THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Application of Machine Learning in Systems Biology

GANG LI

Division of Systems & Synthetic Biology

Department of Biology and Biological Engineering

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2020

Application of Machine Learning in Systems Biology

Gang Li

ISBN 978-91-7905-290-4

Doktorsavhandlingar vid Chalmers tekniska högskola

Ny serie nr 4757

ISSN 0346-718X

Department of Biology and Biological Engineering

Chalmers University of Technology

SE-412 96 Göteborg

Sweden

Telephone + 46 (0)31-772 1000

Cover: Machine learning assisted mathematical modelling in systems biology

Printed by Chalmers Reproservice,

Gothenburg, Sweden 2020

Application of Machine Learning in Systems Biology

Gang Li

Department of Biology and Biological Engineering

Chalmers University of Technology

Abstract

Biological systems are composed of a large number of molecular components. Understanding their behavior as a result of the interactions between the individual components is one of the aims of systems biology. Computational modelling is a powerful tool commonly used in systems biology, which relies on mathematical models that capture the properties and interactions between molecular components to simulate the behavior of the whole system. However, in many biological systems, it becomes challenging to build reliable mathematical models due to the complexity and the poor understanding of the underlying mechanisms. With the breakthrough in big data technologies in biology, data-driven machine learning (ML) approaches offer a promising complement to traditional theory-based models in systems biology. Firstly, ML can be used to model the systems in which the relationships between the components and the system are too complex to be modelled with theory-based models. Two such examples of using ML to resolve the genotype-phenotype relationships are presented in this thesis: (i) predicting yeast phenotypes using genomic features and (ii) predicting the thermal niche of microorganisms based on the proteome features. Secondly, ML naturally complements theory-based models. By applying ML, I improved the performance of the genome-scale metabolic model in describing yeast thermotolerance. In this application, ML was used to estimate the thermal parameters by using a Bayesian statistical learning approach that trains regression models and performs uncertainty quantification and reduction. The predicted bottleneck genes were further validated by experiments in improving yeast thermotolerance.

In such applications, regression models are frequently used, and their performance relies on many factors, including but not limited to feature engineering and quality of response values. Manually engineering sufficient relevant features is particularly challenging in biology due to the lack of knowledge in certain areas. With the increasing volume of big data, deep-transfer learning enables us to learn a statistical summary of the samples from a big dataset which can be used as input to train other ML models. In the present thesis, I applied this approach to first learn a deep representation of enzyme thermal adaptation and then use it for the development of regression models for predicting enzyme optimal and protein melting temperatures. It was demonstrated that the transfer learning-based regression models outperform the classical ones trained on rationally engineered features in both cases. On the other hand, noisy response values are very common in biological datasets due to the variation in experimental measurements and they fundamentally restrict the performance attainable with regression models. I thereby addressed this challenge by deriving a theoretical upper bound for the coefficient of determination (R^2) for regression models. This theoretical upper bound depends on the noise associated with the response variable and variance for a given dataset. It can thus be used to test whether the maximal performance has been reached on a particular dataset, or whether further model improvement is possible.

Keywords: Machine learning, systems biology, genome-scale modelling, uncertainty, regression, deep transfer learning

List of publications

This thesis is based on the work contained in the following papers and manuscripts:

Paper I: The pan-genome of *Saccharomyces cerevisiae*

Li G, Ji B, and Nielsen J. FEMS Yeast Research. 19(7) (2019).

Paper II: Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima

Li G, Rabe KS, Nielsen J and Engqvist MKM. ACS Synthetic Biology. 8(6): 1411-1420 (2019).

Paper III: Bayesian genome scale modeling identifies thermal determinants of yeast metabolism

Li G, Hu Y, Wang H, Zelezniak A, Ji B, Zrimec J, and Nielsen J. Nature Communications. *Revision* (2020).

Paper IV: Performance of regression models as a function of experiment noise

Li G, Zrimec J, Ji B, Geng J, Larsbrink J, Zelezniak A, Nielsen J and Engqvist MKM. *arXiv* (2019).

Paper V: Learning deep representations of enzyme thermal adaptation

Li G, Zrimec J, Viknander S, Zelezniak A, Nielsen J and Engqvist MKM. (2020). *Manuscript*

Additional papers and manuscripts not included in this thesis:

Paper VI: Global Metabolic Network Tuning Enhances Yeast's Fitness During Carbon Source Switch from Glucose to Xylose

Li X, **Li G**, Wang Y, Liu Q, Pereira R, Chen Y and Nielsen J. (2020). *Manuscript*

Paper VII: CODY enables spatiotemporal prediction of in vivo gut microbiome reprogramming induced by diet-switch

Geng J, Ji B, **Li G**, Lopez-Isuza F and Nielsen J. (2020). *Manuscript*

Paper VIII: Genome scale metabolic modelling from yeast to human cells models of complex diseases: latest advances and challenges

Chen Y, **Li G** and Nielsen J. Yeast Systems Biology: Methods and Protocols (Springer, 2019). pp. 329-345

Paper IX: A Consensus *S. cerevisiae* Metabolic Model Yeast8 and Its Ecosystem for Comprehensively Probing Cellular Metabolism

Lu H, Li F, Sánchez BJ, Zhu Z, **Li G,** Domenzain I, Marcišauskas S, Anton PM, Lappa D, Lieven C, Beber ME, Sonnenschein N, Kerkhoven E and Nielsen J. Nature Communications. 10(1) (2019).

Paper X: Engineering of *Saccharomyces cerevisiae* for de novo biosynthesis of isoflavonoids

Liu Q, Liu Y, **Li G**, Savolainen O, Chen Y and Nielsen J (2020), *Manuscript*

Paper XI: Expanded metabolic networks together with accelerated protein evolution render cellular new traits in yeast subphylum

Lu H, Li F, Yuan L, Domenzain I, **Li G**, Chen Y, Ji B, Kerkhoven E and Nielsen J (2020) *Manuscript*

Contribution Summary

Paper I, II, III, IV and V: I co-designed the project, performed the analysis and drafted the paper.

Paper VI: I analyzed part of the transcriptomics data.

Paper VII: I analyzed the time-series data.

Paper VIII: I reviewed and wrote the part about the development history of yeast and human GEM.

Paper IX: I assisted the protein structure data collection.

Paper X: I assisted the ancestral sequence reconstruction.

Paper XI: I assisted the gene family analysis.

Preface

This dissertation serves as partial fulfillment of the requirements to obtain the degree of Doctor of Philosophy at the Department of Biology and Biological Engineering at Chalmers University of Technology. The PhD studies were carried out between August 2016 and August 2020 at the division of Systems and Synthetic Biology (SysBio) under the supervision of Jens Nielsen. The project was co-supervised by Martin Engqvist, Ibrahim EIsemman and Verena Siewers and examined by Stefan Hohmann. It was funded by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant program No 722287.

Gang Li

August 2020

Abbreviations

GEM	Genome-Scale Metabolic Model
PN	Petri Net
FBA	Flux Balance Analysis
ODE	Ordinary differential equation
ML	Machine Learning
DL	Deep Learning
RL	Reinforcement Learning
PCA	Principal component analysis
TBM	Theory-Based Model
GPR	Genotype-Phenotype Relationship
GWAS	Genome-Wide Association Study
SNPs	Single Nucleotide Polymorphisms
P/A	Presence/Absence
CNV	Copy Number Variation
MCA	Multiple Correspondence Analysis
OGT	Optimal Growth Temperature
SVR	Support Vector Regression
etcGEM	Enzyme and Temperature Constrained Genome Scale Metabolic Model
ABC	Approximate Bayesian Computation
MSE	Mean Squared Error
R^2	Coefficient of determination
DNN	Deep Neural Network
ResNet	Residual Neural Network
VNN	Visible Neural Network

Table of Contents

Background

"***What is life?***" (Schrodinger, 1944) has been a fundamental question in biology. The fundamental unit of life is the cell, in which all the information required for duplication is encoded in its genome (Figure 1). Coding sequences of the genome can be transcribed to RNAs (transcriptome) and then messenger RNAs can be translated to proteins (proteome). Proteins are used either as building blocks of the cell or to catalyze the metabolic reactions converting the input nutrients into functional chemical compounds. In a living cell, all those processes have to be precisely regulated and organized to enable its functionality. Understanding how the behaviour of a living cell emerged from the interactions between all those molecular components can have a great impact on basic biology, medicine and bio-industry.

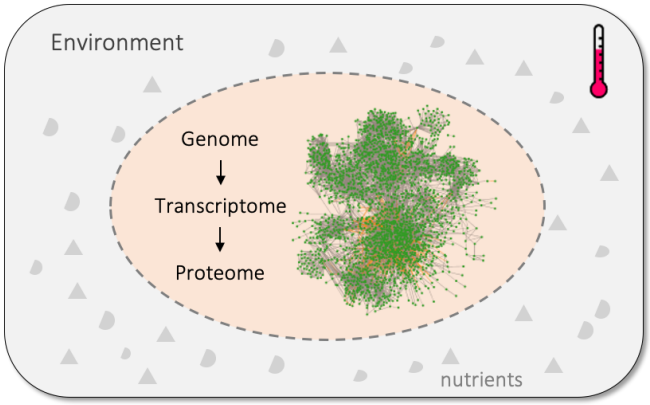

Figure 1. An illustration of a cell system.

