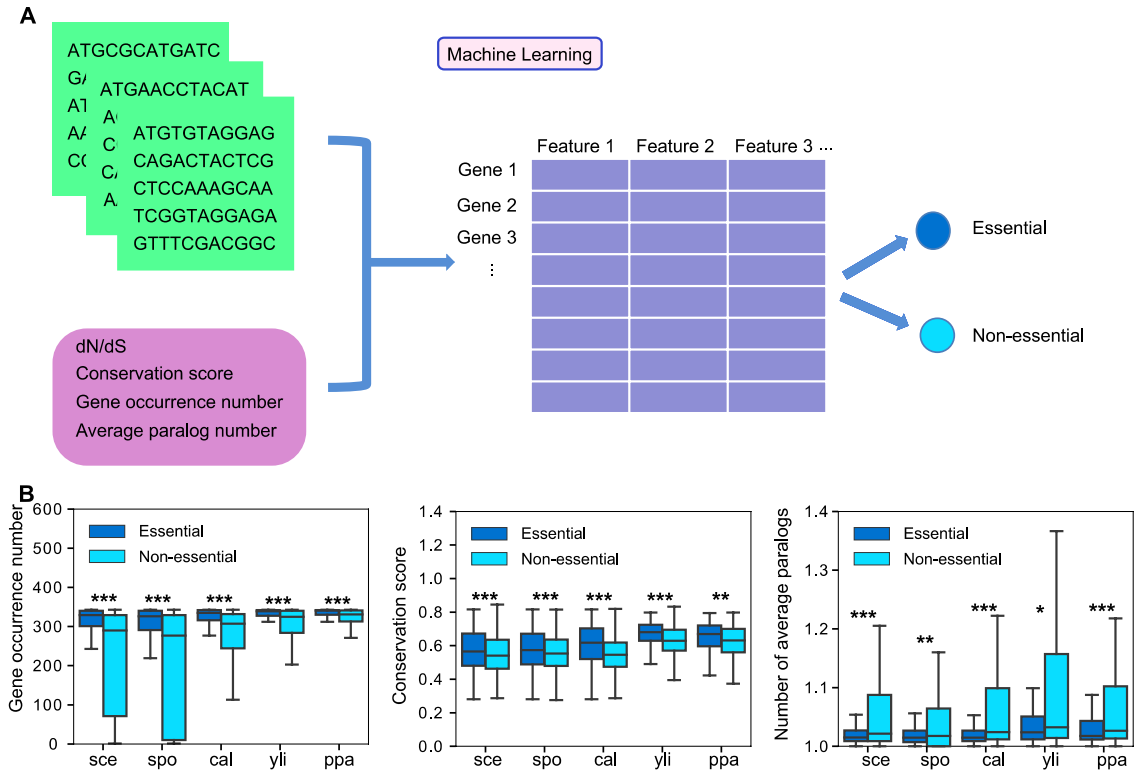

Figure 4 A schematic workflow illustrating essential gene prediction using ML methods and feature analysis between essential and non-essential genes. (A) An overview of gene essentiality prediction based on ML approaches. (B) Evolutionary feature analysis by comparing essential and non-essential genes across various yeast species. Statistical significance was denoted using symbols: ∗ for *P* value <= 0.05, ∗∗ for *P* value <= 0.01, and ∗∗∗ for *P* value <= 0.001. The following species with experimental data were included in the analysis: *S. cerevisiae* (sce), *S. pombe* (spo), *C. albicans* (cal), *Y. lipolytica* (yli), and *K. pastoris* (ppa). Only values within the range of 1-1.8 for the average number of paralogs were displayed.

Through a comparison of available evolutionary features between essential and non-essential genes across several yeast species, it was discovered that these features can indeed be utilized to differentiate between the two types of genes (see **Figure 4B**). For example,

when comparing the gene occurrence number across five species, the median value of essential genes was indeed significantly higher than that for non-essential genes, with a *P* value that was statistically significant.

After computing all sequence-based and evolution-based features, two libraries were generated: one for the training dataset and another for the testing dataset. Each library contained input and label data for multiple gene lists. The input data comprised various features, including sequence-based and evolution-based features, while the label data represented the gene essentiality status. Essential genes were assigned a label of '1', while non-essential genes were assigned a label of '0'. Using the Python-based package, an SVM and RF model were trained using the datasets to classify essential and non-essential genes based on their patterns (**Figure 4A**).

2.1.4 Evaluation of the ML model performance

To assess the predictive capabilities of different ML models, a five-fold cross validation approach was employed, and the receiver operating characteristic (ROC) curve was utilized. The ROC curve was generated by plotting the true positive rate (TPR) on the y-axis against the false positive rate (FPR) on the x-axis. The area under the ROC curve (AUC) was then computed, with a higher value indicating better performance of the ML model in gene essentiality prediction.

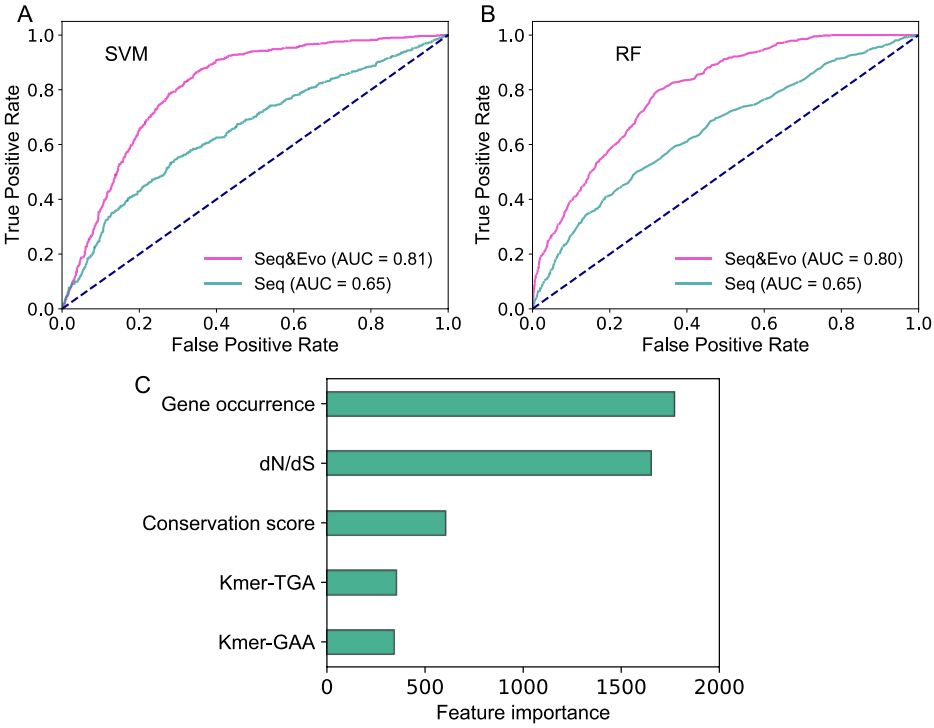

Figure 5 Improved essential gene prediction by integrating evolutionary parameters. The SVM algorithm (A) and the RF algorithm (B) both showed improved accuracy for predicting essential genes on the testing dataset when evolutionary parameters were incorporated, as evidenced by the higher AUC values of the ROC curves. (C) The importance scores of features that contribute to essential gene prediction were determined using the chi-square test method, and the two features Kmer-TGA and Kmer-GAA correspond to specific 3-nucleotide sequence fragments. Only features with high importance scores are shown, while those with lower scores are excluded.

For TPR and FPR, they were calculated as follows:

$$\text{True Positive Rate (or Sensitivity)} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

$$\text{False Positive Rate} = 1 - \text{Specificity} \quad (5)$$

where TP, TN, FP, FN denote true positive, true negative, false positive and false negative, respectively. In other words, TP is the number of essential genes (based on their annotations in the gathered datasets) that were predicted as essential genes, TN is the number of non-essential genes predicted as non-essential genes, FP is the number of non-essential genes predicted as essential genes and FN is the number of essential genes predicted as non-essential genes.

In the comparison of gene essentiality prediction using sequence features alone and in combination with evolution-based features, it was observed that the AUC values for essential gene prediction on the testing dataset were improved from 0.65 to 0.81 and 0.65 to 0.80 for the SVM and RF algorithms, respectively (**Figure 5A-B**). These findings suggest that integrating evolution-based features can indeed lead to a substantial improvement in gene essentiality prediction. This is consistent with previous observations that essential genes tend to be more conserved across species than non-essential genes (**Figure 4B**), and that evolutionary information can thus be used to improve gene essentiality prediction.

One example of an evolution-based feature that can improve gene essentiality prediction is the gene occurrence number. This feature provides insights into the functional importance and conservation of genes across evolutionary time. Essential genes are typically more conserved across different species, as they play fundamental roles in cellular processes that are necessary for the survival of the organism, such as DNA replication, protein translation, and metabolism [99]. As a result, essential genes are more likely to occur in a greater number of species compared to non-essential genes. Another example is dN/dS, which measures the ratio of nonsynonymous substitutions (dN) to synonymous substitutions (dS) in protein-coding genes. Essential genes have evolved to fulfill essential functions that require a specific amino acid sequence, and thus too many nonsynonymous sites would affect these essential functions.

To analyze the detailed contribution of individual features, all features were included as input in the ML model, and a chi-square test [100] was conducted to rank them based on their contribution to gene essentiality prediction (**Figure 5C**). The results revealed that evolution-based features, such as gene occurrence number, dN/dS, and protein-level conservation score, were the top three important features. Gene occurrence number was

identified as the most influential feature. Specifically, two Kmer (indicating gene frequencies) features, Kmer-TGA and Kmer-GAA, were found to have a relatively higher contribution to gene essentiality prediction. This suggests that genes containing these two sequence fragments are more likely to be essential genes compared to genes containing other sequence fragments.

Following the model evaluation and training process using the SVM algorithm, an ML model was generated and applied to predict essential genes for 338 out of 343 fungal species without experimental data. In paper I, these predictions are made publicly available via the Figshare platform, providing a valuable resource for future research in the yeast or fungi community.

2.2 DL on enzyme turnover number (Paper II)

Enzyme turnover number, also known as k_{cat}, refers to the maximum number of substrate molecules that can be converted to product per active site per unit time when the enzyme is saturated with substrate [101]. This crucial parameter of enzyme kinetics, which signifies how fast or efficient an enzyme functions, is necessary for understanding the growth rate, proteome composition, and physiology of organisms [102, 103]. However, obtaining experimental data on k_{cat} is time-consuming and labour-intensive, and for most enzymes, this information is unknown due to the vast array of existing organisms. Despite this challenge, Heckmann et al. previously built predictive ML models by compiling a diverse set of features, including network properties, enzyme structural properties, biochemical information, and assay conditions [104]. Nevertheless, acquiring such features is typically difficult and labour-intensive, limiting this approach to only well-studied organisms like *E. coli*. To overcome this limitation, a DL approach called DLKcat was developed in this study, which utilizes substrate structures and protein sequences as inputs and has demonstrated its ability to predict k_{cat} values on a large scale for various organisms.

2.2.1 Data preparation for the DL model

To collect the k_{cat} data, customized scripts were used to extract the dataset for the DL model construction from the BRENDA [48] and SABIO-RK [49] databases via their respective application programming interfaces (API). During this process, several rounds of data cleaning were performed to ensure data quality (**Figure 6**). The substrate simplified molecular input line entry system (SMILES), which is a string notation used to represent the substrate structure, was extracted using the substrate name to query the PubChem compound database [40]. Two approaches were employed to query protein sequences: for entries with UniProt ID information, the amino acid sequences were obtained using the UniProt API [46] and Biopython v.1.78 (https://biopython.org/), while for entries without UniProt ID, the amino acid sequences were acquired from the UniProt and BRENDA databases based on their EC number and organism information. In this study, k_{cat} values from both wildtype and mutated enzymes were considered. For wildtype enzymes, the

sequences of entries were directly mapped. For mutated enzymes, the sequences of those entries were altered based on the mutated sites.

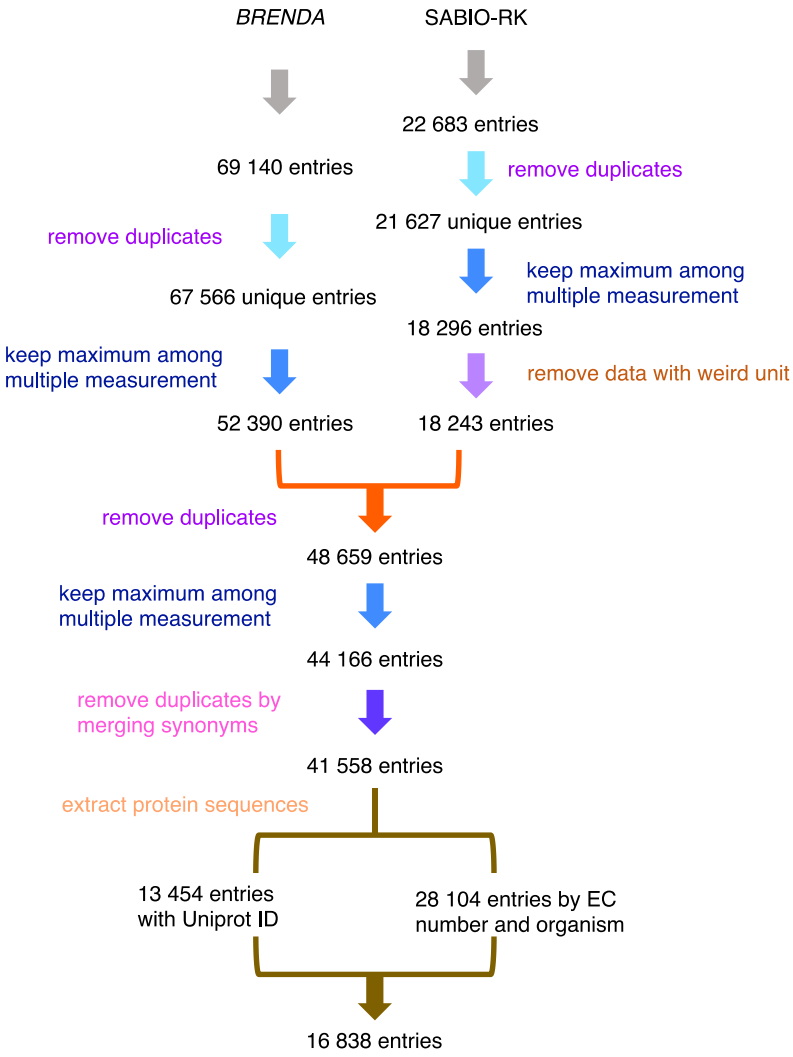

Figure 6 Dataset collection and pre-processing steps involved in constructing the DL model.

The dataset used for the DL model construction includes substrate name, organism information, EC number, protein identifier (UniProt ID), enzyme type, and k_{cat} values. For the majority of data entries, the assay conditions (pH, temperature) were not specified. Including pH and temperature as features would filter out a large part of the dataset and significantly reduce the diversity of enzymes. Therefore, pH and temperature were not included as features in the collected dataset. The final dataset consisted of 16,838 unique entries catalyzed by 7,822 unique protein sequences from 851 organisms and involving 2,672 unique substrates (**Figure 7A-C**).