Systems Biology

Systems biology has evolved as a new discipline that attempts to understand how complex systems underlie life (Breitling, 2010; Ideker et al., 2001; Kitano, 2002; Nielsen, 2017; Vidal, 2009). It combines computational and experimental biology to understand a complex biological system in a quantitative way. There are two different approaches in systems biology (Nielsen, 2017; Shahzad and Loor, 2012): the top-down approach, in which integration analysis of multi-omics and other high-throughput data are used to gain biological insights of the system; and the bottom-up approach, in which detailed theory-based mathematical models are constructed to simulate and predict the behaviours of the system. In a typical top-down approach, omics data (e.g. RNAseq) are generated under different conditions and then statistical tests are performed to identify the significantly changed components (e.g. genes). Gaining biological insights from those components is usually done by analyzing them in the context of functional groups like gene ontologies (GO) (Ashburner et al., 2000; Rhee et al., 2008) and metabolic networks (Patil and Nielsen, 2005), among others. The omics data are usually collected without a specific hypothesis in mind. New hypotheses are made after analyzing all the data and new experiments can then be conducted for the confirmation (Huang et al., 2017).

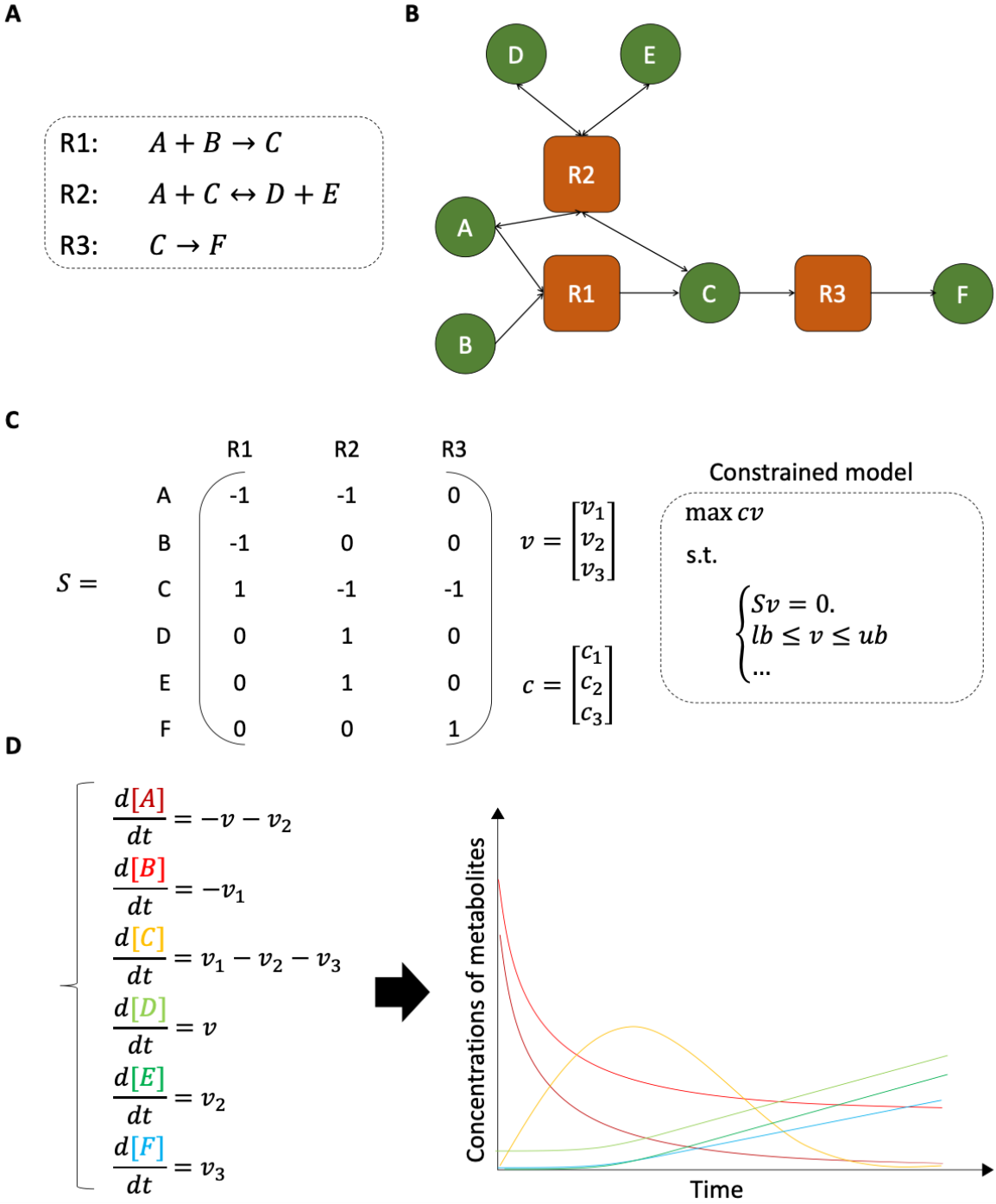

Figure 2. Different approaches to model metabolism. (A) A toy metabolism with only three reactions. A-F represent different metabolites. R1-R3 represent three different reactions. (B) Petri net modelling of biochemical reactions. (C) Constraint-based stoichiometric metabolic model. (E) Ordinary differential equations (ODE) based kinetic model. The line-plot on the right is an example of the metabolite concentrations at different time points after solving those ODEs.

The ultimate goal of the bottom-up approach is to model a biological system (e.g. cell or a subsystem) with theory-based mathematical models. The models can then be used to **understand** and **predict** how a biological system responds to external or internal perturbations. If we choose the modelling of metabolism as an example (Figure 2A), the simplest approach is to reconstruct a

metabolic network, represented as a Petri net (PN) (Lanzeni et al., 2008; Reddy et al., 1996) (Figure 2B). There are two types of nodes in PN: metabolites and reactions. The directed edges connect reactants to reactions and reactions to products. This PN modelling approach has been successfully applied on metabolic networks to (1) predict the viability of mutant strains of *Escherichia coli* and *Saccharomyces cerevisiae* (Wunderlich and Mirny, 2006); (2) reduce the complexity of the metabolic networks for feasible verifications and interpretations (Koch et al., 2017); (3) enumerate metabolic pathways for the production of heterologous target chemicals in the chassis organisms (Carbonell et al., 2012). The advantage of such a topology-based modelling approach is that it doesn't require any experimental data for the simulation. However, its applications are also largely limited to the qualitative descriptions/predictions of metabolisms.

To enable a quantitative description and prediction of metabolic networks, constraints-based genome-scale metabolic modelling (GEM) approaches based on flux balance analysis (FBA) can be applied (Orth et al., 2010) (Figure 2C). FBA is a mathematical approach for analyzing the flow of metabolites through a metabolic network. In this case, a metabolic model is represented with a stoichiometric matrix (*S*) that contains all the stoichiometric coefficients of each reaction in the model. By assuming a steady state of the metabolism, the first constraint becomes that the total production rate of any metabolite equals the total consumption rate (Figure 2C):

$$S \cdot v = 0 \qquad [1]$$

in which v is the flux vector. The second type of common constraints is to define lower and upper bounds of metabolic fluxes. This constraint can be used to: (1) define the reversibility of the reactions in the model; (2) define the input nutrients for the model; (3) force the model to produce certain metabolites; (4) define the gene knockout profiles. With these constraints, an objective function needs to be defined to enable FBA. Among others, biomass production has been the most commonly used objective function, which describes all the biomass precursors like DNA, RNA, etc in the correct proportions (Feist and Palsson, 2010). Now FBA becomes a linear programming problem. By solving it, one can get the flux values through all reactions in the model.

Current efforts made to improve the performance of a GEM fall into two groups: (1) improving the quality of the model by correcting the miss-annotated reactions in the model (Chen et al., 2019; Heavner and Price, 2015; Osterlund et al., 2012) and expanding the scope of the model by including more biological processes, like gene expression process (O'Brien et al., 2013), transcription regulation (Herrgård et al., 2006); (2) introducing additional biological constraints in the GEM, including thermodynamics (Henry et al., 2007), enzyme usage (Sánchez et al., 2017) and membrane constraints (Liu et al., 2014). Expanding the scope of the model and inclusion of any additional constraints comes along with additional parameters required. For example, enzyme-constrained metabolic models (Sánchez et al., 2017) require the enzyme turnover number (k_{cat}) for each enzyme-catalyzed reaction in the model. Accurate estimations or measurements of those parameters are critical for the performance of GEMs.

However, since a constraints-based GEM relies on the steady state assumption of metabolism, it is only suitable for determining fluxes at steady state. It ignores the dynamic nature of the metabolism and cannot predict the metabolite concentrations. Genome-scale kinetic models move one-step further to resolve the shortages of constraints-based GEMs (Jamshidi and Palsson, 2008;

Smallbone et al., 2010). The core of the kinetic models is to use ordinary differential equations (ODE) to describe the changing of metabolite concentrations over time (Figure 2D):

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{S} \cdot \boldsymbol{v}(E; x; k), x(0) = x_0 \quad [2]$$

in which *S* and *v* denote the stoichiometric matrix and the flux vector, respectively. *S* is the same as in the constraints-based GEM, while *v* is a function of many factors including enzyme concentration *E* (for enzymatic reactions), substrate concentrations *x*, and kinetic parameters *k*. There has been much progress made for the mathematical formalisms of enzyme kinetics, as reviewed in (Saa and Nielsen, 2017). Regardless of the difficulty from choosing the right kinetic formula for each enzyme, another challenge comes from the large number of kinetic parameters that need to be determined for genome scale kinetic modeling. These challenges have prevented its development and application. Current approaches have been focused on small scale models like ones for core metabolism (Khodayari et al., 2014) and/or on the development of strategies to improve the estimation of kinetic parameters (Khodayari and Maranas, 2016; Miskovic et al., 2019).

Metabolism is one of the most well-modelled biological systems, which however, only accounts for a part of the whole cell. Different models have been developed for other biological processes, like ODE based models for signaling pathways (Bridge et al., 2018; Shaw et al., 2019), Boolean computational models for gene-regulatory networks (Peter et al., 2012) and stochastic models for gene expression (Raj and van Oudenaarden, 2009). However, the ultimate goal of systems biology is to model the whole cell (Carrera and Covert, 2015; Tomita, 2001). A promising approach to this has been made by combining different models for different cellular processes together for the bacterium *Mycoplasma genitalium* (Karr et al., 2012). For instance, constraints-based GEM is used for metabolism, *Poisson* processes are used to model RNA and protein degradation, etc. This whole-cell model thereby enables the linkage between different biological processes. It was found to be able to describe the life cycle of a single cell from the level of individual molecules and their interactions and lead to various biological discoveries which were later confirmed by experiments. However, since models for different cellular processes were still based on the state-of-art models at that time, the approach did not go much beyond the understanding of the cell at that time.

One would ask whether we really need a whole-cell model? My answer is yes. Truly for many applications, the whole-cell model is not necessary. A simple example is that a PN model or a GEM is enough for the prediction of essentiality of metabolic genes. But to understand the emergent properties in the whole cell system, it is necessary to build whole cell models, since the whole system is more than the sum of its parts (cellular processes) and interactions between the molecular components in different processes need to be captured (Kitano, 2002; Nielsen and Jewett, 2008).

As reviewed above, there have been many attempts and much progress made to build mathematical models for either different cellular processes or even whole cells. The ultimate goal is a whole-cell model that is fully dependent on the kinetics of all processes and also captures the stochastic properties of many biological processes. However, this is only possible when sufficient details of the system are known, including formulating the dynamics of each single component as well as parameters required for the formulation. We are still on the way to this goal. Development of such

theory-based models is a fundamental task in systems biology; however, it is challenging and time-consuming. For instance, although the first GEM for *S. cerevisiae* was developed in 2003 (Förster et al., 2003), the research community is still working on expanding and refining this model even today (Lu et al., 2019).

Machine learning

Machine learning (ML) is at the core of artificial intelligence (AI) which aims to allow machines to perform tasks that require intelligence, like reasoning, learning, planning, problem-solving, perception, etc. (Luxton, 2016) (Figure 3A). It builds mathematical models based on data (training data) in order to make predictions or decisions without being explicitly programmed. The term "machine learning" first appeared in publications in the middle of the 20th century (Samuel, 1959). Growing attention has been given in the 21st century mainly due to two reasons: (1) advances in computational capacity, especially the development of graphic processing units (GPU) that speeded up deep learning (DL) (Oh and Jung, 2004; Raina et al., 2009), which represents the most advanced ML approach (Figure 3A); (2) the availability of big data that fueled the development of ML models as training data.

There are currently three types of learning problems in ML: supervised learning, unsupervised learning and reinforcement learning (Figure 3B). In supervised ML, the main objective is to use ML algorithms to approximate the true function $f(\cdot)$ that maps the input **features** x to the output **labels** y:

$$y = f(x) \quad [3]$$

in which x denotes the sample features and y denotes sample labels. Depending on whether the labels are categorical or continuous, there are two types of learning tasks, classification and regression (Figure 3B). The training approach tries to reduce the discrepancy between predicted labels and true labels.

Supervised ML has been widely used in biology. For instance, a support vector machine based regression model was developed for the prediction of stability changes upon mutation from the protein sequence or structure (Capriotti et al., 2005a); a deep neural network based classifier was developed for the prediction of enzyme commission numbers (Ryu et al., 2019); and many others (Chiu et al., 2019; Heckmann et al., 2018; Märtens et al., 2016; Zhou et al., 2018).

By contrast, unsupervised learning uses unlabeled training samples and can discover the similarities or differences between samples based on the **features** describing those samples (Tarca et al., 2007). In general, there are three types of tasks that can be addressed by unsupervised learning: association rule learning, clustering and dimension reduction. First, association rule learning is a rule-based approach for discovering the relationships between variables in the training dataset (PIATETSKY-SHAPIRO and G, 1991). The most intuitive example of this is market-basket analysis (Cios et al., 2007): if we look into the trade records of a Chinese supermarket, we may find that customers who bought sliced lambs would very likely also buy hotpot ingredients. Retailers can use such information to strategically price or place their products or make recommendations on an on-line shopping platform. Associate rules have also been applied in

biology to analyze gene-expression data (Becquet et al., 2002; Carmona-Saez et al., 2006). Secondly, clustering represents the most commonly used unsupervised learning task. It is a process that groups similar samples together based on their features. For instance, patients with lung adenocarcinoma can be clustered into three groups based on their gene-expression profiles (Beer et al., 2002). Patients belonging to different groups can then receive different treatments. Thirdly, dimensionality reduction is another unsupervised learning approach for condensing or simplifying the features of samples in the training dataset (Meng et al., 2016; Zampieri et al., 2019). It has many advantages: (1) it allows the visualization of data when its dimensionality is reduced to 2D or 3D, for example the first two components in Principal component analysis (PCA) (Meng et al., 2016); (2) in case of $p \gg N$ problem, which means the number of features (p) is far greater than the number of samples (N), with dimensionality reduction, p can be reduced close to or even smaller than N, which can then be used for the development of other ML models (e.g. classifiers) without the concern of $p \gg N$. An example that applies dimensionality reduction for the development of supervised models is the prediction of drug response from multi-omics data (Chiu et al., 2019).

Reinforcement learning (RL) refers to a class of techniques designed to train computational agents to successfully interact with their environment, typically to achieve specific goals (Esteva et al., 2019; Kaelbling et al., 1996). It has been successfully and commonly used in the applications like robotic manipulation (Gu et al., 2017), autonomous driving cars (Shalev-Shwartz et al., 2016), games (Silver et al., 2017), etc. It is not yet as commonly used as supervised and unsupervised learning approaches in biology, but there have been some promising attempts made (Bocicor et al., 2011; Mahmud et al., 2018; Ralha et al., 2010; Wang et al., 2018).

The performance of obtained ML models from supervised and unsupervised learning relies on many factors involved in the different steps (Figure 3C). (1) The quality of the sample labels (for supervised learning) and the engineered features are critical, since the maxim "garbage in, garbage out" remains true for all ML algorithms (Beam and Kohane, 2018). This is particularly critical when applying ML to biological datasets, where labels in biological datasets usually come from experimental measurements and are thus noisy, and engineering relevant features is based on domain knowledge (Chen et al., 2020), which is usually lacking. (2) The choice and optimization of different ML algorithms. There are many different algorithms developed for both supervised and unsupervised learning. Take regression as an example, the popular algorithms include Gaussian process (Seeger, 2004), linear, LASSO (Santosa and Symes, 1986), Elastic Net (Zou and Hastie, 2005), Bayesian linear (Goldstein and Wooff, 2007), support vector machine (Cortes and Vapnik, 1995), decision tree (Breiman et al., 1984), random forest (Breiman, 2001; Tin Kam Ho, 1995), gradient boosting (Breiman, 1997), artificial neural networks (Schmidhuber, 2015), etc. In addition to the choice of different algorithms, it is also very important to optimize the hyperparameters of the model (Claesen and De Moor, 2015).

Generalization error, which is a measure of how accurate the model is when testing it on the unseen samples (Mohri et al., 2018), is used to evaluate the performance of ML models. In the development of ML models, especially supervised models, the original dataset is splitted into training, validation and test datasets. The training dataset is used to train models with different hyperparameters, and the validation dataset is used to select the hyperparameter set with which the

model achieves the best validation score. This best model is finally evaluated on the test dataset and the test score is reported as a measure of the generalization error. In cases when the training of the model is not time-consuming, nested cross-validation (Cawley and Talbot, 2010; Stone, 1974) be used instead of the single split of train-validation-test (Figure 3D).