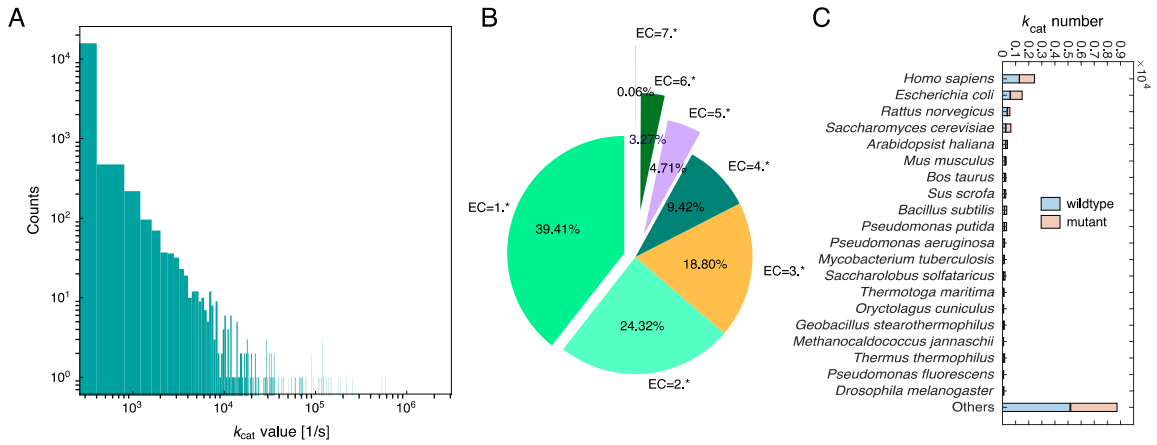

Figure 7 Analysis of in vitro k_{cat} values from the BRENDA and SABIO-RK databases after several rounds of data pre-processing and cleaning. (A) Data distribution of in vitro k_{cat} values. (B) In vitro k_{cat} values classification by the first digit of the EC number. (C) In vitro k_{cat} values classification by species.

2.2.2 Construction of the DL model pipeline

The DL approach for in vitro k_{cat} value prediction was developed by integrating a GNN for substrates and a CNN for proteins (**Figure 8**). The combination of GNN and CNN is particularly effective for processing pairs of data with varying structures, such as molecular graphs and protein sequences. The molecular graphs used to represent substrates consist of vertices that represent atoms and edges that represent chemical bonds, while the protein sequences comprise a string of characters representing amino acids.

Since substrates typically consist of a limited range of chemical atoms and bonds, additional learning parameters were required. To accomplish this, r-radius subgraphs were used to obtain vector representations, which were induced by neighbouring vertices and edges within a given radius r from a vertex [105]. To begin, the substrate SMILES information was transformed into a molecular graph using RDKit v.2020.09.1 (https://www.rdkit.org). Next, the GNN updated each atom vector and its neighbouring atom vectors, which were transformed by the neural network using a nonlinear function (i.e., ReLU [61]). Additionally, two transitions were implemented in the GNN: vertex transitions and edge transitions. These transitions were designed to ensure that local vertex and edge information was propagated throughout the graph by iteratively processing and summing neighbouring embeddings. Ultimately, the output of the GNN was a set of real-valued molecular vector representations for substrates.

To obtain the protein sequences representations, the CNN framework was utilized to scan protein sequences. The neural network transformed the protein sequences through a nonlinear function (i.e., ReLU), to generate the vector representations. To apply the CNN to proteins, "words" were defined in the protein sequence, and the sequence was split into overlapping n-grams of amino acids with n set to 1, 2, or 3 to prevent low-frequency words in the learning representations [106]. The protein sequences were then translated into

various word embeddings, and the CNN used a filter function with a weight matrix to compute hidden vectors from the input word embeddings. Subsequently, a set of hidden vectors for the split subsequences was obtained based on the n-gram amino acid splitting.

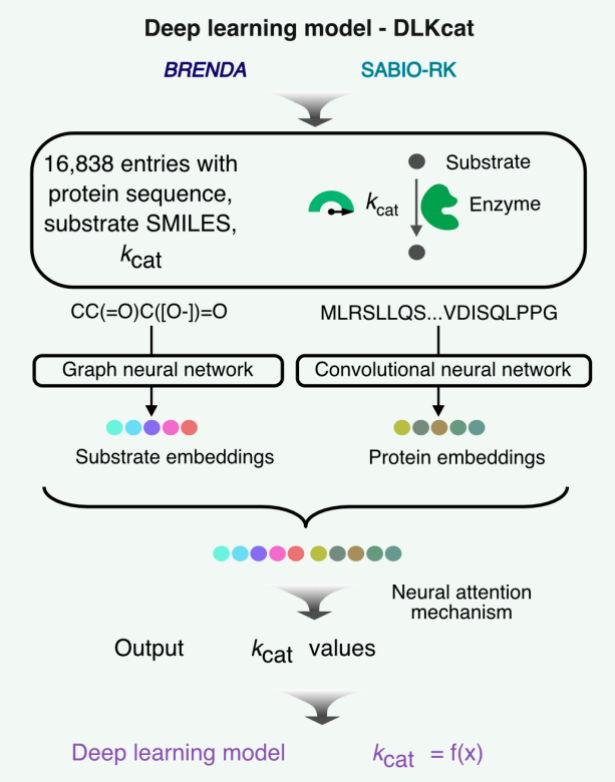

Figure 8 Schematic overview of the DL approach developed to predict k_{cat} values by integrating a GNN for substrates and a CNN for proteins.

After obtaining the substrate representations and protein sequence representations, they were concatenated along with an output vector (k_{cat} value) to train the DL model based on the neural attention mechanism [105]. During training, the total dataset was shuffled before being randomly split into a training dataset, validation dataset, and test dataset in an 80:10:10 ratio. The objective of the training process was to minimize the loss function using the substrate-protein pairs and k_{cat} values in the training dataset. PyTorch software package was utilized for building and training models, accessed through the Python package interface under the CUDA environment.

2.2.3 DL model performance for k_{cat} prediction

The DL model training architecture presented in the above section includes various hyperparameters that should be tuned to optimize the model performance. Specifically, I explored different values for the r-radius (0, 1, or 2), the n-gram (1, 2, or 3), the vector dimensionality (5, 10, or 20), the number of layers in GNN (2, 3, or 4), and the number of layers in CNN (2, 3, or 4) to identify the optimal settings that would influence the DL model performance (**Figure 9A-C**), where the vector dimensionality refers to the number of input neurons in the neural network. After hyperparameter tuning, the optimal

hyperparameter settings were found to be an r-radius of 2, n-gram of 3, vector dimensionality of 20, 3 layers in GNN, and 3 layers in CNN. Using these settings, the DL model was trained for its optimal performance.

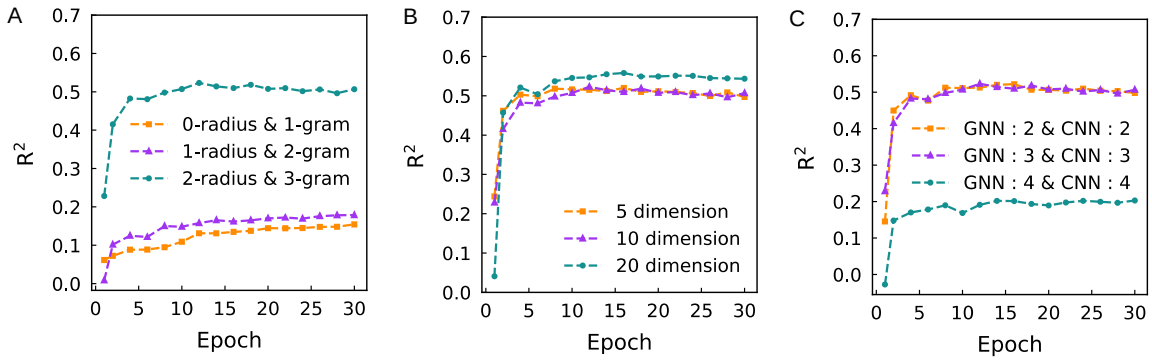

Figure 9 Hyperparameter tuning on the validation dataset. (A) Learning curves for various r-radius subgraphs and n-gram amino acids. (B) Learning curves for different vector dimensionality. (C) Learning curves for varying numbers of layers in GNN and CNN.

Once the DL model was trained using the optimal hyperparameter settings, the resulting model was further evaluated to determine its detailed performance. The final model's performance was initially checked on the entire dataset (including the training, validation, and test datasets). The results showed that the DL model had an excellent performance, as evidenced by a high Pearson's r value of 0.88 and a *P* value of 0 (**Figure 10A**). Moreover, the model demonstrated high predictive accuracy when it was tested on the test dataset, as indicated by a Pearson's r value of 0.71 and a *P* value of 1.5e-262 (**Figure 10B**). The model also performed well on the subset of the test dataset where at least either the substrate or enzyme was not present in the training dataset, with a Pearson's r value of 0.70 and a *P* value of 8.0e-88 (**Figure 10C**). The success of the DL model is mainly attributed to two critical factors: the large size and high quality of the dataset used in this work, and the use of complex models by combining GNN and CNN.

Apart from the conventional performance evaluation on the test dataset, the DL model in this study was also evaluated based on its ability to predict k_{cat} values in different metabolic contexts. To explore the metabolic contexts of all wildtype enzymes in the entire dataset, the enzymes were categorized into four modules based on the KEGG database [43]. These modules include primary metabolism-CE (carbohydrate and energy), which encompasses the main carbon and energy metabolism pathways such as glycolysis/gluconeogenesis, TCA cycle, and pentose phosphate pathway; primary metabolism-AFN (amino acids, fatty acids, and nucleotides); intermediate metabolism, related to the biosynthesis and degradation of cellular components, such as coenzymes and cofactors; and secondary metabolism, which is associated with metabolites produced in specific cells or tissues, e.g., flavonoid biosynthesis, caffeine metabolism, bile acid biosynthesis, etc. Based on the trained DL model, it was found that enzymes associated with primary-CE metabolism on average exhibited a higher predicted k_{cat} value than those of primary-AFN, secondary, and

intermediate metabolism (**Figure 10D**). Additionally, enzymes associated with intermediate metabolism exhibited a slightly lower k_{cat} value on average compared to those of primary-AFN and secondary metabolism (**Figure 10D**). These results are highly consistent with previous reported findings that enzyme-substrate pairs from central carbon metabolism tend to have relatively higher k_{cat} values than secondary and intermediate metabolism [107].

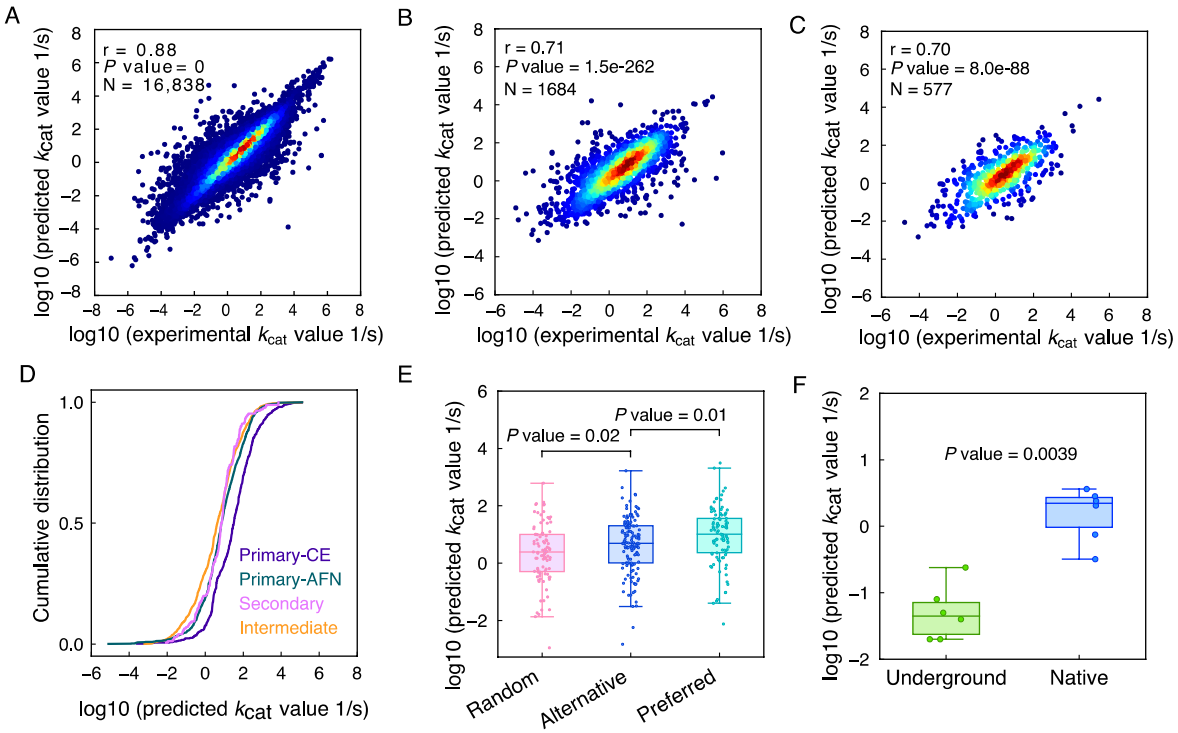

Figure 10 Performance evaluation of DLKcat. Performance of the final DL model on (A) the entire dataset; (B) the test dataset; and (C) a subset of the test dataset where either the protein sequence or the substrate was not present in the training dataset. The correlation between predicted k_{cat} values and experimental k_{cat} values was evaluated, with the temperature of the color representing the density of data points. Student's t-test was used to calculate the *P* value for Pearson's correlation. (D) Cumulative distribution of DL-based k_{cat} values for enzyme-substrate pairs belonging to different metabolic contexts. CE: carbohydrate and energy; AFN: amino acids, fatty acids, and nucleotides. (E) Enzyme promiscuity analysis on the test dataset. For enzymes with multiple substrates, the substrates were divided into preferred and alternative based on their experimental measured k_{cat} values, and then used the predicted k_{cat} values for this boxplot. The random substrates were chosen randomly from the compound dataset in the training data, except for the documented substrates and products for the tested enzyme. (F) Comparison of predicted k_{cat} values for native substrates and underground substrates with the human aldo–keto reductase enzyme as a case study. In each box plot (E and F), the central band represents the median value, the box represents the upper and lower quartiles, and the whiskers extend up to 1.5 times the interquartile range beyond the box range. A two-sided Wilcoxon rank sum test was used to calculate the *P* values (E and F).

Enzyme promiscuity refers to the capability of an enzyme to catalyze multiple reactions or substrates. Understanding enzyme promiscuity and related underground metabolism is a crucial topic in evolutionary biology [108], with potential implications for protein and metabolic engineering. Enzyme promiscuity is an important factor in enzyme evolution and can be harnessed to generate novel enzymes with desired catalytic properties and broader substrate ranges, aiding in the development of new industrial applications [109]. In this regard, I validated the performance of the DL model in predicting enzyme

promiscuity. To achieve this, a compound dataset was created using compound names and SMILES based on the substrate information in the training data. For enzymes that had k_{cat} values reported from different substrates, the substrates were categorized into preferred and alternative based on their experimentally measured k_{cat} values. The substrates that were not documented as substrates or products in the training data were randomly selected as the random substrates. Through a comparison of the predicted k_{cat} values for preferred substrates, alternative substrates, and random substrates of promiscuous enzymes (**Figure 10E**), it was found that the DL model was capable of predicting that enzymes indeed have a higher k_{cat} value for the preferred substrates in comparison to the alternative substrates (*P* value = 0.01). Moreover, the DL model was able to predict that enzymes have a higher k_{cat} value for the alternative substrates compared with the random substrates (*P* value = 0.02). These findings validate the predictive power of the DL model in identifying preferred substrates in enzyme promiscuity.