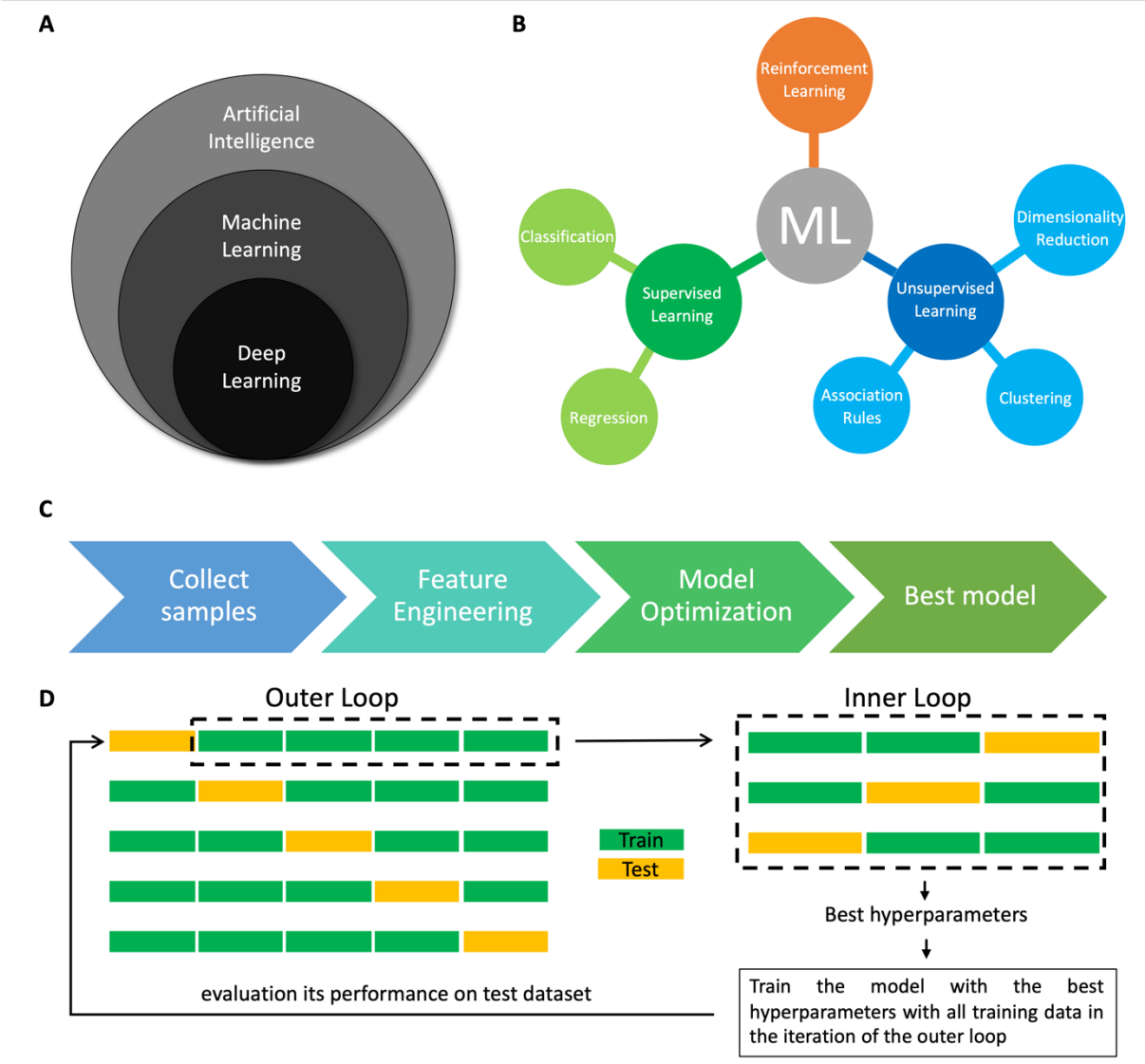

Figure 3. Machine learning. (A) Relationships between Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL). (B) Different categories of ML problems. (C) General steps for the development of a ML model. (D) An illustration of nested cross-validation approach. 5-fold for the outer loop and 3-fold for the inner loop are shown.

Once a ML model is developed, in addition to its primary application in making predictions for unseen samples, interpreting what the model has learned becomes very important before applying it at an industrial scale (Doshi-Velez and Kim, 2017). Different ML algorithms have different levels of interpretability. For instance, it is very easy to interpret a linear model by just analyzing the coefficients while it is very challenging to interpret a deep learning model. There has been some promising progress made for interpreting complex ML models (Molnar, 2020). For example, identification of important features (Altmann et al., 2010; Breiman, 2001; Hooker et al., 2019;

Zheng et al., 2017) would give hints about what features globally drive the prediction of the model; evaluation of how a model makes a specific prediction with techniques like Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017); depiction of the functional relationship between certain input features and predictions with Partial dependence (PD) and individual conditional expectation (ICE) plots (Molnar, 2020). It should be noted that all the relationships revealed by ML approaches point to correlation instead of causality. Conversion from correlation to causal relationships needs further investigation and validation.

When machine learning meets systems biology

In systems biology, building mathematical models enables us to **understand** and **predict** how a system responds to external or internal signals and perturbations. Development of theory-based models requires our deep understanding of the system but relies less on the experimental data about the systems behavior. On the contrary, data-driven ML models require lots of observed data rather than detailed understanding of the system (Figure 4A). Thereby, the application of ML in systems biology can be divided into the following scenarios (Figure 4B).

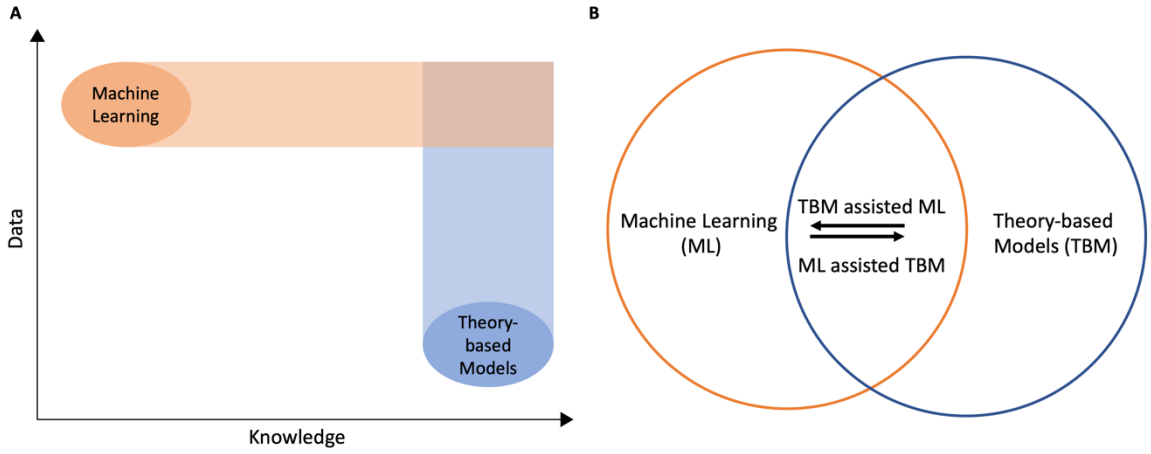

Figure 4. ML and theory-based models (TBM). (A) ML and TBM in relation to the usage of biological knowledge and experimental data. (B) Combination of ML and TBM modelling approaches.

Firstly, for a given biological system, if there is sufficient data but lack of biological knowledge for the development of theory-based models, training and then interpreting ML models can help predict the system behavior and identify the most important factors (features) that contribute the most to the model's predictive power. With the breakthrough in high-throughput technologies and accumulated biological data in the past decades, lots of data are currently available or easy-to-generate. However, the development of underlying theory is still challenging for many biological systems. An example is the relationship between genotype and phenotype of *S. cerevisiae*. Despite there being thousands of strains with characterized genomes and phenomes (Gallone et al., 2016; Peter et al., 2018), there is still a lack of theory-based models that can describe how the variations

in genomes determine the strain phenotypes. Instead, ML alone can be used to associate genotype-phenotype relationships by training supervised models using genomes as input and phenome as output. Interpretation of the resulting models would give hints about the key mutations or genes. In addition, the model could also be used to suggest specific genetic manipulations to achieve desired phenotypes. Examples for this kind of application were shown in **Paper I & II**.

Secondly, in case that both data and knowledge are currently lacking, there have been many discussions around how to combine the advantages of data-driven ML and theory-based models in different scientific fields (Alber et al., 2019; Berthold et al., 2020; Karpatne et al., 2017; Montáns et al., 2019; Peng et al., 2020; Zampieri et al., 2019). In summary, they can be combined in two different ways: (1) to develop theory-based models and use ML to assist this development (ML assisted TBM); (2) to develop ML models and use the features derived from theory-based models (TBM assisted ML). In the former way, the objective is to develop a theory-based model. The major advantage of theory-based models is that it reveals the causality relationships and it is fully interpretable. However, simplified assumptions have to be made to enable the construction of theory-based models when the knowledge about the system is lacking. This would not only lead to the poor performance of the model but also make the model difficult to interpret and analyze (Karpatne et al., 2017). ML can thereby be used to resolve those challenges in many ways including parameter estimation, uncertainty quantification and reduction, analysis of simulation results from theory-based models (e.g. by unsupervised learning like PCA) (**Paper III**). On the contrary, the objective of the TBM assisted ML approach is to develop a predictive ML model. Although theory-based models should always be preferable due to the above-mentioned advantages, in some cases there can be a gap and a part of the system cannot be modelled with theory-based models. In this case, the missing part can be filled with ML models, which take the output or features derived from the theory-based models as input. An example for this is the prediction of drug side effects through the combination of ML and genome scale models (GEM) (Shaked et al., 2016). In this application, the flux bounds for different drugs was first obtained by simulating GEMs with flux variance analysis. Then those bounds were used as input features to train a support vector machine to predict the drug side effects. This approach not only provides a way to estimate the drug side effects, but also enables identification of key reactions and pathways that are important for identifying specific side effects. Another promising approach is to develop theory-informed deep learning models by designing the connections between nodes in different layers in the deep neural networks based on biological networks (Crawford and Greene, 2020; Gazestani and Lewis, 2019; Lin et al., 2017; Yu et al., 2018), so called visible neural networks (VNN) (Ma et al., 2018).