The evaluation of native and underground metabolism can be illustrated through the analysis of k_{cat} data for the human aldo-keto reductase and its multiple substrates [110]. In this study, the substrates with the top 10% catalytic ability (experimental k_{cat} value) were defined as native substrates (n = 6), while those with the last 10% catalytic ability (experimental k_{cat} value) were considered underground substrates (n = 6), as defined in the reference [110]. The predicted k_{cat} values by DLKcat revealed a significant difference (*P* = 0.0039) between the native substrates (top 10% of k_{cat} values with a median of 2.22 s^-1) and the underground substrates (bottom 10% of k_{cat} values with a median of 0.04 s^-1) (**Figure 10F**).

2.2.4 Prediction of k_{cat} values for mutated enzymes

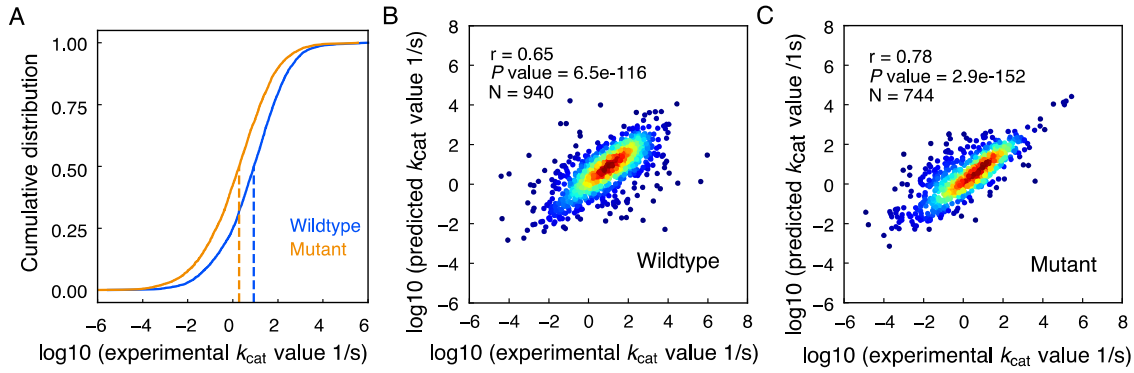

Figure 11 Prediction performance for wildtype and mutated enzymes on the test dataset. (A) Cumulative distribution of experimentally measured k_{cat} values for wildtype and mutated enzymes. (B-C) Prediction performance of k_{cat} values for (B) all wildtype enzymes and (C) all mutated enzymes via the DL model. The temperature of the color represents the density of data points. Student's t test was performed to calculate the *P* value for Pearson's correlation in B-C.

In addition to its overall good performance for predicting k_{cat} values, the next thing that I explored is whether the DL model could capture more details, such as the effects of amino acid substitutions on k_{cat} values of individual enzymes. To explore this, the original

annotated dataset was divided into two categories: one containing wildtype enzymes and the other containing mutated enzymes with amino acid substitutions. It can be observed that the median k_{cat} value of mutant enzymes is lower than that for wildtype enzymes (**Figure 11A**). Moreover, the DL model was found to be a good predictor of k_{cat} values for both wildtype enzymes (Pearson's r = 0.65) and mutated enzymes (Pearson's r = 0.78) when evaluating its performance on the test dataset (**Figure 11B-C**). These results suggest that the DL model can be a reliable tool for predicting k_{cat} values of both wildtype and mutated enzymes.

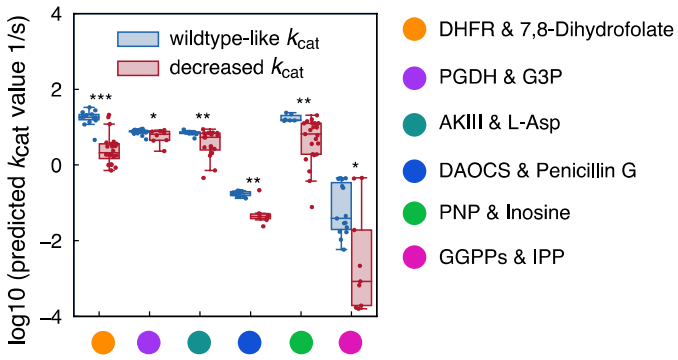

Figure 12 Comparison of predicted k_{cat} values on several mutated enzyme-substrate pairs. Two categories were used for comparison: enzymes with wildtype-like k_{cat} and enzymes with decreased k_{cat}. Enzyme abbreviations used are DHFR (dihydrofolate reductase), PGDH (D-3-phosphoglycerate dehydrogenase), AKIII (aspartokinase III), DAOCS (deacetoxycephalosporin C synthase), PNP (purine nucleoside phosphorylase), and GGPPs (geranylgeranyl pyrophosphate synthase). Substrate abbreviations used are G3P (glycerate 3-phosphate), L-Asp (L-Aspartate), and IPP (isopentenyl diphosphate). Significance levels are indicated as follows: *P* value < 0.05 (*), *P* value < 0.01 (**) and *P* value < 0.001 (***). A two-sided Wilcoxon rank sum test was used to calculate the *P* value for the two-group comparisons in this analysis. In the boxplot, the central band represents the median value, the box represents the upper and lower quartiles, and the whiskers extend up to 1.5 times the interquartile range beyond the box range.

Next, several extensively investigated enzyme-substrate pairs were collected from literature and open-access databases where each enzyme-substrate pair had k_{cat} values reported for multiple amino acid substitutions. The entries for each enzyme-substrate pair were subsequently divided into two groups based on their experimentally measured k_{cat} values: (i) wildtype-like k_{cat} group, where k_{cat} values were within 0.5-2.0 fold change of the wildtype k_{cat} value; or (ii) decreased k_{cat} group, where k_{cat} values were less than 0.5 fold change of the wildtype k_{cat} value. Scarcity of mutated enzymes with k_{cat} values over 2-fold of wildtype k_{cat} precluded defining the increased k_{cat} group. After that, by using the DL-predicted k_{cat} values, it was validated that enzymes from the decreased k_{cat} group indeed showed significantly lower k_{cat} values compared to those of enzymes from the wildtype-like k_{cat} group for all of the enzyme-substrate pairs (**Figure 12**). Therefore, the DL model can effectively capture the effects of minor changes in protein sequences on the activities of individual enzymes.

2.2.5 Interpretation of the DL model

Unraveling the black box of DL is a great challenge for DL-based applications in biology and chemistry [111]. One approach to investigate which subsequences in a protein are more

important for the substrate is to integrate a neural attention mechanism that assigns attention weights to each subsequence [105]. This neural attention mechanism traces important signals from the output of the neural network to the input, with the input being a molecular vector and a set of vectors of subsequences in one protein produced by the substrate and the protein, respectively, and the output being the attention weight assigned to each subsequence.

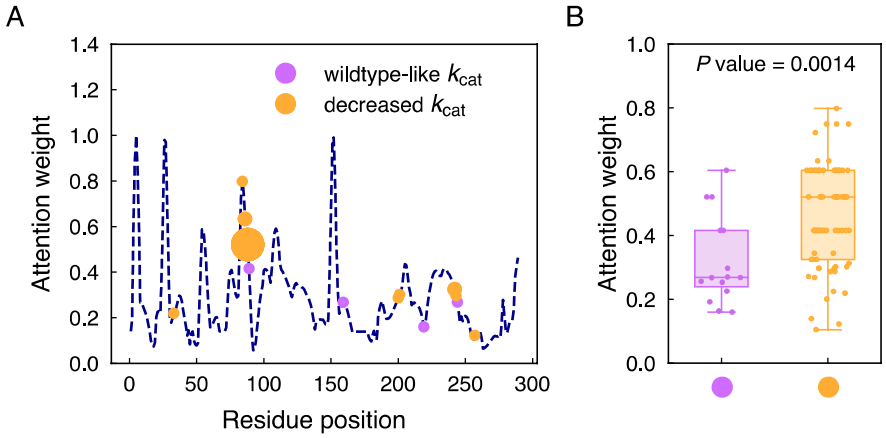

Figure 13 Interpretation of the DL model using the purine nucleoside phosphorylase (PNP) enzyme as a case study. (A) Attention weight of sequence position in the wildtype PNP enzyme, using inosine as the substrate. The mutated residues in each of the mutated enzymes (both wildtype-like k_{cat} and decreased k_{cat}) were marked on the curve according to their mutated residue. The dot size indicates the number of mutated enzymes occurring in that mutated position. (B) Boxplot comparing the overall attention weight for the PNP-Inosine pair between enzymes with wildtype-like k_{cat} and decreased k_{cat}. The *P* value was calculated using a two-sided Wilcoxon rank sum test. In this boxplot, the central band represents the median value, the box represents the upper and lower quartiles, and the whiskers extend up to 1.5 times the interquartile range beyond the box range.

In the case of the wildtype purine nucleoside phosphorylase (PNP) on the substrate inosine, which is an important enzyme of *Homo sapiens* with rich experimental mutagenesis data and rich mutation sites data, DL can capture which sequence position or residue is more important for the enzyme's catalytic capability through the output-attention weight (**Figure 13A**). Enzymes from the decreased k_{cat} group in this enzyme-substrate pair presented a significantly higher attention weight compared to those of enzymes from the wildtype-like k_{cat} group (**Figure 13B**). By marking these enzymes from both the decreased k_{cat} group and the wildtype-like k_{cat} group into the curve based on the mutated position, it was found that residues that were mutated in the decreased k_{cat} group had significantly higher attention weights (**Figure 13A**). This indicates that the calculation of attention weights from the DL model has the potential to identify amino acid residues whose mutation would likely have a more substantial effect on enzyme activity.

2.2.6 Biological insights gained with the aid of the DL model

Enzyme-constrained genome-scale metabolic models (ecGEMs) are computational tools used in systems biology to study the metabolic capabilities and functions of organisms at a genome-scale level [103, 112]. The distinctive feature of ecGEMs is that they incorporate enzyme kinetics information to constrain the metabolic fluxes and predict the metabolic

behavior of the organism under different conditions. This property allows for a more realistic representation of the metabolic network and more accurate predictions of the organism's phenotypes. In addition, the ecGEMs have played a critical role in accurately simulating maximum growth abilities, metabolic shifts, and proteome allocations, as the whole-cell metabolic network in ecGEMs is constrained by enzyme catalytic capacities.

The ecGEMs rely heavily on genome-scale k_{cat} values, making the enzyme k_{cat} parameter one of the most significant factors that affect their reconstruction. Previously, GEMs were reconstructed for 332 yeast species and 11 outgroup fungi [34], but only 14 of those GEMs were successfully expanded with enzyme constraints using an ecGEM reconstruction pipeline. This pipeline was customized and relied solely on k_{cat} values reported in the BRENDA database, as DL tools or other related algorithms were unavailable at the time. The limited availability of k_{cat} values has prevented the reconstruction of ecGEMs for more species [113] [36]. As the DL model developed in this study allows prediction of almost all k_{cat} values for metabolic enzymes against any substrates for any species (except for enzyme-substrate pairs with generic substrates lacking detailed SMILES information), this enabled generation of ecGEMs for all 343 yeast and fungi species. Finally, through using the metabolite and enzyme information extracted from the 343 GEMs as the input of the DL model for k_{cat} prediction, I predicted k_{cat} values for around three million protein-substrate pairs in 343 yeast/fungi species.

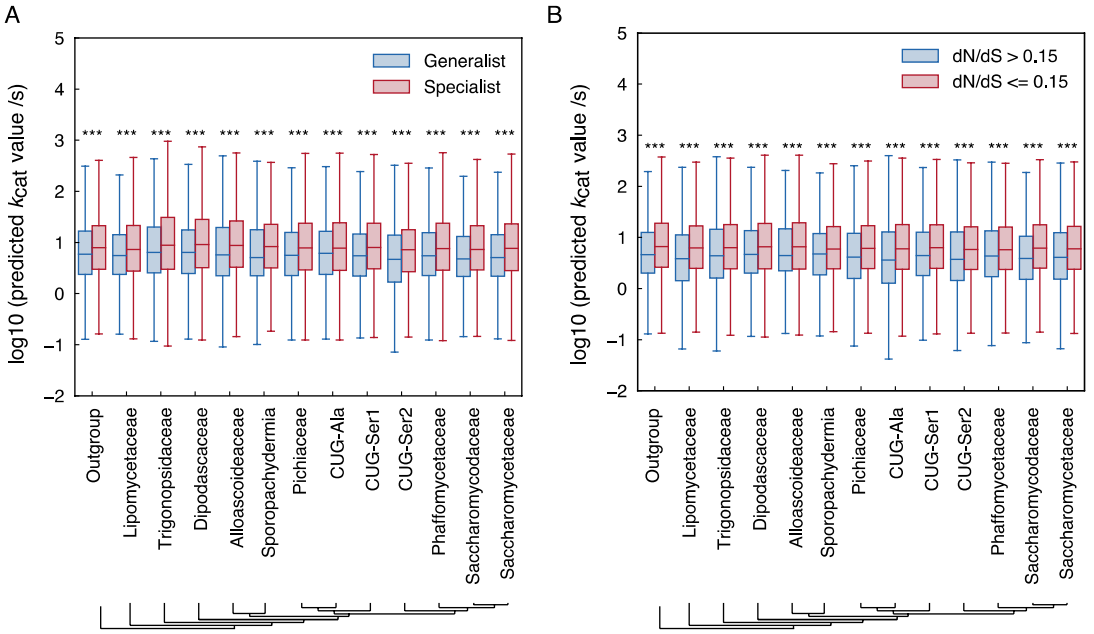

Figure 14 Analysis of the evolutionary patterns in predicted k_{cat} values for 343 yeast/fungi species. (A) The enzyme k_{cat} values associated with generalist and specialist metabolism were evaluated for all 343 species. (B) The relationship between enzyme k_{cat} values and the ratio of non-synonymous to synonymous substitutions (dN/dS) for all 343 species. The x-axis shows the genus-level phylogeny for 332 yeast species, divided into 12 major clades, with 11 outgroup species included. The cutoff of 0.15 was set according to the distribution of dN/dS values in these species. Statistical significance was indicated by *** (*P* value < 0.001). The *P* value was calculated using a two-sided Wilcoxon rank sum test. In each boxplot, the central band represents the median value, the box represents the upper and lower quartiles, and the whiskers extend up to 1.5 times the interquartile range beyond the box range.

Furthermore, comprehensive analysis of these predicted k_{cat} values revealed a global trend showing that specialist enzymes with narrow substrate specificity exhibited higher k_{cat} values than generalist enzymes that catalyze multiple reactions (**Figure 14A**). This trend aligns with the hypothesis that ancestral enzymes with broad substrate specificity and low catalytic efficiency improved their k_{cat} values as they evolved into specialists through mutation, gene duplication and HGT [114]. These findings also hold true for fungi and are consistent with those reported for *E. coli* [114]. In addition, the potential link between enzyme k_{cat} values and dN/dS was further evaluated based on these predicted k_{cat} values (**Figure 14B**). It can be observed that conserved enzymes with lower dN/dS values have significantly higher k_{cat} values compared to relatively less conserved enzymes with high dN/dS. This implies that conserved yeast and fungi enzymes under evolutionary pressure are adapted to have higher k_{cat} values.

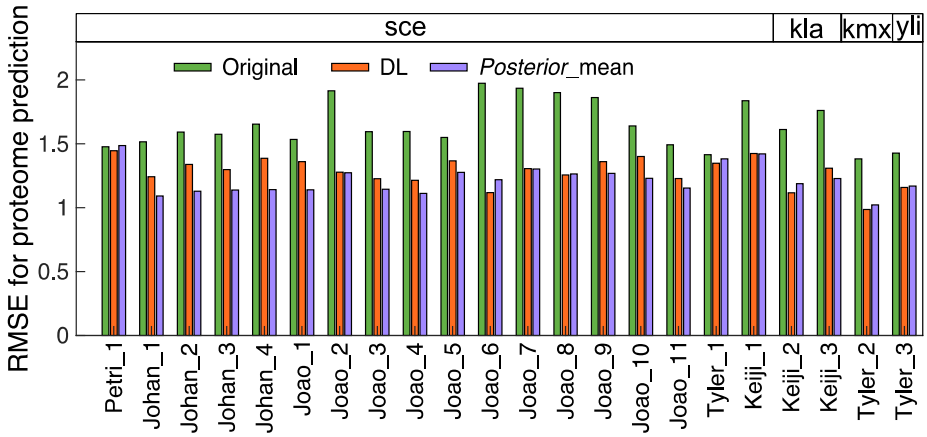

Figure 15 Comparison of three different ecGEM modelling pipelines, namely original-ecGEM, DL-ecGEM, and posterior-mean-ecGEM, in their ability to predict quantitative proteome. Four species with known absolute proteome data were assessed in this evaluation. The x-axis represents various proteome datasets, which are available on the GitHub repository (https://github.com/SysBioChalmers/DLKcat). The evaluated species were *S. cerevisiae* (sce), *Kluyveromyces lactis* (kla), *Kluyveromyces marxianus* (kmx), and *Y. lipolytica* (yli).

The efficacy of the DLKcat computational tool in predicting phenotypes was also assessed through proteome predictions with ecGEMs. To do this, three types of ecGEMs were reconstructed: Original-ecGEMs, which were built using k_{cat} profiles extracted from the BRENDA and SABIO-RK databases; DL-ecGEMs, which were reconstructed using k_{cat} profiles predicted by DLKcat; and posterior-mean-ecGEMs, which were parameterized with mean k_{cat} values from 100 posterior datasets after the Bayesian training process. These three models were used to predict protein abundances and were compared with published quantitative proteomics data from four species under different carbon sources, culture modes, and medium setups. Regarding the protein abundance simulation, the medium was set to match the experimental condition. In the case of the chemostat condition, the growth rate was fixed to the dilution rate, and the carbon source uptake rate was minimized, which is a standard configuration for simulating chemostat conditions. On the other hand, in the batch condition, the growth rate maximization was used as the objective. The simulated

protein abundances, which could be extracted from the fluxes, were then compared with those in collected proteome datasets.

By comparison, posterior-mean-ecGEMs generally had the lowest root mean square error (RMSE), indicating the best performance in predicting proteome data, while DL-ecGEMs reduced RMSE by 30% compared to Original-ecGEMs (**Figure 15**). These results demonstrate the potential value of DLKcat-predicted k_{cat} profiles in ecGEM reconstruction, as it can serve as a useful tool for connecting genotype and phenotype.

2.3 ML on protein production (Paper III)

Since the emergence of the commercial recombinant human insulin production by *Escherichia coli* in 1982, the biopharmaceutical industry has grown rapidly [115]. Nowadays, biotechnology-based pharmaceutical production has formed a global biopharmaceutical market. With the development of advanced synthetic biology and metabolic engineering tools, the production of recombinant proteins by yeast has become crucial in the biopharmaceutical industry, and the development of yeast platform strains capable of overproducing various biopharmaceutical proteins is highly desirable [116]. However, achieving this requires a fundamental understanding of the cellular machinery, particularly the protein secretory pathway. The secretory pathway spans several different organelles that carry out peptide translocation, folding, Endoplasmic reticulum (ER)-associated protein degradation (ERAD), sorting processes, as well as different post-translational modifications (PTMs) to ensure proper protein functionality [117]. In this study, a proteome-constrained genome-scale protein secretory model of yeast *S. cerevisiae* (pcSecYeast) was reconstructed, which allows simulating and explaining phenotypes related to the secretory capacity. The model was then used to simulate the production of eight different recombinant proteins, and a ML approach was integrated to analyze feature importance towards the production of recombinant proteins.

2.3.1 Simulation of the production of recombinant proteins in yeast

Recombinant proteins are transported and modified by various components of the secretory pathway in the yeast *S. cerevisiae*, depending on their amino acid composition and PTMs. In order to determine the factors that affect the levels of secreted proteins, the pcSecYeast model was enhanced to describe the production of eight different recombinant proteins by incorporating the corresponding recombinant protein production and secretion reactions, respectively. Subsequently, eight specific models were generated to simulate the maximum secretion of each of the eight recombinant proteins under varying growth rates. These proteins possess varying sizes and PTMs (**Figure 16A**), and their specific PTM profile dictates the combination of multiple processes required for their efficient production and secretion. As such, the secretory pathway can be seen as a complex production line. This provided an excellent opportunity to investigate how these factors affect the levels of secreted proteins.

A

Protein	abbr.	DSB	NG	OG	GPI	Length
Insulin precursor	IP	3	0	0	0	53
Human granulocyte colony stimulating factor	hGCSF	2	0	1	0	174
Hemoglobin	Hemoglobin	0	0	0	0	299
β-glucosidase	BGL	0	0	0	0	421
α-amylase	α-amylase	4	1	0	0	478
Acid phosphatase	PHO	8	9	0	0	435
Human serum albumin	HSA	17	1	0	0	585
Human transferin	HTF	19	0	1	0	679

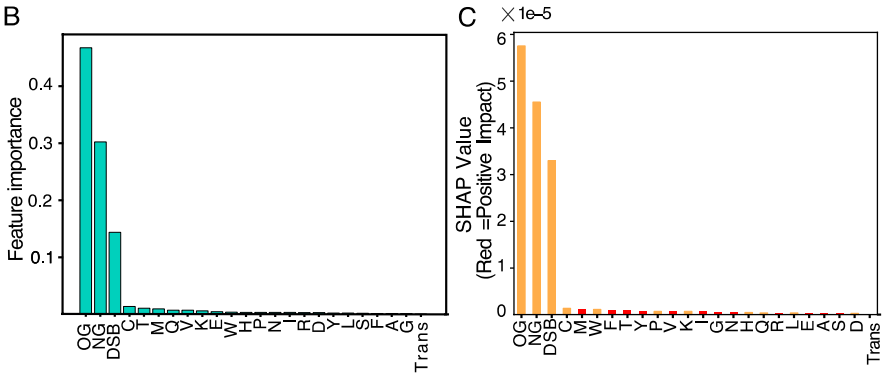

Figure 16 Feature importance analysis for recombinant protein production using ML. (A) Overview of protein features for eight recombinant proteins produced by *S. cerevisiae*. (B) Feature importance analysis towards recombinant protein production based on the built-in function in the random forest (RF) algorithm. (C) Feature importance analysis towards recombinant protein production by combining SHapley Additive exPlanations (SHAP) and RF. Abbreviations used: OG O-glycosylation site, NG N-glycosylation site, DSB disulfide bond number, Trans transmembrane domain, single letters stand for specific amino acids, abbr.: abbreviation.

2.3.2 ML for feature importance analysis towards protein production

To determine the protein features with the greatest impact on recombinant protein production, I integrated a ML approach to score the importance of factors. In this approach, various factors (PTMs, amino acid compositions) were used as input features and the maximum recombinant protein production rate from the pcSecYeast model simulation was used as the target label. The dataset was randomly divided into a training dataset, which constituted 80% of the total data, and a testing dataset, which constituted 20% of the data. A random forest (RF) regressor with 10 estimators was then used to train the ML model. Two techniques were utilized to compute the feature importance scores. The first method involved using the built-in feature_importances_ attribute of the RF model to obtain feature importance scores directly. The second method involved using the SHapley Additive exPlanations (SHAP) interpretation [118], which is based on game theory principles and estimates the contribution of each feature to the prediction.

The built-in feature_importances_ attribute in the RF model revealed that O-glycosylation and N-glycosylation are the two most influential features affecting recombinant protein production. In contrast, the frequencies of specific amino acids had a relatively minor impact on protein production (**Figure 16B**). This suggests that PTMs have a greater influence on protein production than amino acid composition. However, the built-in

function in RF only identifies which features have more or less influence and cannot indicate whether the influence is positive or negative.

To investigate the directionality of feature influence, the SHAP framework was applied to calculate feature importance scores from the RF algorithm. A higher SHAP value indicates a greater contribution to protein production. The correlations between the target protein production and various factors were visualized using color coding, where the red color indicates a positive effect on protein production, and the orange color indicates a negative effect. The analysis results revealed that O-glycosylation and N-glycosylation had a large negative impact on recombinant protein production (**Figure 16C**).

In addition, the significance of the feature importance analysis for protein production was further evaluated by performing five-fold cross validation on the training dataset (**Figure 17).** The dataset was partitioned randomly into five subsets of equal size, with four subsets used for training and the remaining subset for testing in each fold. This procedure was repeated five times, and each subset served as the testing set once. The feature importance scores were then visualized, and it was found that PTMs had a significantly greater impact on protein production than amino acid compositions. Specifically, the negative impact of O-glycosylation and N-glycosylation on protein production was further consolidated, suggesting that having more glycosylation sites may increase the burden on the cell.

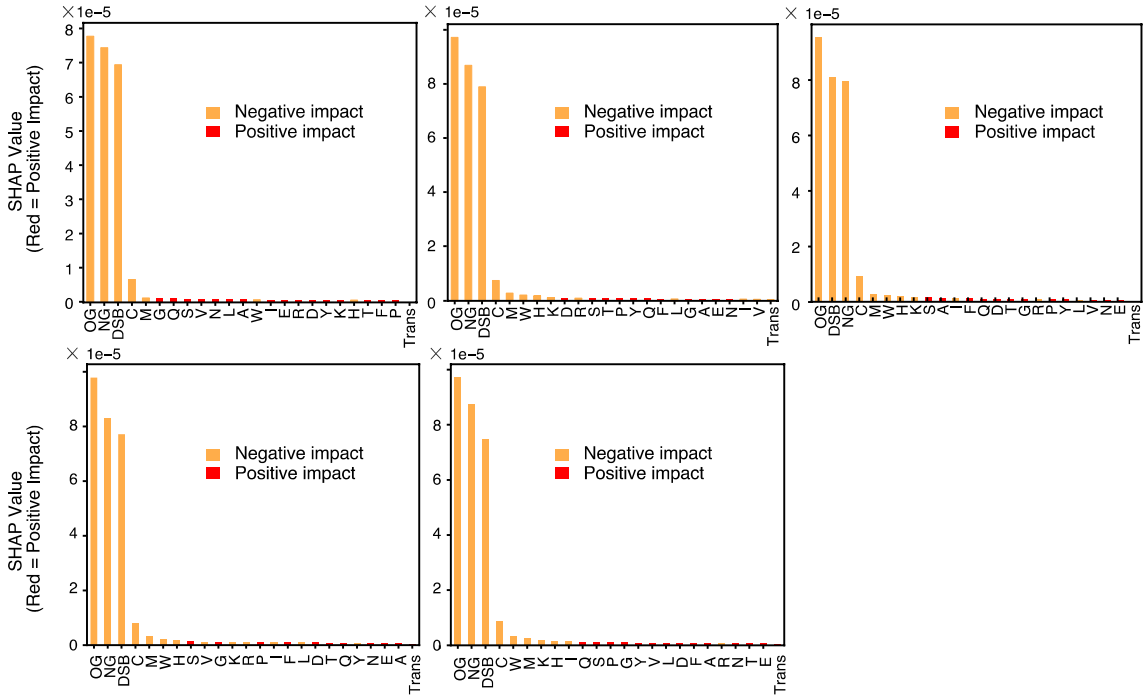

Figure 17 Assessment of the feature importance scores based on five-fold cross validation.

3. Development and applications of comparative genomics tools on yeasts

As introduced in the background section, comparative genomics can provide meaningful insights into the genetic basis of complex traits and is widely employed in diverse fields, such as evolutionary biology, systems biology, and biomedical research. This chapter consists of two sections focusing on the design and applications of comparative genomics tools on yeasts. The first study (**Paper IV**) introduces the construction of HGTphyloDetect, a computational toolbox aimed at facilitating the identification and phylogenetic analysis of horizontal gene transfer (HGT) events. The second study (**Paper I**) presents a comprehensive analysis on large-scale yeast species by combining HGT analysis, gene family expansion and contraction and GEM simulation, aiming to explore the underlying mechanisms of substrate utilization.

3.1 HGTphyloDetect - Detection of horizontal gene transfer (Paper IV)

As described in the section 1.6, HGT is a crucial factor in shaping genome evolution and facilitating gain-of-function abilities, as well as metabolic adaptation to different environmental niches. With the rapid expansion of genomic data, it has become increasingly feasible to identify putative HGT events on a genome-wide scale.

Despite significant advancements in this field, there are only a limited number of computational approaches available for predicting HGT events. One such method, HGTector, utilizes sequence homology search hit distribution statistics to detect HGT events on a genome-wide scale. However, it lacks the ability to provide a detailed phylogeny analysis to understand the underlying mechanisms of HGT [119]. Although HGT-Finder can determine the horizontal transfer index and probability value for each queried gene using phyletic distribution, this software is regrettably unavailable for download [120], necessitating the exploration of alternative tools. AvP is another technique that uses a phylogenetic framework to automate the identification of potential HGT events [121], but the quality of the produced phylogenetic trees is not particularly high. Additionally, it remains unclear whether AvP can detect HGT events involving evolutionarily closely related species.

To overcome these limitations, I have developed HGTphyloDetect, an open-source computational toolbox that combines high-throughput analysis with phylogenetic inference, to analyze HGT events. High throughput algorithms were employed to detect HGT events, irrespective of the evolutionary distance between the donors and the horizontally acquired genes. This emphasizes the versatility of HGTphyloDetect in detecting HGT events among genes from both closely and distantly related species.

3.1.1 Detecting horizontal gene transfer from phylogenetically distant organisms

In order to detect HGT events from phylogenetically distant organisms, such as prokaryotes to eukaryotes, a reliable and phylogeny-based workflow was developed as shown in **Figure 18**.

The first step in this workflow involves using the BLASTP algorithm to search for one particular gene or multiple genes of interest against the NCBI non-redundant (nr) protein database. After obtaining the BLASTP hits, taxonomic information associated with them is extracted from the NCBI taxonomy database using the ETE v3 toolkit [122]. Together with this information, the Alien Index scores are computed based on bbhG and bbhO as illustrated in **Figure 18**. In this context, bbhG and bbhO denote the E-values of the top BLAST hit in the ingroup and outgroup lineages, respectively. The ingroup lineage refers to the species within the kingdom, but outside of the subphylum. Conversely, the outgroup lineage comprises all species outside of the kingdom. The mathematical formula for the Alien Index used here was originally introduced in a notable study by Gladyshev et al [123]. In their work, they established that an Alien Index value of 45 or higher is a reliable indication of foreign origin.