Lastly, even if there is sufficient knowledge that enables the construction of a very detailed theory-based model, it can still benefit from ML approach. Simulating a large-scale detailed theory-based model can be time- and resource-consuming, especially in some cases it has to be simulated for thousands or millions of times. In such a case, a ML model can be used as a surrogate model to speed-up the simulation process. Unsupervised learning techniques can also be used to interpret/analyze the simulation results of theory-based models.

Aims and scope

In this thesis, I firstly demonstrated that ML alone can be used to model the genotype-phenotype relationships with two examples: ML applied to (1) resolving genotype–phenotype relationships in *S. cerevisiae* (**Paper I**). (2) predicting optimal growth temperature (OGT) of microbes from their proteomes (**Paper II**). In the second section, I used the example of modelling the thermosensitivity of yeast metabolism to demonstrate how ML techniques can be used to assist the development of an enzyme and temperature constrained genome-scale metabolic model (**Paper III**). In many applications, regression is among the most frequently used approaches. As discussed in the **Background** section**,** the performance of regression models depends on many factors, including but not limited to feature engineering and quality of response values. In the third section, to estimate how the presence of noise in response values affects the development of regression models, I mathematically derived a theoretical upper bound for the coefficient of determination (R^2) for regression models and applied it to aid the development of several regression models (**Paper IV**). In the end, I used an example of applying deep transfer learning to predicting protein melting temperatures and enzyme catalytic temperature optima, to demonstrate the application of deep learning on small biological datasets with limiting biological knowledge-based features (**Paper V**).

Genotype-phenotype relationships (GPR): abundant data with limited theory

Understanding the relationship between genotype, which refers to the genomic background of an organism, and phenotype, which is the characteristics or traits of an organism, is fundamental in biology (Kemble et al., 2019). The abilities for determination of the genotype of an organism becomes easier and cheaper than ever before with the breakthrough in genome-sequencing technologies (Schuster, 2008). Resolving GPR enables both understanding how genotypes determine phenotypes and predicting phenotypes when new genomic perturbations are introduced. Theory-based models are lacking due to current limited understanding about GPR. Current approaches use top-down systems biology approaches by first generating lots of genomic and phenotypic data (Figure 5) and then applying genome-wide association study (GWAS) to identify the key genetic variants associated with the trait of interest by applying massive numbers of statistical tests (Pearson and Manolio, 2008; Zeng et al., 2015). GWAS is particularly useful when applying it to study the genetic variations between two groups of samples. ML approaches could be a promising complement to GWAS, especially when the number of samples is large. The genomic features can be extracted for the development of supervised ML models (classifiers or regressors). The resulting model can be used in two distinct ways: (1) predicting the phenotypes when new genomic perturbations are introduced, which is exactly one of the objectives in GPR study; (2) prioritizing genomic features according to the feature importance scores derived from the model. In this chapter, two examples for each of the above two applications will be discussed (**Paper I & II**).

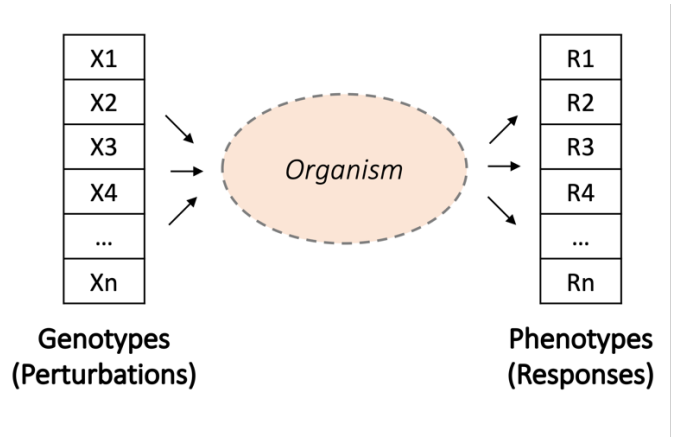

Figure 5. A classical regime for the generation of genomic-phenotypic data.

Case 1: Resolving yeast GPR with pan-genome reconstruction and ML

Dataset *S. cerevisiae* is among the species with the most abundant genomic and phenomic data since it serves as a powerful model system to study eukaryotic biology (Botstein and Fink, 2011; Goffeau et al., 1996) and widely used platform strain for the production of fuels, chemicals and pharmaceuticals (Nielsen, 2015). In this section, a total of 1,392 genomes from different *S. cerevisiae* strains were collected from GenBank and published literature, thus treated as different perturbations in the genome (Figure 5). The quality of these genomes was scrutinized, and low-quality genomes were removed. 1,364 genomes were finally obtained. 767/1,364 strains were selected and categorized based on their phenotypic features and industrial applications (Wine, Beer, Clinical, Sake, Bakery and Bioethanol) (**Paper I**).

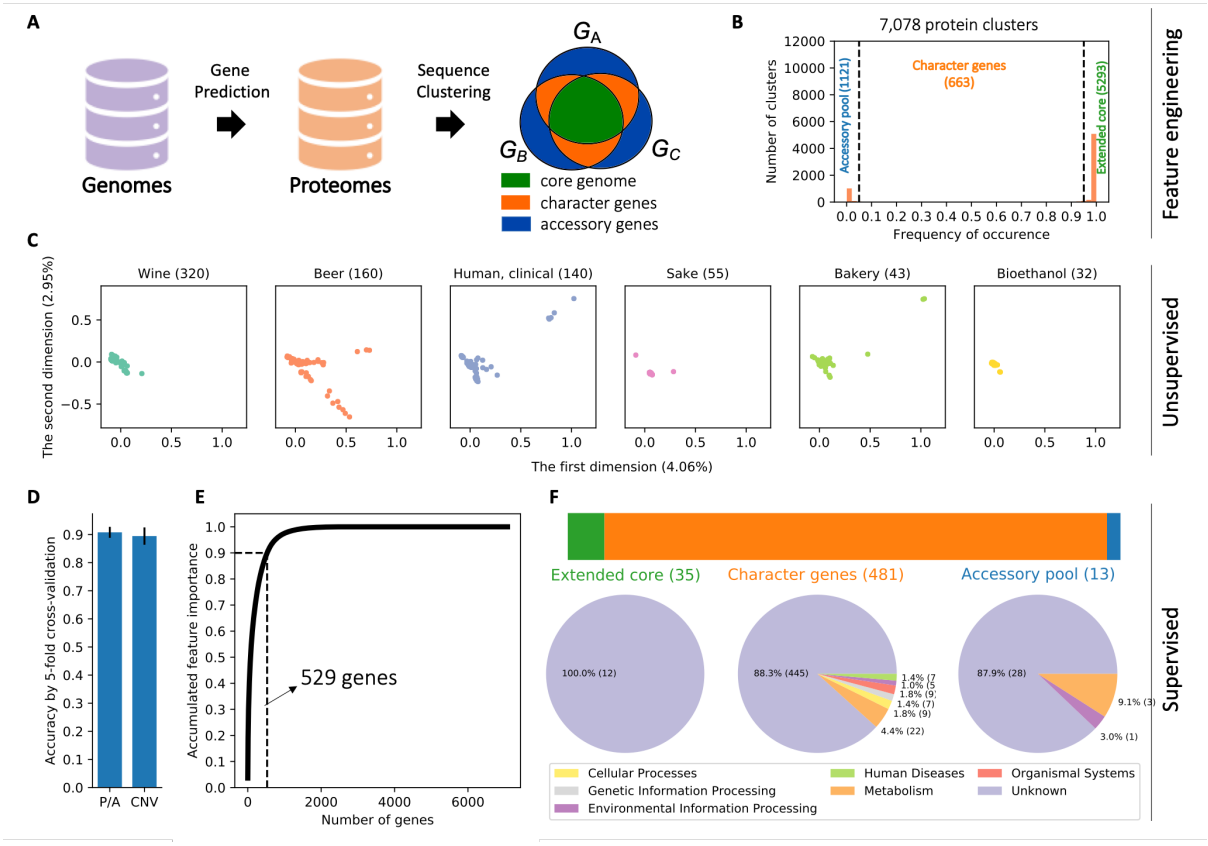

Figure 6. Resolving yeast GPR with pan-genome and ML. (A) The pan-genome reconstruction process. (B) Histogram chart that describes the distribution of the fraction of strains that each gene cluster covers. Those gene clusters were manually categorized into three groups: extended core (gene clusters that cover at least 95% of strains), accessory pool (gene clusters that cover only 5% or less of strains) and character genes (genes that cover 5–95% of strains). The number of gene clusters belonging to each category was shown in the parentheses. (C) Multiple correspondence analysis on the presence/absence of genes in the pan-genome. (D) Accuracy score obtained by 5-fold cross-validation with a random forest classifier using gene P/A or CNV features. (E) Accumulated feature importance curve from (D) on P/A dataset. The top genes contribute to 90% of the prediction power. (F) The pan-genome categories and KEGG function categories of the top genes.

Feature engineering via pan-genome reconstruction

Once samples were collected, the next step was to manually extract representative features to represent those samples (Figure 3C). There are many features that can be extracted, like single nucleotide polymorphisms (SNPs), chromosomal rearrangements, gene presence/absence (P/A), gene copy number variation (CNV), etc. P/A and CNV are the simplest but informative features associated with many phenotypes among others (Peter et al., 2018). The extraction of P/A and CNV can be done by the reconstruction of the pan-genome, which accounts for a set of all genes across all strains within this species (Tettelin and Masignani, 2005). The reconstruction was done at the protein level by first predicting protein-coding genes from genome sequences and then clustering the resulting translated protein sequences based on the similarities (Figure 6A). This approach clustered the 8.9 million protein sequences from 1,364 genomes of different strains into 7,078 protein clusters (Figure 6B). We refer to these 7,078 protein clusters as 7,078 genes in the pan-genome hereafter. This pan-genome can then be encoded with either gene P/A or CNV tables, which can be used as input features for later ML analysis.

Unsupervised learning reveals strain similarities

Gene P/A of 767 labeled strains were analyzed with Multiple Correspondence Analysis (MCA) (Abdi and Valentin, 2007), which is a supervised ML method like PCA, but designed for categorical variables. The results reveal some interesting patterns (Figure 6C): (1) most of the strains are clustered close to each other, indicating that most strains are very similar to each other in terms of the gene content; (2) the MCA was also able to distinguish strains from subclasses 'Beer1' and 'Beer2' as previously described (Gallone et al., 2016) (**Paper I**).

Supervised learning prioritizes the important genes

We next applied supervised ML to identify the major differences across strain types. The random forest classifier algorithm was chosen as it can score the feature importance. The model was trained on either gene P/A or CNV as input features to classify strains into these 6 different strain types. CNV contains additional copy number information in addition to gene P/A. Both models on P/A or CNV achieved accuracies up to 90% (Figure 6D). The model does not perform better when using CNV than only P/A, which indicates that gain/loss of some genes is more likely to determine the strain types rather than increase/decrease of the gene copy number. To prioritize the important genes that contribute to the differences among different strain types, the feature importance scores were extracted from the model trained on P/A and the top 527 genes that contribute 90% of the predictive power were obtained (Figure 6E). Further functional analysis (Figure 6F) of those genes revealed that (1) those important genes mainly belong to the group of character genes in the pan-genome; (2) function of most genes are poorly annotated so far; (3) the largest group of genes with known functions are metabolic genes. This makes sense as most of those selected strains were used for the production of different products in industry and hence have been selected for having distinct metabolism, e.g. for utilization of certain metabolites in the medium or producing specific flavors. Further analysis of those strains thereby can be focused on the metabolism with strain-specific theory-based models (e.g. GEM) (**Paper XI**).

Case 2: Annotating thermal niche of microorganisms with ML

In the above example, ML was used mainly to interpret the relationships between genomic features and sample strain phenotypes. There are also cases in GPR study that our objective is to develop a predictive model for re-use. An example for this is predicting the optimal growth temperature (OGT), which is an important characteristic of microorganisms, directly from their genomes. The OGT has been widely used in various applications, including enzyme engineering (Demirjian et al., 2001), microbiology (Kato et al., 2019), evolution biology (Green et al., 2013; Nguyen et al., 2017). However, experimental determination of OGT remains challenging since it is a laborious process that requires cultivation in temperature-controlled conditions or most microorganisms that cannot even be cultured in the laboratory (Rappé and Giovannoni, 2003). The direct link between genome and OGT so far is not well understood yet, even though huge efforts have been paid in the past decades to uncover the factors that determine OGT at various levels of DNA, RNA, proteins and metabolic pathways (Engqvist, 2018; Hickey and Singer, 2004). Current knowledge is still far from enough to develop theory-based models that describe causal link between strain genome and OGT.

Notably, there has accumulated lots of traits data, including OGT for thousands of microorganisms in the past decades, which are collected and stored in different public databases like ATCC (http://www.lgcstandards-atcc.org), DSMZ (http://www.dsmz.de), NCTC (http://www.phe-culturecollections.org.uk), NIES (http://www.shigen.nig.ac.jp) and BacDive database (Söhngen et al., 2016), etc. OGTs from those databases have been combined into a single and well-structured dataset with 21,498 microorganisms by Engqvist MKM (Engqvist, 2018). In addition, with the development of genome sequencing technologies, genomes of many microorganisms in this dataset have been sequenced and annotated. This enables us to develop ML models for the prediction of OGT from the genome of microorganisms (**Paper II**).

Dataset A high-quality dataset is crucial to the development of ML models. In this application, a dataset containing OGTs of 21,498 microorganisms was collected (Engqvist, 2018). Since characteristics of proteomes, especially the amino acid compositions, were found to be strongly correlated with organism OGT (Hickey and Singer, 2004), proteomes of 5,761/21,498 microorganisms were collected from various public databases. After removal of low-quality protomes, a dataset with 5,532 organisms was obtained (Figure 7A).

Model development and validation Once the training dataset is obtained, the next step is to develop and validate ML models. This is a supervised learning problem that aims to develop a predictive regression model. Six different regression algorithms were tested with a nested-cross validation approach (As illustrated in Figure 3D). The coefficient of determination score (R^2) was used as the measure of model performance. Two sets of features: amino acid compositions and dipeptide compositions, were tested individually as well as in combination. The best model (R^2 of 0.88) was achieved by training a support vector machine repressor using the feature set of dipeptide compositions (Figure 7B). It shows the highest accuracy compared with models developed by other approaches on the same dataset (Nakashima et al., 2003; Zeldovich et al., 2007). The final SVR model was then trained with all the samples.

Availability In this application, it's important to make our model publicly accessible and reusable. Thereby, we developed a command-line based tool called Tome for this OGT model

(https://github.com/EngqvistLab/Tome)). It can be executed with a single command by taking all the proteins for a given organism in FASTA format as input (Figure 7C).

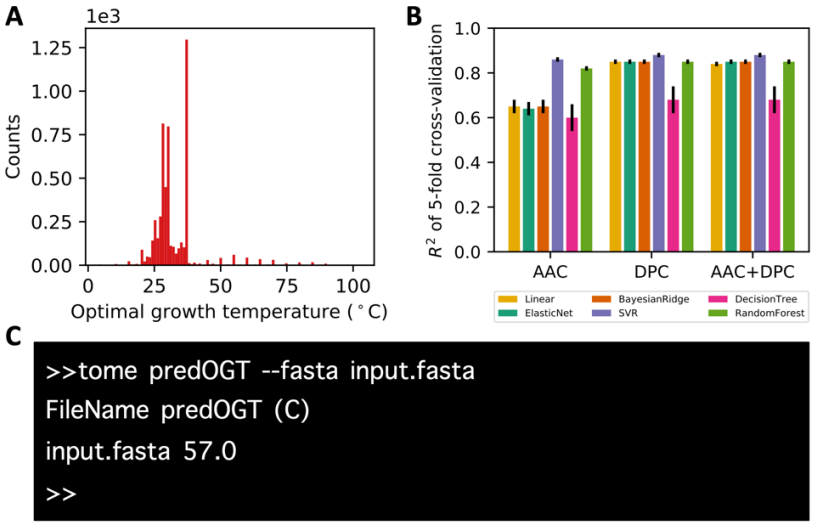

Figure 7. Development and release of OGT prediction model. (A) Distribution of OGTs in the training dataset; (B) Coefficient of determination score (R^2) obtained by a 5-fold cross-validation for six different regression models. Error bars represent the standard deviation of R^2 scores. AAC, amino acid composition; DPC, dipeptide composition; (C) A command line tool (Tome) for OGT prediction was developed.

Thermosensitivity of yeast metabolism: some theory and some data

Temperature is one of the most important environmental factors that can dramatically affect the physiology of organisms. As shown in the last chapter, each microorganism has evolved to its own optimal growth temperature (OGT), where minor deviations from the optimal temperature by merely a few degrees can dramatically impair cell growth. For instance, the model eukaryotic organism *S. cerevisiae* has an optimal growth temperature of ~30°C, whereas a temperature of 42°C is already lethal to the organism (Caspeta and Nielsen, 2015; Zakhartsev et al., 2015). Functions of all the cellular components, such as DNA, RNA, proteins and lipids are affected by temperatures to different extent (Driessen et al., 2014; Leuenberger et al., 2017; Neidleman, 1987; Slivka et al., 2012). As a matter of fact, the temperature dependence of the cell physiology is a function of the temperature dependences of all cellular components. However, mapping the temperature effects on single cellular components to ones on cell physiology has been a long-standing question in systems biology, where the main obstacles are the lack of sufficient data and theory.

Many models that describe the temperature dependence of cell growth were developed based on the very simplified assumptions. For example, in the textbook for biological engineering (Villadsen et al., 2011), the temperature dependence of growth rate is considered as a function of enzyme activity and protein denaturation at a proteome scale:

$$r(T) = \frac{Ae^{-E_g/RT}}{1+Be^{-\Delta G_d/RT}} \quad [4]$$

in which A and B are two constants; E_g is the activation energy of the growth process; ΔG_d is free energy change of protein denaturation; R is the universal gas constant; T is the temperature. In another work, Dill K. et al used the protein length distribution in the proteome of a microorganism as well as a dominant activation energy term (ΔH^*) to describe the temperature dependence of cell growth (Dill et al., 2011):

$$r(T) = r_0 e^{-\Delta H^*/RT} \prod_{i=1}^{\Gamma} f(N_i, T) \quad [5]$$

where r_0 denotes a growth-rate constraint. Γ is the number of proteins that are essential to growth. $f(N_i, T)$ is the probability that the i-th essential protein is in the folded state, which can be estimated from the protein length. This model is useful to give an estimate of the number of essential proteins for cell growth which can be used to interpret the differences between strains with different genetic backgrounds (Caspeta and Nielsen, 2015). There are many other similar models as reviewed in (Grimaud et al., 2017). All these models use very few proteome-wide parameters to describe the temperature dependence of cell growth. However, it captures very little information about the temperature dependence of cellular components like proteins and their associations, which limits their application to further our understanding of temperature effects on cell physiology from a molecule level.