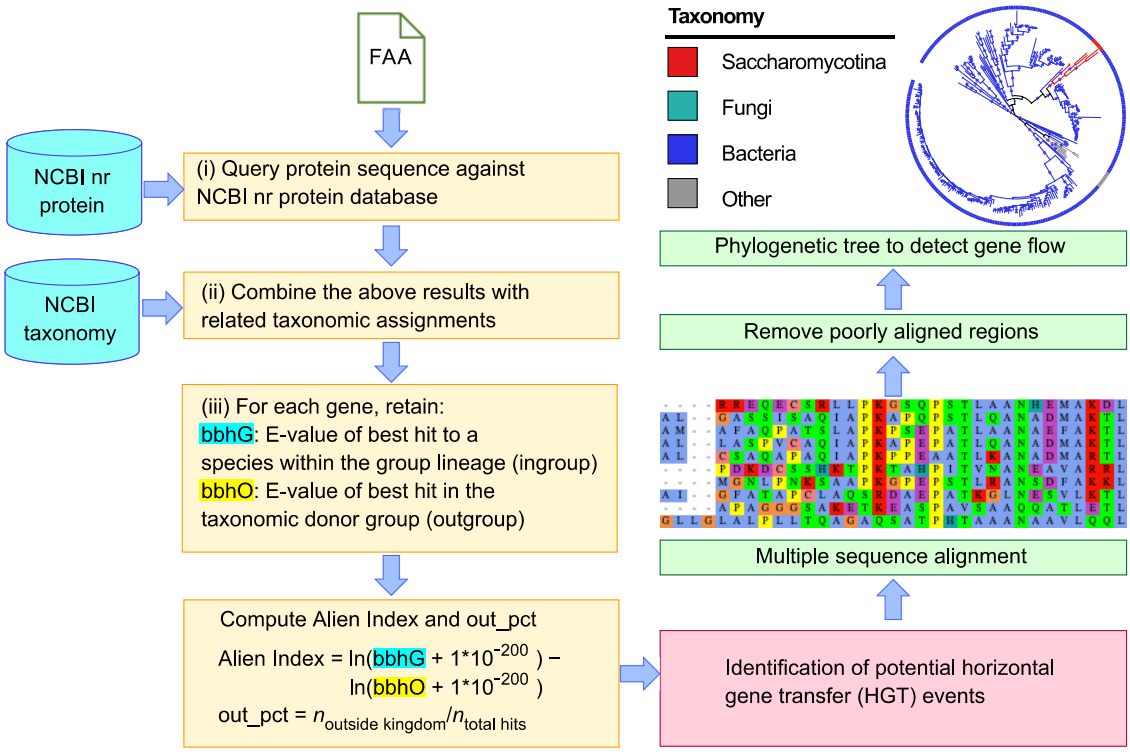

Figure 18 Overview of the HGTphyloDetect workflow for automated detection of HGT events from distantly related organisms (e.g., prokaryotes to eukaryotes)

Furthermore, in order to eliminate erroneous results in the detection of HGT, HGTphyloDetect computes the percentage of outgroup hits with distinct taxonomic species names for each gene, as illustrated in **Figure 18**. Finally, genes that meet both the

criteria of an Alien Index value ⩾ 45 and out_pct ⩾ 90% are considered probable HGT events from distantly related species [35, 123]. The threshold value for out_pct was established based on a significant study by Shen et al., in which the parameter was assessed for its effectiveness in eliminating erroneous HGT events [35]. With HGTphyloDetect, users can easily adjust their predictions by defining customized values for the AI and out_pct parameters, in addition to using the default settings.

3.1.2 Detecting horizontal gene transfer from closely related organisms

While the above-described workflow is effective in detecting HGT events from organisms that are distantly related evolutionarily, a complementary workflow has been developed for the automated detection of HGT events from more closely related organisms (**Figure 19**), such as eukaryote-to-eukaryote transfers. This expands the versatility of the computational toolbox to include a wider range of HGT detection capabilities. Unlike the previous workflow that required defining the recipient, ingroup, and outgroup lineage, it is difficult to define them in detecting HGT events from closely related organisms if using the same approach.

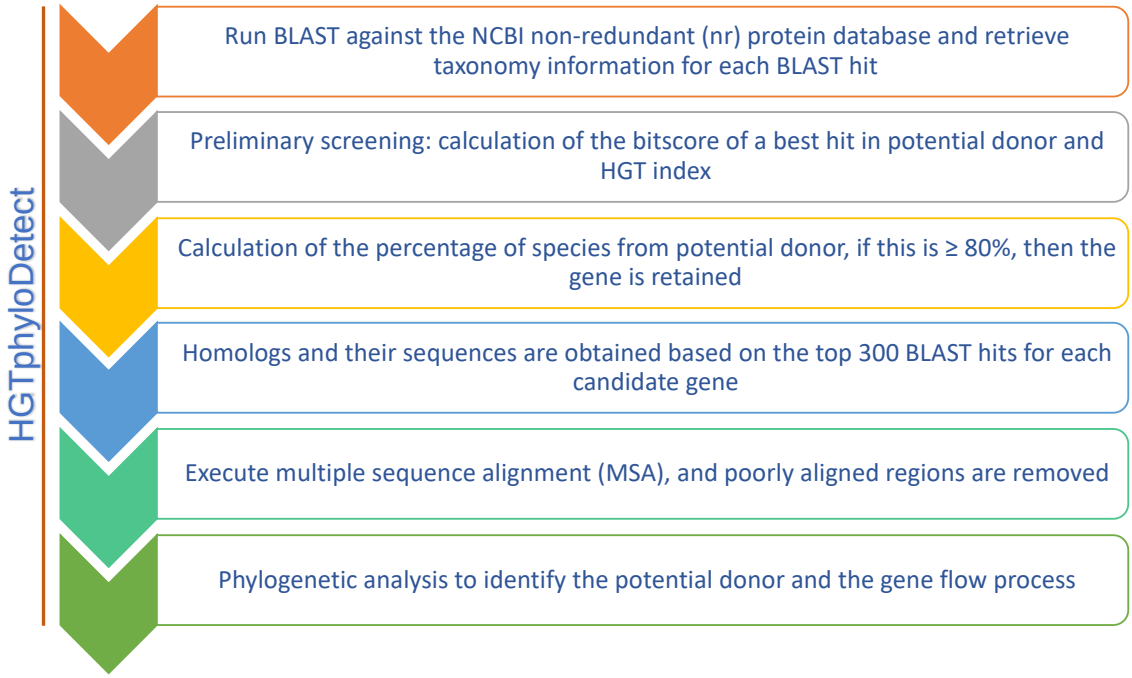

Figure 19 HGTphyloDetect workflow for automated detection of HGT events from closely related organisms (e.g., eukaryote-to-eukaryote transfers).

The workflow depicted in **Figure 19** involves several steps aiming at identifying potential horizontally acquired genes acquired from closely related organisms. Firstly, a set of genes is used as input for a BLASTP process against the NCBI nr protein database, with taxonomic information retrieved for each gene hit. In the first round of screening, genes with a best hit in the kingdom lineage (excluding the recipient subphylum lineage) and a bitscore of at least 100 are selected. The HGT index (also known as comparative similarity

index) is then calculated by dividing the bitscore of the best hit in a potential donor (i.e., a species inside the kingdom but outside the subphylum) by the bitscore of the best hit in the recipient (i.e., a species inside the subphylum). Genes with an HGT index of at least 50% are retained, indicating a strong match to genes in potential donors. Next, for each gene, the percentage of species from potential donors that have distinct taxonomic species names is determined. If this value is greater than or equal to 80%, the gene is retained. These threshold values were selected based on previous studies [36, 124, 125], but users may adjust these parameters as needed to optimize their analysis. Finally, the remaining genes are considered to be horizontally acquired genes from closely related organisms.

3.1.3 Basic usage and applications of the HGTphyloDetect toolbox

HGTphyloDetect, an open-source and user-friendly tool, can be downloaded from https://github.com/SysBioChalmers/HGTphyloDetect. Using the tool is straightforward - users simply need to provide a FASTA file containing both the protein identifier and sequence as input. By accessing the large NCBI nr protein and taxonomy databases remotely on demand, HGTphyloDetect eliminates the need to download these large databases locally. The installation process is simple and requires only a few dependencies. Moreover, a detailed user tutorial is provided in the GitHub repository to help users navigate the tool with ease.

Additionally, HGTphyloDetect is a versatile tool for HGT detection, offering users the ability to adjust parameter threshold values to customize their analyses. With a user-friendly example, the tool demonstrates its ability to identify potential donors and HGT events for one gene or all genes in a single species, or even for hundreds of species. The output of HGTphyloDetect provides detailed information on the potential donors of horizontally transferred genes, which can shed light on the evolutionary history of the organisms being studied. HGTphyloDetect is not limited to prokaryotes; it can also be applied to eukaryotes, enabling large-scale genome wide HGT analyses in both types of genomes. This scalability allows HGTphyloDetect to be seamlessly integrated into larger analytical workflows, making it a flexible and valuable tool for HGT detection in diverse research contexts.

3.1.4 Testing the performance of the HGTphyloDetect toolbox

To assess the predictive capability of the HGTphyloDetect tool, this toolbox was applied to two species (*S. cerevisiae* and *C. versatilis*) that have been previously shown to have horizontally acquired genes in manually curated studies [35, 126]. This benchmark evaluation allows for a comparison of the performance of HGTphyloDetect with previously published approaches.

Previous studies have identified 10 horizontally acquired genes transferred from bacteria in *S. cerevisiae* [126]. To comprehensively detect HGT events in *S. cerevisiae*, HGTphyloDetect was then applied to analyze all 6,000+ genes using default parameters.

As a result, 23 HGT gene candidates from bacteria were predicted, of which 8 candidates were previously reported (**Figure 20A**): YNR058W (BIO3), YDR540C, YJL217W, YKL216W (URA1), YFR055W, YOL164W (BDS1), YMR090W, and YNR057C (BIO4). The remaining 15 genes identified by HGTphyloDetect were not previously associated with HGT, indicating that they may have been overlooked in previous studies due to a lack of sufficient data and appropriate computational methods. However, HGTphyloDetect provided strong evidence for their bacterial origin, as demonstrated by their Alien Index values, out_pct, and E-values. Moreover, HGTphyloDetect demonstrated a high degree of accuracy in identifying HGT events in *S. cerevisiae* (**Figure 20A**).

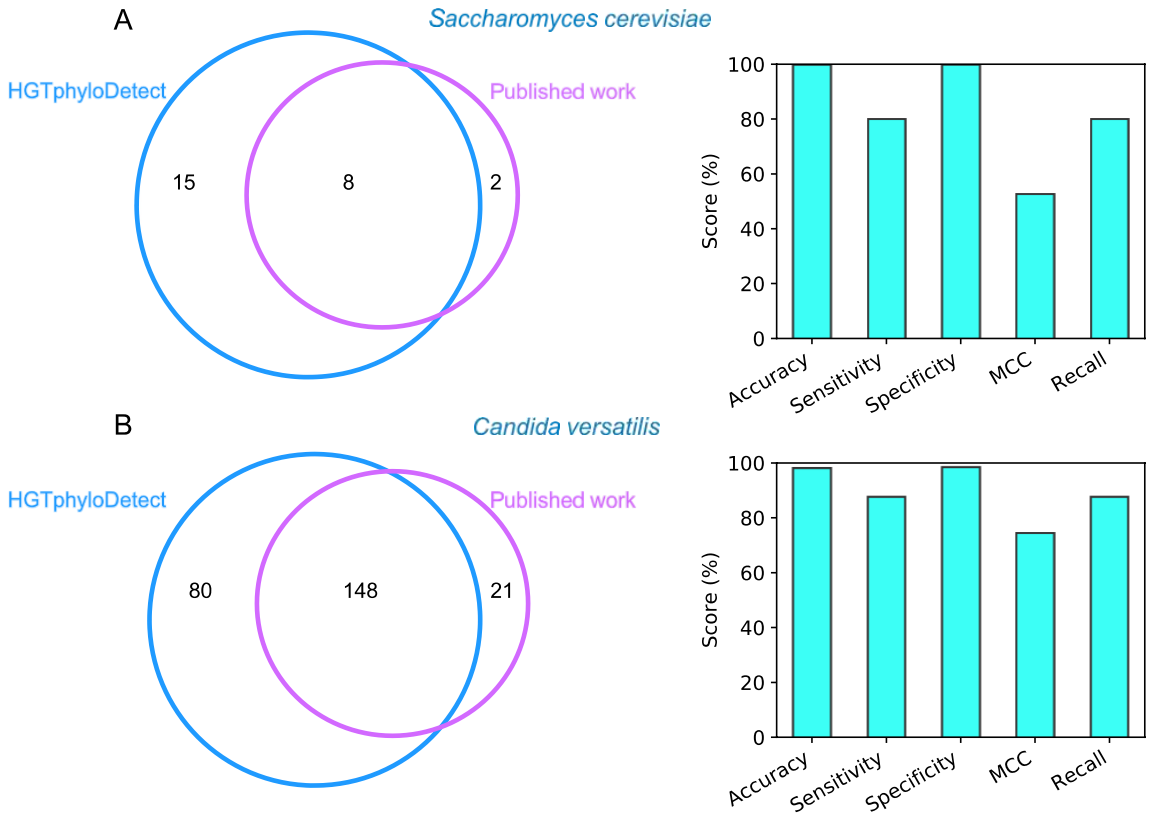

Figure 20 Evaluation of the HGTphyloDetect computational toolbox by two case studies. (A) Comparison of the number of horizontally acquired genes in *S. cerevisiae* as identified by HGTphyloDetect with those reported by previously published work. (B) Comparison of the number of horizontally acquired genes in *C. versatilis* as identified by HGTphyloDetect with those reported by previously published work.

Also, HGTphyloDetect revealed that 27 genes in *S. cerevisiae* were possibly obtained through horizontal transfer from fungal species that are more closely related to it. It is noteworthy that, among the 27 genes identified, only five were anticipated to have originated from the Taphrinomycotina, Ustilaginomycotina, and Agaricomycotina lineages, while the remaining 22 genes were probably acquired through horizontal transfer from the Pezizomycotina subphylum, which comprises numerous filamentous species, further suggesting that these species from the Pezizomycotina subphylum have relatively close interaction with *S. cerevisiae*. It should also be highlighted that this study has

systematically predicted HGT events between eukaryotes in the widely studied *S. cerevisiae*. This is quite meaningful since existing computational tools for detecting HGT events were unable to achieve this task.

Regarding *C. versatilis*, it has been reported to have a larger number of horizontally acquired genes (169 in total) [35], making it a suitable candidate for further testing of HGTphyloDetect. To evaluate this, the high-throughput pipeline was applied to all of the genes in *C. versatilis* (over 5,000 genes in total) using default parameters. This analysis identified that 148 out of the 169 genes in *C. versatilis* were horizontally acquired (**Figure 20B).** Afterward, HGTphyloDetect's prediction performance was evaluated using various standard evaluation metrics such as sensitivity, specificity, and accuracy, which were based on true positive, true negative, false positive, and false negative. For example, true positive refers to the situation where HGTphyloDetect correctly predicted a horizontally acquired gene that had been previously curated as such in peer-reviewed literature. Upon calculation, HGTphyloDetect's accuracy, sensitivity, and specificity were found to be 98.16%, 87.57%, and 98.49%, respectively (**Figure 20B**). This outcome demonstrates that HGTphyloDetect accurately predicts HGT gene candidates in *C. versatilis*, with high-quality performance that matches well with previous reports in the literature.

3.1.5 Comparison with other existing approaches for HGT detection

HGTphyloDetect was further compared with other existing computational tools, such as the HGTector toolbox [119], which is also capable of detecting HGT events in a high-throughput manner.

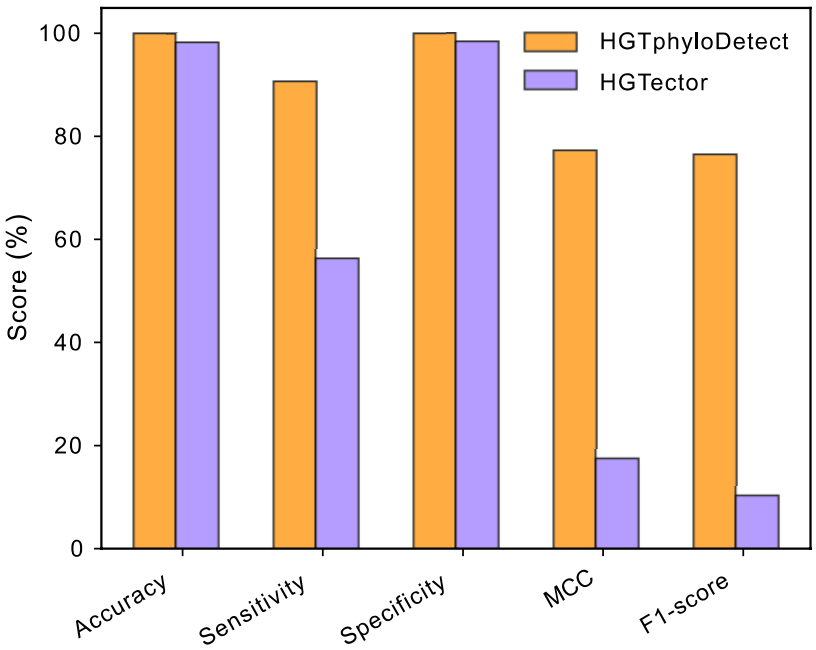

Figure 21 Comparison of the HGT detection performance between HGTphyloDetect and HGTector.

To explore this, I employed the benchmark dataset previously published by the Rokas group [35]. This dataset contains a comprehensive analysis of HGT events across a wide range of over 300 yeast species, which were systematically evaluated and manually inspected. Due to the computational requirements of HGT identification, three yeast species (*Lipomyces kononenkoae*, *Kluyveromyces lactis*, *Lachancea fermentati*) were randomly selected from this dataset for which HGT events had been identified, totaling over 15,000 unique genes. HGT detection workflows in HGTphyloDetect and HGTector were then executed on all these genes, and their performance was evaluated using various metrics, including accuracy, sensitivity, specificity, and others. The final results revealed that HGTphyloDetect outperformed HGTector in terms of accuracy and specificity (as shown in **Figure 21**). However, the most notable difference was the considerable increase in sensitivity, Matthews correlation coefficient (MCC), and F1-score achieved by HGTphyloDetect compared to HGTector (**Figure 21**). As sensitivity means how much HGT events can be detected by the software in this case, this indicates the great power of HGTphyloDetect in identifying HGT events.

3.1.6 Phylogenetic analysis via HGTphyloDetect

The most precise and widely accepted method for identifying horizontally acquired genes is gene-by-gene phylogenetic analysis [127]. This approach involves comparing the phylogeny of the target gene with similar genes from other species. Therefore, for additional validation of the HGT gene identification by HGTphyloDetect, a phylogenetic analysis pipeline was integrated into the whole workflow in detecting HGT events from phylogenetically distant organisms or closely related organisms. Firstly, the query genes of great interest were subjected to BLASTP against the NCBI nr protein database, and for each gene, the top 300 homologs with different taxonomic species names were selected. These homologs were aligned using MAFFT v7.310 [128] with default settings for multiple sequence alignment (MSA), and any ambiguously aligned regions were removed using the '-automated1' option of trimAl v1.4 [129]. To ensure the reliability and high quality of the resulting phylogenetic trees, the alignments were used to construct the trees using IQ-TREE v1.6.12 [130] with 1000 ultrafast bootstrapping replicates. The internal branch bootstrap scores were calculated based on IQ-TREE v1.6.12. Next, each phylogenetic tree was then rooted at the midpoint with the help of R packages: ape v5.4-1 [131] and phangorn v2.5.5 [132]. Finally, iTol v5 (https://itol.embl.de/) [127] was utilized to visualize the resulting phylogenies and evaluate the mode of transmission of each gene.

As an example of the phylogenetic analysis conducted with the help of HGTphyloDetect, the maximum likelihood phylogeny of YOL164W in *S. cerevisiae* was examined in this study. This protein is thought to have acquired alkyl sulfatase and arylsulfatase activity through HGT [127]. In order to gain insights into the evolutionary history of YOL164W and its possible origins, the wrapped pipeline in HGTphyloDetect as shown above was utilized to construct a detailed phylogenetic tree. This involved using the protein sequence

of YOL164W as a query to obtain the top homolog hits, which were then used to reconstruct the maximum likelihood phylogeny. The resulting tree, generated with ease, was reliable and of high quality (**Figure 22A**), clearly suggesting that the protein was horizontally acquired from a bacterial species. By examining the pruned phylogenetic tree, it was feasible to identify the bacterial donor and explore the phylogenetic relationship between this protein and its close relatives from proteobacteria. Notably, all internal branches proximal to the query protein had bootstrap scores exceeding 95%, underscoring the accuracy of the HGT event detection by HGTphyloDetect (**Figure 22B**). The case study presented here demonstrates the utility of the phylogenetic analysis with HGTphyloDetect in elucidating the mechanism of gene transfer for suspected HGT events.

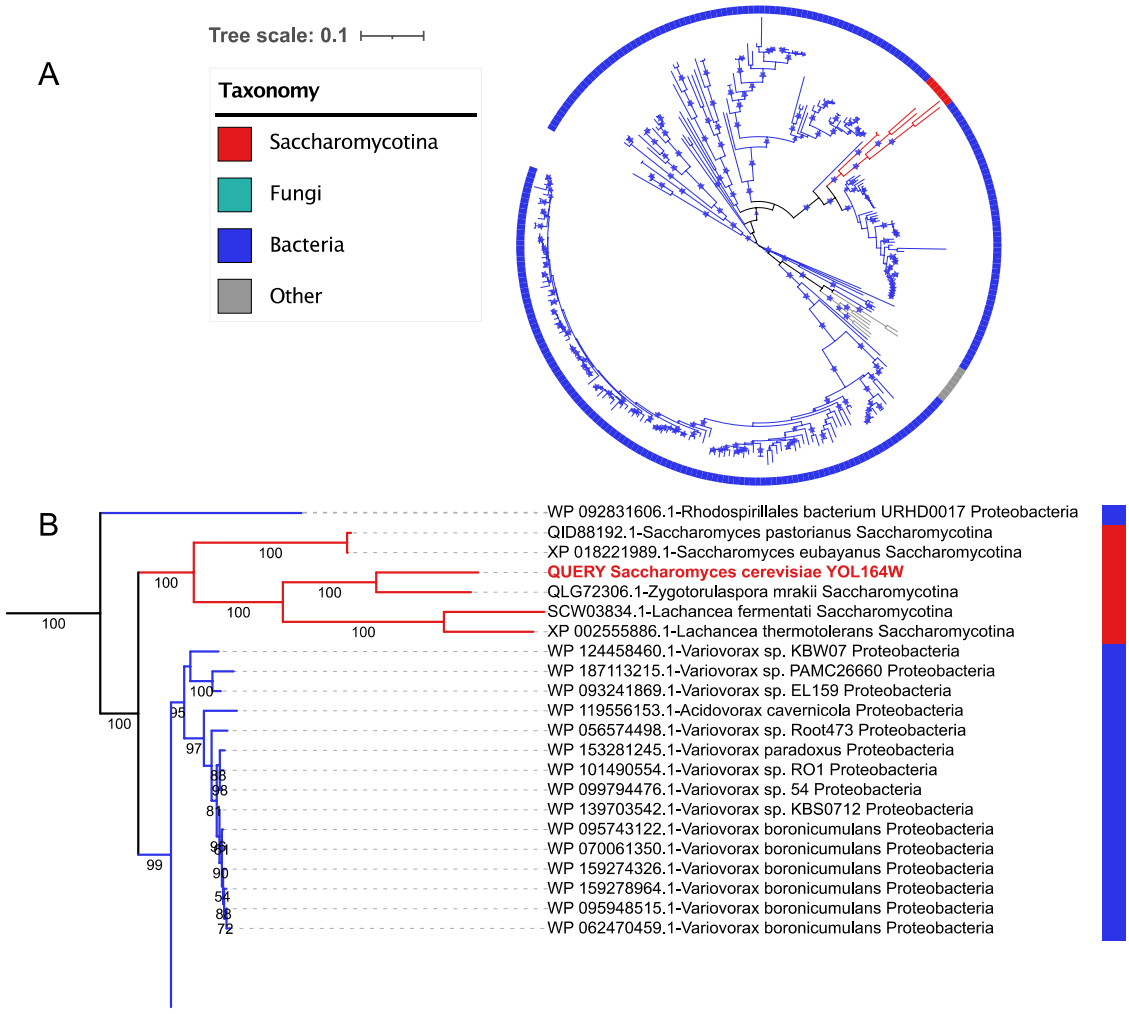

Figure 22 Phylogenetic analysis example of an HGT event from prokaryote to eukaryote via HGTphyloDetect. (A) The maximum likelihood phylogeny of a protein YOL164W in *S. cerevisiae*, with branches having bootstrap support greater than 80% indicated by a star. (B) A pruned maximum-likelihood phylogeny showing the relationship between this protein and its closely related homologs from other bacterial species, providing evidence for a prokaryotic origin of the HGT gene.

3.2 Substrate utilization analysis on large-scale yeast species (Paper I)

Substrate utilization in yeast refers to the process of metabolizing and utilizing various sources for energy and growth. Yeasts are known for their versatile substrate utilization abilities, allowing them to survive and thrive in a wide range of environments, including soil, plant surfaces, and animal tissues. Yeast cells utilize different substrates through a complex network of genes and pathways, which are responsive to changes in the environment and nutritional conditions [133]. Thus, understanding the mechanism of substrate utilization in yeasts is of great significance.

In this study (**Paper I**), the evolutionary mechanisms that underlie the trait diversity in substrate utilization across 332 yeast species were explored. This was accomplished by combining several analytical approaches, including HGT analysis (the pipeline has been shown in the above section), gene family expansion and contraction analysis, and GEMs simulations. These analyses were used to identify the genetic and metabolic features associated with substrate utilization in yeast, providing insights into the molecular and evolutionary mechanisms that underlie this trait.

3.2.1 Gain of new traits in substrate utilization occurring in yeast species

To investigate the mechanism of how yeasts gain new traits, the experimental evidence on substrate usage for 332 yeast species was firstly obtained from various literature sources [35, 134]. The substrate utilization dataset of each species was then compared with their ancestral budding yeast common ancestor (BYCA) phenotype to determine the number of gain and loss events in substrate utilization. The posterior probability of ancestral state in BYCA for each metabolic trait was obtained from a previous study [35]. For this analysis, a posterior probability of 0.85 was used as a cut-off for the threshold to indicate the existence of a phenotype in BYCA, while a probability lower than 0.15 was interpreted as non-existence. As a result, among the 32 traits in substrate utilization that could be linked to the metabolites in GEMs, five traits exhibited a gain of new function in utilizing carbon (2-Keto-D-gluconate, D-arabinose, D-ribose, methanol) and nitrogen (nitrite) sources.

3.2.2 Gene family expansion and contraction analysis

Gene family expansion and contraction analysis across large-scale yeast species were investigated using CAFÉ v4.2.1 [135] with default parameters. The software CAFÉ uses a birth and death process to model the evolution of gene family sizes by a phylogenetic tree, in which gene family sizes were obtained by a customized script based on the ortholog group (OG) defined in a previous study [35]. CAFÉ generated a family-wide *P* value along specific species or branches for each gene family, with a *P* value below 0.05 considered statistically significant, indicating a possible gene family expansion and contraction event. By analyzing the number of gene families that have undergone expansion and contraction at the species level and the clade level, it was demonstrated that there was a higher likelihood of gene family contraction rather than expansion across various yeast species

(**Figure 23A-B**), which aligns with a previous finding that suggests reductive evolution is the predominant mode of evolutionary diversification [35].

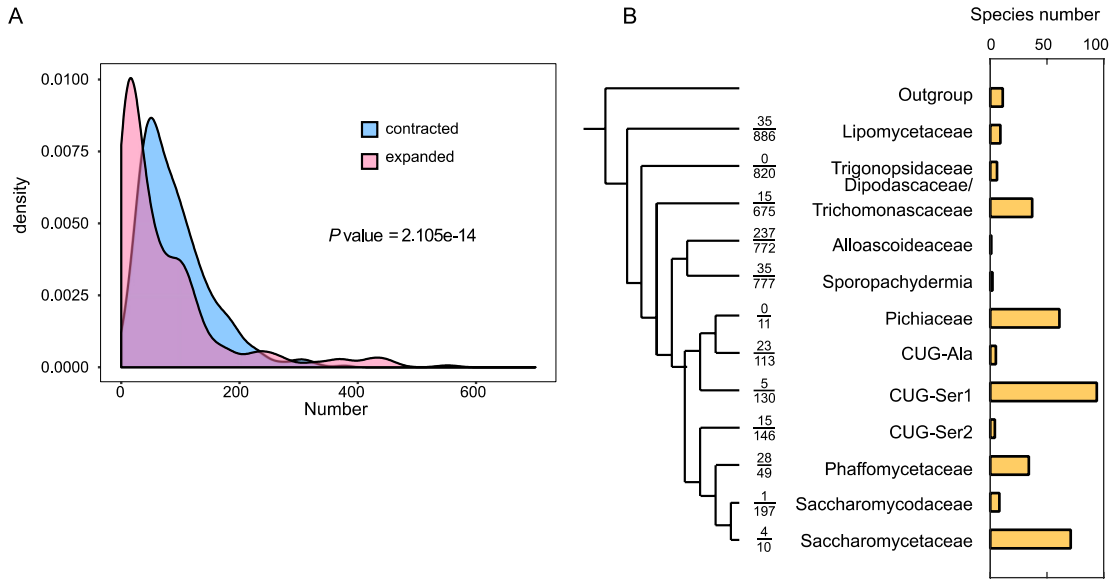

Figure 23 Gene family expansion and contraction analysis across large-scale yeast species. (A) Comparison between the number of gene families that have undergone expansion and contraction at the species level. (B) The number of gene families that have experienced expansion (upper number) and contraction (bottom number) within each clade.

3.2.3 Evolutionary mechanisms underlying the trait diversity in substrate utilization

To investigate the underlying mechanisms associated with substrate utilization, the substrate utilization of each species was compared with the inferred traits of the BYCA [35], and the gains and losses of these metabolic traits were identified (**Figure 24A**). Next, I performed systematic evolution analyses at the gene level and the clade level, including HGT analysis, gene family expansion and contraction analysis, as described in the preceding section. Subsequently, for each change in substrate utilization, the reason was explored to know whether this was brought about by HGT, expansion of a gene family, or a promiscuous enzyme. Finally, the results suggested that HGT plays a relatively minor role in the gains or losses of metabolic traits (**Figure 24A**), and many genes acquired from HGT events were found to be transporters or extracellular substrate degradation enzymes (**Figure 24B**), suggesting that they may have contributed to the expansion of substrate usage in yeast.