To this end, Chang R et al (Chang et al., 2013) developed a genome-scale metabolic model (GEM) for *Escherichia coli* that incorporates the temperature dependence of enzyme activities. This was done by firstly determining the maximal flux through each reaction (V_{max}) and then making this V_{max} temperature dependent and using it as the new upper bound in the constraints. The temperature dependence of V_{max} is described as a product of protein denaturation for which a two-state

denaturation was assumed, and activity for which the *Arrhenius* equation was used. The resulting model was able to predict the growth-limiting reactions for thermotolerance which were then validated by experiments. Later, the model was expanded by Chen K et al (Chen et al., 2017) by including the protein-folding and chaperone network into the GEM that contains both metabolism and protein expression networks. Additional temperature dependence of folding kinetics and aggregation propensity was included for each protein in the model. The resulting model was able to predict the chaperone-mediated proteome reallocation of *E. coli* at different temperatures.

With the increasing complexity of the models, the number of parameters increases from several (*e.g.* 4 in Eq 4 and 3 in Eq 5) to a few thousand (e.g. GEM-based). Most of the parameters are unknown and have to be estimated empirically or computationally. Even with experimentally determined values, they are not very accurate due to the noise in the experimental settings and inherent difference between in *vitro* and in *vivo* conditions. This leads to large statistical uncertainties in model parameters and can make the models unreliable. The situation becomes even worse when modelling more complex organisms like *S. cerevisiae* and other ones which are not as well-studied as *E. coli*.

In this chapter, I will discuss how ML technologies can help resolve those challenges when modelling the thermosensitivity of yeast metabolism (**Paper III**).

Development of enzyme and temperature-constrained GEM

In principle, the genome-scale kinetics model is preferred for modelling the thermosensitivity of yeast metabolism. However, development of such models is not feasible at the moment due to the lack of detailed kinetics of enzymatic reactions. Thereby, the simplified version constraints-based GEM under the steady-state assumption was chosen (Figure 8). To model the temperature dependence of yeast metabolism at the molecular level, modelling how temperature affects enzyme activities is the primary step. A classical view is that it's a combination of temperature effects on enzyme denaturation and *turnover number* k_{cat} (Figure 8A).

Denaturation The two-step denaturation assumption is the simplest way to model the temperature dependence of the protein denaturation process (Chang et al., 2013; Chen et al., 2017; Kumar and Nussinov, 2001). Under this assumption, a protein molecule is in either native or denatured state and the transition between two states is reversible (Figure 8B). Thereby, the concentration of an enzyme in the cell at different temperatures can be modelled as

$$[E]_{N,i} = \frac{[E]_{t,i}}{1+e^{-\frac{\Delta G_u(T)}{RT}}} \quad [6]$$

in which $[E]_{N,i}$ is the concentration of enzyme i in the native state; $[E]_{t,i}$ is the total enzyme concentration including both native and denatured enzymes; $\Delta G_u(T)$ is the free energy difference between the denatured state and the native state, which can be obtained from

$$\Delta G_u(T) = \Delta H^* + \Delta C_{p,u}(T - T_H^*) - T\Delta S^* - T\Delta C_{p,u} log(\frac{T}{T_S^*}) \quad [7]$$

in which ΔH^* and ΔS^* are the enthalpy and entropy changes between the denatured and native states at convergence temperatures T_H^* (373.5 K) and T_S^* (385 K) (Murphy and Gill, 1991; Robertson and Murphy, 1997; Sawle and Ghosh, 2011); $\Delta C_{p,u}$ is the difference in heat-capacity change between the denatured and native states. Thereby, to obtain the temperature dependence of enzyme denaturation, three parameters ΔH^*, ΔS^* and $\Delta C_{p,u}$ are required.

Turnover number k_{cat} The classical view of temperature effects on chemical reactions is usually described with *Arrhenius* equation (Eq 6):

$$k = Ae^{-E_a/RT} \quad [8]$$

where A is a constant, R is the universal gas constant and E_a is the activation energy. However, many studies have found that it is insufficient to explain the temperature dependence of enzyme activities together with protein denaturation (Buchanan et al., 1999; Daniel and Danson, 2010; Hobbs et al., 2017; van der Kamp et al., 2018). Thereby, an expanded *Arrhenius* equation (macromolecular rate theory) was used (Figure 8C), by including a non-zero heat-capacity change ($\Delta C_p^{\ddagger}$) between the transition state and the ground state of the enzyme catalytic process (Hobbs et al., 2017; van der Kamp et al., 2018):

$$k_{cat}(T) \propto \frac{k_B T}{h} e^{-\frac{\Delta G^{\ddagger}(T)}{RT}} \quad [9]$$

in which k_B is the Boltzmann constant, h is Planck's constant and $\Delta G^{\ddagger}(T)$ is the free energy difference between the ground state and the transition state which can be expressed as

$$\Delta G^{\ddagger}(T) = \Delta H_{T_0}^{\ddagger} + \Delta C_p^{\ddagger}(T - T_0) - T\left(\Delta S_{T_0}^{\ddagger} + \Delta C_p^{\ddagger} ln\left(\frac{T}{T_0}\right)\right) \quad [10]$$

in which $\Delta H_{T_0}^{\ddagger}$, $\Delta S_{T_0}^{\ddagger}$ and $\Delta C_p^{\ddagger}$ are the differences in enthalpy, entropy and heat capacity change between the transition and ground states, respectively, and T_0 is the reference temperature.

Enzyme and temperature constrained GEM (etcGEM) An enzyme constrained model has been previously developed for yeast (Sánchez et al., 2017). Its central concepts are: 1) the flux through each reaction cannot exceed the capacity of its catalytic enzyme: $v_i \leq k_{cat,i} \cdot [E]_i$, where $[E]_i$ is the concentration of enzyme i; 2) the total enzyme amount is limited in the cell: $\sum [E]_i \leq [E]_t$. Once the temperature dependent denaturation and k_{cat} were considered, $[E]_i$ in the first constraint should be $[E]_{N,i}$ which is the concentration of individual active enzymes. $[E]_i$ in the second constraint should be $[E]_{t,i} = [E]_{N,i} + [E]_{U,i}$, which is the total concentration of enzymes in both active and denatured forms (Figure 8A). In addition, to capture the increased expenditure for maintenance under increased heat stress, a temperature dependent Non-Growth Associated ATP maintenance term can be assumed from experimental measurements. All the constraints in etcGEM are summarized in Figure 8A.