Besides, the main donor of HGT events is other fungi, rather than bacterial species (**Figure 24C**), indicating that there is a more frequent gene flow between these yeast species and other fungal species. Also, there exist obvious differences in these HGT events involved in substrate usage among various clades (**Figure 24A**). In the Wickerhamiella/Starmerella clade and its phylogenetically close relatives, such as Dipodascaceae/ Trichomonascaceae, Trigonopsidaceae and Lipomycetaceae, there are relatively more HGT events. This is likely due to the fact that the vast majority of species in these clades are ecologically

associated with other fungal species or eukaryotes [136]. On the other hand, there are very few or even zero HGT events related to substrate usage in the CUG group and its relatives (e.g., CUG-Ser1, CUG-Ser2, Phaffomycetaceae), indicating that genetic code alteration may act as a barrier to HGT [137].

By further analyzing the evolutionary events involved in substrate gain and substrate loss at a holistic level, it was found that the expansion of gene families and promiscuous enzymes are predominant factors driving substrate gain (**Figure 24D**). In contrast, highly correlated and non-highly correlated reactions were found to be the primary driving force of substrate loss (**Figure 24E**). Highly correlated reactions indicate consistency between the presence of a reaction and the phenotype of substrate utilization, while non-highly correlated reactions may not show the same consistency. Interestingly, non-highly correlated reactions had a strong effect on substrate loss, suggesting that the loss of metabolic traits does not always correspond to the loss of the same reactions across various yeast species.

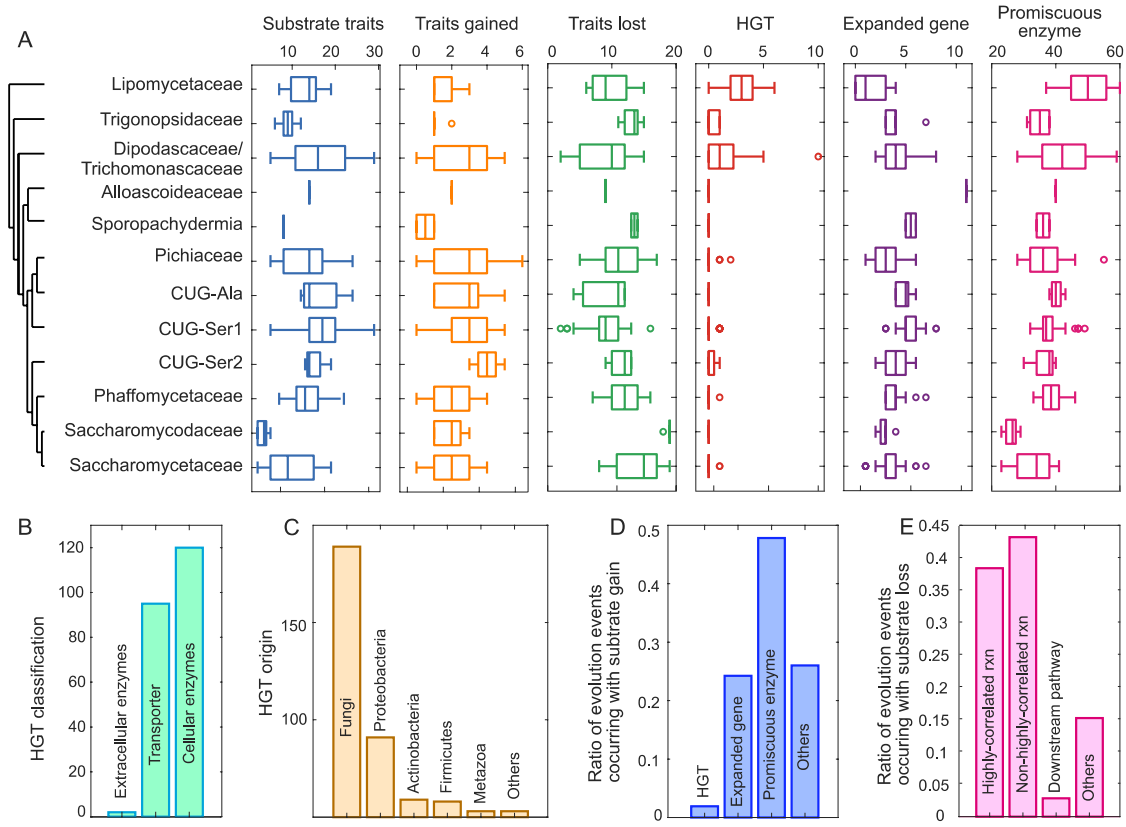

Figure 24 Exploration of the evolutionary mechanisms underlying the trait diversity in substrate utilization across 332 yeast species. (A) Comparison of the number of substrate traits, gain and loss of substrate utilization, as well as HGT events associated with substrate utilization, expanded genes associated with substrate utilization, and promiscuous enzymes in the substrate utilization pathway. The y-axis shows the genus-level phylogeny for 332 yeast species, divided into 12 major clades. (B) HGT classification for those genes associated with substrate utilization based on compartmental annotation. (C) Classification of donors for those HGT events in substrate utilization. (D) Proportions of various evolutionary processes linked to the gain of substrate utilization. (E) Ratios of evolutionary events occurring in substrate loss. Highly correlated and non-highly correlated reactions refer to whether it is consistent between reaction existence with the substrate utilization phenotype existence. Downstream pathway is defined as when all enzymes and reactions are included in the original substrate utilization pathway, but specific enzymes are absent in a distantly related pathway.

4. Conclusions

In this thesis, I have integrated state-of-the-art ML and DL approaches to enhance the understanding of yeast species in various dimensions, i.e., essential genes, k_{cat} and protein production. In the first study (**Paper I**), I developed ML approaches to predict gene essentiality and applied the tool to 343 yeast/fungi species. Interestingly, evolution-based features were found to be important factors that can substantially improve essential gene prediction. In the second study (**Paper II**), I constructed a high-quality DL model named DLKcat for the prediction of k_{cat} by combining a GNN for substrates and a CNN for proteins. This model can potentially identify amino acid residues that may have a more influential effect on enzyme activity. Since k_{cat} is an important enzyme kinetics parameter in reconstructing ecGEMs, DLKcat was further applied for the reconstruction of ecGEMs for 332 yeast species, enabling the elucidation of cellular metabolism systematically. In the third study (**Paper III**), ML was employed to explore the feature importance on recombinant protein production, and it was observed that PTMs can have a greater impact than amino acid compositions.

Furthermore, I utilized comparative genomics techniques to investigate the evolution of yeast species. In **Paper IV**, I developed a novel computational tool called HGTphyloDetect for the identification of HGT events by integrating phylogenetic analysis. HGTphyloDetect can be used to identify HGT events from both phylogenetically distant and closely related species. Case studies on *S. cerevisiae* and *C. versatilis* indicate the high accuracy of HGTphyloDetect in the detection of HGT events. Additionally, HGTphyloDetect allows users to explore the gene flow process with the aid of phylogenetic analysis. Last but not least, I investigated the underlying mechanisms for substrate utilization by performing a comprehensive analysis on large-scale yeast species (**Paper I**). It was found that gene family expansion and enzyme promiscuity are prominent mechanisms for substrate trait gains, while HGT plays a relatively minor role in substrate gains.

Taken together, the ML and comparative genomics tools and techniques implemented in this study represent a significant contribution to the development of yeast systems biology. These findings are not only valuable for the yeast community, but also have the potential for broader applications in biotechnology.

5. Future perspectives

In the first part of my thesis, I utilized ML approaches to predict essential genes by integrating sequence features and evolution-based features. However, the gene essentiality data used for model training only included information from five species, which limited the dataset and impacted the model's performance. Currently, the model's AUC value on the testing dataset is around 0.8, indicating that there is still room for improvement. To improve the model's performance, I plan to explore advanced DL frameworks such as CNN, transformers, etc. These frameworks can help capture complex features or embeddings that may be missed by traditional ML algorithms adopted in the study. Additionally, to improve the performance of the model, I intend to acquire more data from various sources to increase the dataset's size and diversity. By incorporating data from more species and optimizing the model's architecture, I believe that the prediction performance of the model could be further enhanced. Furthermore, the hypothesis that evolutionary information is beneficial for essential gene prediction, which was proposed and validated in this work, may also be applicable to other gene-related or enzyme-related problems, such as enzyme affinity prediction.

DLKcat is a powerful approach for estimating k_{cat} values based on DL that is further used to reconstruct ecGEMs for more than 300 yeast species [101]. In addition to predicting k_{cat} values, DLKcat has the advantage of calculating attention weights derived from the neural network to identify sequence residues that have an influential effect on enzyme catalytic activity. Although DLKcat performs well in k_{cat} prediction, challenges still remain, such as not considering environmental conditions like temperature and pH. However, combining DLKcat with other emerging ML tools, such as the model for predicting enzyme optimal temperature [82], can enable the investigation of the impact of environmental parameters on enzyme activities. As more and more experimental data on k_{cat} values become available, the DL model can be retrained to improve the performance. The future version of DLKcat could further incorporate representations obtained from protein 3D structures with the aid of AlphaFold [138], which would enhance the model's interpretability and may improve its performance. Additionally, integrating pre-trained language models in the future version of DLKcat may also contribute to improving the model's performance.

With the rapidly increasing amount of newly sequenced genome data, HGTphyloDetect has proven to be an effective toolbox that can meet the growing demand for biological applications across various fields. It can be used to help interpret pathogen phenotypes in fungi [139], analyze antibiotic resistance determinants in bacteria [140], and explore new functionalities in the gut microbiome [141]. With HGTphyloDetect, it is now possible to investigate these different phenotypes on a large scale and determine which genes are likely acquired through HGT, and whether these HGT genes are involved in the generation of these important phenotypes. Although HGTphyloDetect is already enabling novel

analyses, there is still room for improvement in terms of performance. Enhancements to computation speed could make it even more suitable for large-scale analyses in the future.

ML has become increasingly popular in biology due to its ability to analyze large datasets and extract meaningful insights that may not be apparent through other traditional analytical approaches. However, there are several challenges that ML may not be able to solve. One of the major challenges is the limited availability of large and diverse datasets, which are essential for training accurate models. Without a sufficient amount of data, ML models may not be able to generate accurate or reliable predictions. For another thing, ML models are often seen as black boxes, making it difficult to understand how the models make predictions. This can limit the interpretability of the results and hinder further investigations into the underlying mechanisms behind the predictions. Although ML has demonstrated its great potential in biology, it is not a panacea for all biological problems. ML is a powerful tool that can help generate hypotheses and facilitate data-driven discoveries, but it cannot replace the need for experiments. The combination of experimental data with ML-based analyses can help accelerate progress in the field of systems biology and synthetic biology.

References

1. Hanczyc MM. Engineering life: A review of synthetic biology, Artificial life 2020;26:260-273.
2. Zhang X-E, Liu C, Dai J et al. Enabling technology and core theory of synthetic biology, Science China Life Sciences 2023:1-44.
3. Stephanopoulos G, Sinskey AJ. Metabolic engineering-methodologies and future prospects, Trends Biotechnol 1993;11:392-396.
4. Nielsen J, Keasling JD. Engineering cellular metabolism, Cell 2016;164:1185-1197.
5. Nielsen J. Metabolic engineering, Applied microbiology and biotechnology 2001;55:263-283.
6. Stephanopoulos G, Aristidou AA, Nielsen J. Metabolic engineering: principles and methodologies 1998.
7. Keasling JD. Manufacturing molecules through metabolic engineering, Science 2010;330:1355-1358.
8. Lin Y, Jain R, Yan Y. Microbial production of antioxidant food ingredients via metabolic engineering, Curr Opin Biotechnol 2014;26:71-78.
9. Ofosu FK, Daliri EB-M, Elahi F et al. New insights on the use of polyphenols as natural preservatives and their emerging safety concerns, Frontiers in Sustainable Food Systems 2020;4:525810.
10. Chartrain M, Salmon PM, Robinson DK et al. Metabolic engineering and directed evolution for the production of pharmaceuticals, Curr Opin Biotechnol 2000;11:209-214.
11. Zhou YJ, Kerkhoven EJ, Nielsen J. Barriers and opportunities in bio-based production of hydrocarbons, Nature Energy 2018;3:925-935.
12. Zhou YJ, Buijs NA, Zhu Z et al. Production of fatty acid-derived oleochemicals and biofuels by synthetic yeast cell factories, Nat Commun 2016;7:11709.
13. Ingram L, Gomez P, Lai X et al. Metabolic engineering of bacteria for ethanol production, Biotechnology and bioengineering 1998;58:204-214.
14. Jojima T, Noburyu R, Sasaki M et al. Metabolic engineering for improved production of ethanol by *Corynebacterium glutamicum*, Applied microbiology and biotechnology 2015;99:1165-1172.
15. Achimón F, Areco VA, Brito VD et al. Plants as Bioreactors for the Production of Biopesticides, Plants as Bioreactors for Industrial Molecules 2023:337-366.
16. Ke J, Wang B, Yoshikuni Y. Microbiome engineering: synthetic biology of plant-associated microbiomes in sustainable agriculture, Trends Biotechnol 2021;39:244-261.
17. Wang Y, Yin J, Chen G-Q. Polyhydroxyalkanoates, challenges and opportunities, Curr Opin Biotechnol 2014;30:59-65.
18. Zhang X, Lin Y, Wu Q et al. Synthetic biology and genome-editing tools for improving PHA metabolic engineering, Trends Biotechnol 2020;38:689-700.
19. Tan D, Wang Y, Tong Y et al. Grand challenges for industrializing polyhydroxyalkanoates (PHAs), Trends Biotechnol 2021;39:953-963.

20. French K. Harnessing synthetic biology for sustainable development, Nature Sustainability 2019;2:250-252.
21. Nielsen J, Jewett MC. Impact of systems biology on metabolic engineering of *Saccharomyces cerevisiae*, FEMS Yeast Res 2008;8:122-131.
22. Kitano H. Systems biology: a brief overview, Science 2002;295:1662-1664.
23. Yu R, Nielsen J. Big data in yeast systems biology, FEMS Yeast Res 2019;19:foz070.
24. Kirschner MW. The meaning of systems biology, Cell 2005;121:503-504.
25. Mardinoglu A, Nielsen J. Systems medicine and metabolic modelling, J Intern Med 2012;271:142-154.
26. Thak EJ, Yoo SJ, Moon HY et al. Yeast synthetic biology for designed cell factories producing secretory recombinant proteins, FEMS Yeast Res 2020;20:foaa009.
27. Botstein D, Fink GR. Yeast: an experimental organism for 21st century biology, Genetics 2011;189:695-704.
28. Smith MG, Snyder M. Yeast as a model for human disease, Current protocols in human genetics 2006;48:15.16. 11-15.16. 18.
29. Chen X, Ji B, Hao X et al. FMN reduces Amyloid-β toxicity in yeast by regulating redox status and cellular metabolism, Nat Commun 2020;11:867.
30. Lian J, Mishra S, Zhao H. Recent advances in metabolic engineering of *Saccharomyces cerevisiae*: new tools and their applications, Metab Eng 2018;50:85-108.
31. Nielsen J. Yeast systems biology: model organism and cell factory, Biotechnol J 2019;14:1800421.
32. Vieira Gomes AM, Souza Carmo T, Silva Carvalho L et al. Comparison of yeasts as hosts for recombinant protein production, Microorganisms 2018;6:38.
33. Manfrão-Netto JHC, Gomes AMV, Parachin NS. Advances in using Hansenula polymorpha as chassis for recombinant protein production, Frontiers in bioengineering and biotechnology 2019;7:94.
34. Peter J, De Chiara M, Friedrich A et al. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates, Nature 2018;556:339-344.
35. Shen X-X, Opulente DA, Kominek J et al. Tempo and mode of genome evolution in the budding yeast subphylum, Cell 2018;175:1533-1545. e1520.
36. Lu H, Li F, Yuan L et al. Yeast metabolic innovations emerged via expanded metabolic network and gene positive selection, Mol Syst Biol 2021;17:e10427.
37. Eldarov MA, Beletsky AV, Tanashchuk TN et al. Whole-genome analysis of three yeast strains used for production of sherry-like wines revealed genetic traits specific to flor yeasts, Front Microbiol 2018;9:965.
38. Dujon BA, Louis EJ. Genome diversity and evolution in the budding yeasts (Saccharomycotina), Genetics 2017;206:717-750.
39. Chen F, Yuan L, Ding S et al. Data-driven rational biosynthesis design: from molecules to cell factories, Brief Bioinform 2020;21:1238-1248.
40. Kim S, Chen J, Cheng T et al. PubChem in 2021: new data content and improved web interfaces, Nucleic acids research 2021;49:D1388-D1395.

41. Hastings J, Owen G, Dekker A et al. ChEBI in 2016: Improved services and an expanding collection of metabolites, Nucleic Acids Res. 2016;44:D1214-1219.
42. Wishart DS, Feunang YD, Guo AC et al. DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res. 2018;46:D1074-d1082.
43. Kanehisa M, Furumichi M, Sato Y et al. KEGG: integrating viruses and cellular organisms, Nucleic acids research 2021;49:D545-D551.
44. Bansal P, Morgat A, Axelsen KB et al. Rhea, the reaction knowledgebase in 2022, Nucleic acids research 2022;50:D693-D700.
45. Caspi R, Billington R, Keseler IM et al. The MetaCyc database of metabolic pathways and enzymes-a 2019 update, Nucleic acids research 2020;48:D445-D453.
46. Consortium U. UniProt: a worldwide hub of protein knowledge, Nucleic acids research 2019;47:D506-D515.
47. Burley SK, Berman HM, Bhikadiya C et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy, Nucleic acids research 2019;47:D464-D474.
48. Chang A, Jeske L, Ulbrich S et al. BRENDA, the ELIXIR core data resource in 2021: new developments and updates, Nucleic acids research 2021;49:D498-D508.
49. Wittig U, Rey M, Weidemann A et al. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics, Nucleic Acids Res. 2018;46:D656-d660.
50. Rudolph J, Tan S, Tan S. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?, Journal of Applied Learning and Teaching 2023;6.
51. Jin W. Research on machine learning and its algorithms and development. In: Journal of Physics: Conference Series. 2020, p. 012003. IOP Publishing.
52. Liu B, Liu B. Supervised learning. Springer, 2011.
53. Celebi ME, Aydin K. Unsupervised learning algorithms. Springer, 2016.
54. Neftci EO, Averbeck BB. Reinforcement learning in artificial and biological systems, Nature Machine Intelligence 2019;1:133-143.
55. Breiman L. Random forests, Machine learning 2001;45:5-32.
56. Noble WS. What is a support vector machine?, Nat Biotechnol 2006;24:1565-1567.
57. Almansour NA, Syed HF, Khayat NR et al. Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study, Computers in biology and medicine 2019;109:101-111.
58. Liu B, Li C-C, Yan K. DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks, Brief Bioinform 2020;21:1733-1741.
59. Boopathi V, Subramaniyam S, Malik A et al. mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides, Int J Mol Sci 2019;20:1964.
60. Zhang Z. Introduction to machine learning: k-nearest neighbors, Annals of translational medicine 2016;4.
61. LeCun Y, Bengio Y, Hinton G. Deep learning, Nature 2015;521:436-444.
62. O'Shea K, Nash R. An introduction to convolutional neural networks, arXiv preprint arXiv:1511.08458 2015.

63. Reiser P, Neubert M, Eberhard A et al. Graph neural networks for materials science and chemistry, Communications Materials 2022;3:93.
64. Wang Y, Wang J, Cao Z et al. Molecular contrastive learning of representations via graph neural networks, Nature Machine Intelligence 2022;4:279-287.
65. Xiong J, Xiong Z, Chen K et al. Graph neural networks for automated de novo drug design, Drug Discovery Today 2021;26:1382-1393.
66. Jha K, Saha S, Singh H. Prediction of protein-protein interaction using graph neural networks, Sci Rep 2022;12:8360.
67. Goodfellow I, Pouget-Abadie J, Mirza M et al. Generative adversarial networks, Communications of the ACM 2020;63:139-144.
68. Marsit S, Leducq J-B, Durand É et al. Evolutionary biology through the lens of budding yeast comparative genomics, Nature Reviews Genetics 2017;18:581-598.
69. Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS, PLoS genetics 2008;4:e1000304.
70. MacEachern S, McEwan J, McCulloch A et al. Molecular evolution of the Bovini tribe (Bovidae, Bovinae): Is there evidence of rapid evolution or reduced selective constraint in Domestic cattle?, BMC Genomics 2009;10:1-14.
71. Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life, Nature Reviews Genetics 2015;16:472-482.
72. Van Etten J, Bhattacharya D. Horizontal gene transfer in eukaryotes: not if, but how much?, Trends in Genetics 2020;36:915-925.
73. Fitzpatrick DA. Horizontal gene transfer in fungi, FEMS Microbiol Lett 2012;329:1-8.
74. Power JJ, Pinheiro F, Pompei S et al. Adaptive evolution of hybrid bacteria by horizontal gene transfer, Proc Natl Acad Sci U S A 2021;118:e2007873118.
75. Chen I, Dubnau D. DNA uptake during bacterial transformation, Nat Rev Microbiol 2004;2:241-249.
76. Hall RJ, Whelan FJ, McInerney JO et al. Horizontal gene transfer as a source of conflict and cooperation in prokaryotes, Front Microbiol 2020;11:1569.
77. Demuth JP, Bie TD, Stajich JE et al. The evolution of mammalian gene families, PLoS One 2006;1:e85.
78. Fajardo D, Saint Jean R, Lyons PJ. Acquisition of new function through gene duplication in the metallocarboxypeptidase family, Sci Rep 2023;13:2512.
79. Zhang Q-J, Zhu T, Xia E-H et al. Rapid diversification of five Oryza AA genomes associated with rice adaptation, Proc Natl Acad Sci U S A 2014;111:E4954-E4962.
80. Zrimec J, Börlin CS, Buric F et al. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure, Nat Commun 2020;11:6141.
81. Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers, Proc Natl Acad Sci U S A 2019;116:13996-14001.

82. Li G, Rabe KS, Nielsen J et al. Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima, ACS synthetic biology 2019;8:1411-1420.
83. Yu T, Cui H, Li JC et al. Enzyme function prediction using contrastive learning, Science 2023;379:1358-1363.
84. Deng J, Deng L, Su S et al. Investigating the predictability of essential genes across distantly related organisms using an integrative approach, Nucleic acids research 2011;39:795-807.
85. Zhang X, Xiao W, Xiao W. DeepHE: Accurately predicting human essential genes based on deep learning, PLoS Computational Biology 2020;16:e1008229.
86. Giaever G, Chu AM, Ni L et al. Functional profiling of the *Saccharomyces cerevisiae* genome, Nature 2002;418:387-391.
87. Cherry JM, Adler C, Ball C et al. SGD: Saccharomyces genome database, Nucleic acids research 1998;26:73-79.
88. Arnaud MB, Costanzo MC, Skrzypek MS et al. The Candida Genome Database (CGD), a community resource for *Candida albicans* gene and protein information, Nucleic Acids Research 2005;33:D358-D363.
89. Lock A, Rutherford K, Harris MA et al. PomBase 2018: user-driven reimplementation of the fission yeast database provides rapid and intuitive access to diverse, interconnected information, Nucleic acids research 2019;47:D821-D827.
90. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic acids research 2007;35:D61-D65.
91. Chen W-H, Minguez P, Lercher MJ et al. OGEE: an online gene essentiality database, Nucleic acids research 2012;40:D901-D906.
92. O'Meara TR, Veri AO, Ketela T et al. Global analysis of fungal morphology exposes mechanisms of host cell escape, Nat Commun 2015;6:1-10.
93. Cankorur-Cetinkaya A, Dikicioglu D, Oliver SG. Metabolic modeling to identify engineering targets for *Komagataella phaffii*: The effect of biomass composition on gene target identification, Biotechnology and bioengineering 2017;114:2605-2615.
94. Wei S, Jian X, Chen J et al. Reconstruction of genome-scale metabolic model of *Yarrowia lipolytica* and its application in overproduction of triacylglycerol, Bioresources and Bioprocessing 2017;4:1-9.
95. Ning L, Lin H, Ding H et al. Predicting bacterial essential genes using only sequence composition information, Genet. Mol. Res 2014;13:4564-4572.
96. Jordan IK, Rogozin IB, Wolf YI et al. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria, Genome Res 2002;12:962-968.
97. Capra JA, Singh M. Predicting functionally important residues from sequence conservation, Bioinformatics 2007;23:1875-1882.
98. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood, Molecular biology and evolution 2007;24:1586-1591.
99. Zhang Z, Ren Q. Why are essential genes essential?-The essentiality of Saccharomyces genes, Microbial Cell 2015;2:280.

100. Chen Z, Zhao P, Li F et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data, Brief Bioinform 2020;21:1047-1057.
101. Li F, Yuan L, Lu H et al. Deep learning-based k_{cat} prediction enables improved enzyme-constrained model reconstruction, Nat Catal 2022:1-11.
102. Chen Y, Nielsen J. Energy metabolism controls phenotypes by protein efficiency and allocation, Proc Natl Acad Sci U S A 2019;116:17592-17597.
103. Sánchez BJ, Zhang C, Nilsson A et al. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints, Mol Syst Biol 2017;13:935.
104. Heckmann D, Lloyd CJ, Mih N et al. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models, Nat Commun 2018;9:1-10.
105. Tsubaki M, Tomii K, Sese J. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences, Bioinformatics 2019;35:309-318.
106. Dong Q-W, Wang X-l, Lin L. Application of latent semantic analysis to protein remote homology detection, Bioinformatics 2006;22:285-290.
107. Bar-Even A, Noor E, Savir Y et al. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters, Biochemistry 2011;50:4402-4410.
108. Notebaart RA, Kintses B, Feist AM et al. Underground metabolism: network-level perspective and biotechnological potential, Curr Opin Biotechnol 2018;49:108-114.
109. Glasner ME, Truong DP, Morse BC. How enzyme promiscuity and horizontal gene transfer contribute to metabolic innovation, The FEBS journal 2020;287:1323-1342.
110. Notebaart RA, Szappanos B, Kintses B et al. Network-level architecture and the evolutionary potential of underground metabolism, Proc Natl Acad Sci U S A 2014;111:11762-11767.
111. Goh GB, Hodas NO, Vishnu A. Deep learning for computational chemistry, J Comput Chem 2017;38:1291-1307.
112. Chen Y, Nielsen J. Mathematical modeling of proteome constraints within metabolism, Current Opinion in Systems Biology 2021;25:50-56.
113. Domenzain I, Sánchez B, Anton M et al. Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0, Nat Commun 2022;13:3766.
114. Nam H, Lewis NE, Lerman JA et al. Network context and selection in the evolution to enzyme specificity, Science 2012;337:1101-1104.
115. Williams DC, Van Frank RM, Muth WL et al. Cytoplasmic inclusion bodies in *Escherichia coli* producing biosynthetic human insulin proteins, Science 1982;215:687-689.
116. Wang G, Huang M, Nielsen J. Exploring the potential of *Saccharomyces cerevisiae* for biopharmaceutical protein production, Curr Opin Biotechnol 2017;48:77-84.

117. Feizi A, Österlund T, Petranovic D et al. Genome-scale modeling of the protein secretory machinery in yeast, PLoS One 2013;8:e63284.
118. Lundberg SM, Erion G, Chen H et al. From local explanations to global understanding with explainable AI for trees, Nature machine intelligence 2020;2:56-67.
119. Zhu Q, Kosoy M, Dittmar K. HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers, BMC Genomics 2014;15:1-18.
120. Nguyen M, Ekstrom A, Li X et al. HGT-Finder: A new tool for horizontal gene transfer finding and application to Aspergillus genomes, Toxins (Basel) 2015;7:4035-4053.
121. Koutsovoulos GD, Granjeon Noriot S, Bailly-Bechet M et al. AvP: A software package for automatic phylogenetic detection of candidate horizontal gene transfers, PLOS Computational Biology 2022;18:e1010686.
122. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data, Molecular biology and evolution 2016;33:1635-1638.
123. Gladyshev EA, Meselson M, Arkhipova IR. Massive horizontal gene transfer in bdelloid rotifers, Science 2008;320:1210-1213.
124. Marcet-Houben M, Gabaldón T. Acquisition of prokaryotic genes by fungal genomes, Trends in Genetics 2010;26:5-8.
125. Crisp A, Boschetti C, Perry M et al. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes, Genome biology 2015;16:1-13.
126. Hall C, Brachat S, Dietrich FS. Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*, Eukaryotic Cell 2005;4:1102-1115.
127. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation, Nucleic acids research 2021;49:W293-W296.
128. Katoh K, Kuma K-i, Toh H et al. MAFFT version 5: improvement in accuracy of multiple sequence alignment, Nucleic acids research 2005;33:511-518.
129. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses, Bioinformatics 2009;25:1972-1973.
130. Nguyen L-T, Schmidt HA, Von Haeseler A et al. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies, Molecular biology and evolution 2015;32:268-274.
131. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language, Bioinformatics 2004;20:289-290.
132. Schliep KP. phangorn: phylogenetic analysis in R, Bioinformatics 2011;27:592-593.
133. Lu H, Li F, Sánchez BJ et al. A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism, Nat Commun 2019;10:1-13.
134. Opulente DA, Rollinson EJ, Bernick-Roehr C et al. Factors driving metabolic diversity in the budding yeast subphylum, BMC biology 2018;16:1-15.

135. Han MV, Thomas GW, Lugo-Martinez J et al. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3, Molecular biology and evolution 2013;30:1987-1997.
136. Gonçalves P, Gonçalves C, Brito PH et al. The Wickerhamiella/Starmerella clade—a treasure trove for the study of the evolution of yeast metabolism, Yeast 2020;37:313-320.
137. Richards TA, Leonard G, Soanes DM et al. Gene transfer into the fungi, Fungal Biol Rev 2011;25:98-110.
138. Jumper J, Evans R, Pritzel A et al. Highly accurate protein structure prediction with AlphaFold, Nature 2021;596:583-589.
139. Alexander WG, Wisecaver JH, Rokas A et al. Horizontally acquired genes in early-diverging pathogenic fungi enable the use of host nucleosides and nucleotides, Proc Natl Acad Sci U S A 2016;113:4116-4121.
140. Lehtinen S, Chewapreecha C, Lees J et al. Horizontal gene transfer rate is not the primary determinant of observed antibiotic resistance frequencies in Streptococcus pneumoniae, Science advances 2020;6:eaaz6137.
141. Groussin M, Poyet M, Sistiaga A et al. Elevated rates of horizontal gene transfer in the industrialized human microbiome, Cell 2021;184:2053-2067. e2018.