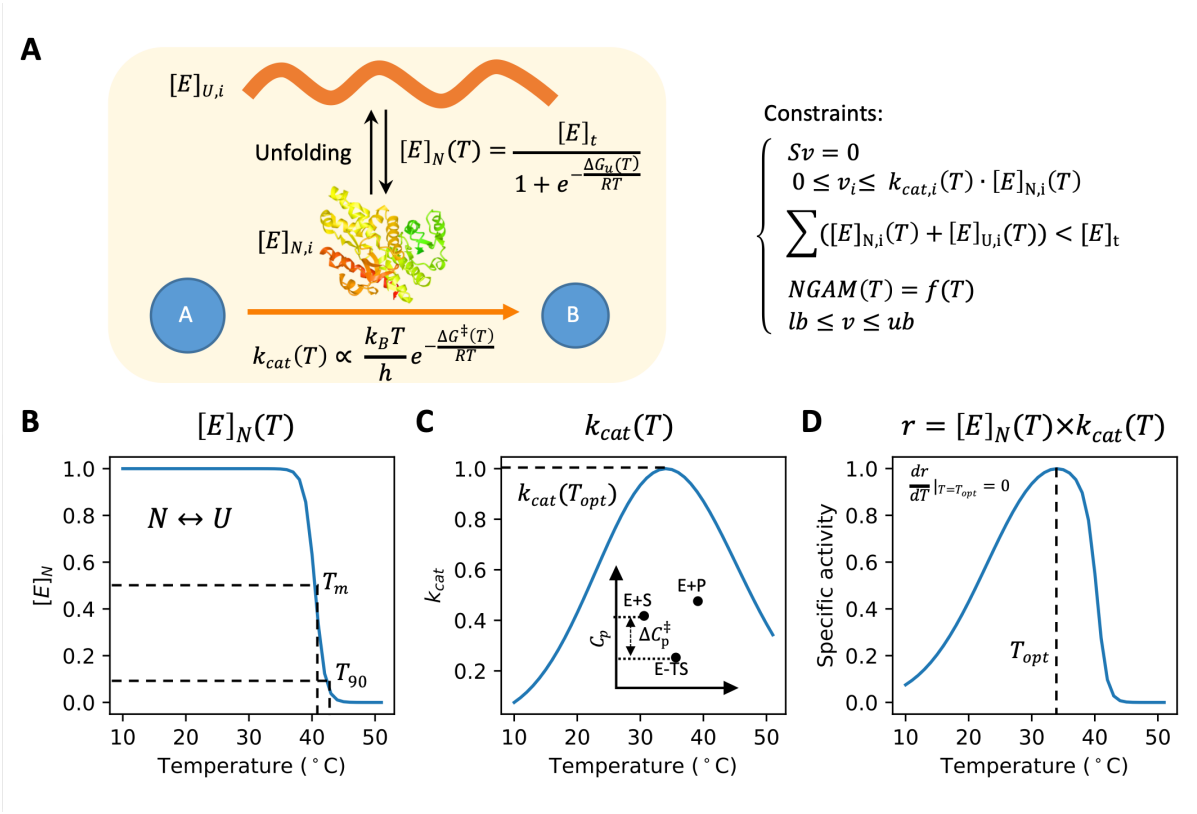

Figure 8. Enzyme and temperature constrained GEM (etcGEM). (A) An illustration of the temperature effects on enzyme-catalyzed reactions and their integration into a constrained GEM. (B) A two-state denaturation model was used to describe the temperature dependent unfolding process. $[E]_N$ is the concentration of the enzyme in native state; T_{opt} is the optimal temperature at which the specific activity is maximized; T_m and T_{90} are temperatures at which there is a 50% and 90% probability that an enzyme is in the denatured state, respectively. (C) Macromolecular rate theory describing the temperature dependence of enzyme *turnover number* k_{cat}. Inset shows the heat capacity difference between ground state (E+S) and transition state (E-TS), adapted from Hobbs J., *et al* (Hobbs et al., 2017). (D) Temperature dependence of enzyme *specific activity* r, which is a product of (B) and (C).

In the etcGEM, to quantitatively describe the temperature dependence of enzyme activities, two Gibbs free energy expressions ($\Delta G_u(T)$ in Eq 7 and $\Delta G^{\ddagger}(T)$ in Eq 10) respectively for protein unfolding and catalytic process are required. Thereby, six thermal parameters $\Delta H^{\ddagger}_{T_0}$, $\Delta S^{\ddagger}_{T_0}$, $\Delta C^{\ddagger}_p$, ΔH^*, ΔS^*, $\Delta C_{p,u}$ for each enzyme and its catalyzed reaction have to be determined. However, the direct experimental measurement of those thermal parameters is not possible. They have to be obtained indirectly. In this case, we need to use theories already developed:

(1) the ΔG_u at melting temperature (T_m) (the temperature at which there is a 50% possibility that an enzyme is in the denatured state) is 0 (Figure 8B);

(2) the ΔG_u at T_{90} (the temperature at which there is a 90% possibility that an enzyme is in the denatured state) is $-RTln9$ (Figure 8B);

(3) the k_{cat} value at T_{opt} (the temperature at which the specific activity is maximized) is known and can be incorporated into Eq 9 (Figure 8C);

(4) at T_{opt} the first order derivative of specific activity with respect to temperature is 0 (Figure 8D);

(5) ΔH^* and ΔS^* can be estimated directly from protein sequence length (Sawle and Ghosh, 2011).

After applying all those theories (check details in **Paper III**), there are three parameters that have to be determined for each enzyme in etcGEM: T_{opt}, T_m and $\Delta C_p^{\ddagger}$.

ML for parameter estimation

In the resulting yeast etcGEM, there are 764 enzymes described with 2,292 parameters (T_{opt}, T_m and $\Delta C_p^{\ddagger}$ for each enzyme). However, values of most parameters are unknown. For example, there are only around 14 enzymes with T_{opt} values (as of Feb 2018) from BRENDA (Jeske et al., 2019) which is the main resource for enzyme data, that can be successfully mapped to etcGEM through Uniprot ID. With T_m, 266/764 enzymes have an experimentally determined value (Leuenberger et al., 2017). While there is no experimental data for $\Delta C_p^{\ddagger}$of any yeast enzymes. Determination of those missing values becomes the first challenge in modelling the thermosensitivity of yeast metabolism with etcGEM. I therefore used data-driven ML approaches to estimate those missing values.

Enzyme T_{opt} (Papers II and IV) Although there are very limited records for yeast enzymes in BRENDA (release 2018), there are about 33,000 T_{opt} records for enzymes from different organisms, of which around 5,300 can be associated with a Uniprot ID which can then be used to get protein sequences. Those data can be used to train a regression model for the prediction of enzyme T_{opt} directly from primary sequences. Domain knowledge can be incorporated to engineer the relevant features for enzyme T_{opt} prediction. In biotechnology and protein engineering OGT is typically used directly to guide the discovery of thermostable enzymes (Vieille and Zeikus, 2001). This is under the fact that each enzyme should be at least functional at the OGT of its source organism. Then we can expect that OGT can be used as one of informative features for the prediction of enzyme T_{opt}. Thereby, an enzyme T_{opt} dataset with 2,609 enzymes with known sequences and OGTs of their source organisms were collected (Figure 9A).

Three sequence-based feature sets were extracted: amino acid composition, dipeptide composition and basic protein properties, including sequence length, isoelectric point, molecular weight, aromaticity (Lobry and Gautier, 1994), instability index (Guruprasad et al., 1990), gravy (Kyte and Doolittle, 1982), and fraction of three secondary structure units: helix, turn, and sheet. Five regression models were tested on those feature sets (Figure 9B). It showed that inclusion of OGT as an additional feature to the sequence-based features greatly improved the model performance (R^2 improved from 0.3 to over 0.5), which is in line with our original expectation about the relationship between enzyme T_{opt} and OGT. The combination of amino acid composition and OGT already achieved the highest R^2 score with a random forest regressor in a 5-fold cross-validation. Further inclusion of any other feature sets such as dipeptide composition did not further improve the model performance. This means that although dipeptide composition itself showed some predictive power (R^2 of 0.25) to the enzyme T_{opt}, they do not provide additional information compared with amino acid frequencies for the prediction of enzyme T_{opt}.

Two straightforward ways that are promising to further improve the prediction of enzyme T_{opt} are: (1) collection of more samples and (2) engineering more features. To collect more samples, T_{opt} of 5,675 enzymes with known protein sequences were collected from the newly released BRENDA

(2019). Of these 3,096 enzymes were successfully mapped to a microbial OGT database (Engqvist, 2018). This new release provided around 400 more training samples. Two large feature sets were extracted: (1) 5,494 domain knowledge-based features belonging to 20 different subsets were extracted with iFeature (Chen et al., 2018); (2) UniRep (Alley et al., 2019), a deep-learning based sequence embedding with 5,700 descriptors was extracted. Those different sub-feature sets showed different prediction power to enzyme T_{opt} (Figure 9C). Interestingly, the amino acid composition alone already achieved the best model performance (R^2 of ~0.4) among other feature sets including UniRep as well as the combination of 20 sub-feature sets extracted from iFeature (Figure 9C). Inclusion of OGT as an additional feature into any feature sets again boosted the performance of ML models (Figure 9C). The best model achieved an R^2 score of 0.55 in a 5-fold cross-validation.

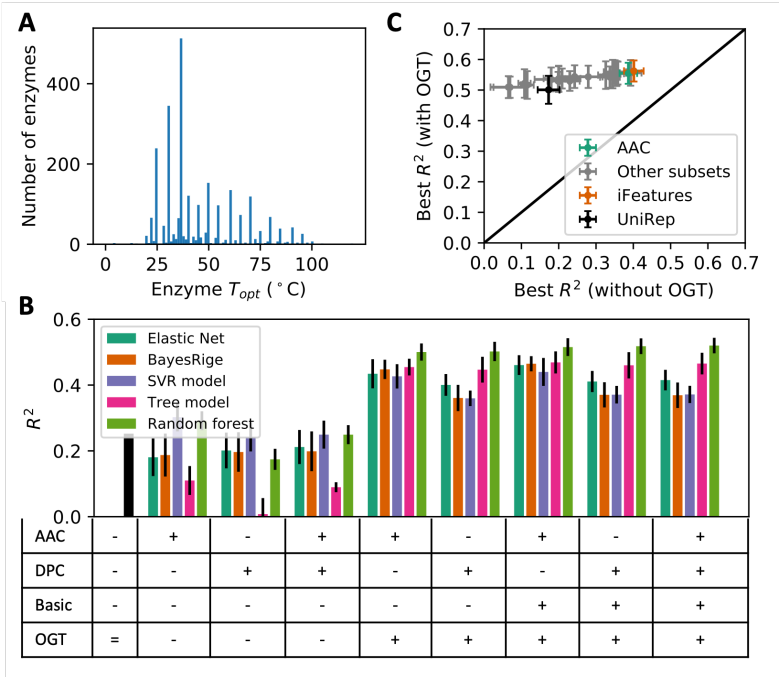

Figure 9. Development of enzyme T_{opt} prediction model. (A) Distribution of T_{opt} values in the training dataset. (B) 5-fold cross-validation results for five regression models on different feature sets. The "=" shows the explained variance when using OGT as the estimation of enzyme T_{opt}. "+" and "-" denote the presence and absence of feature sets used for ML analysis. Error bars show the standard deviation of R^2 scores obtained in 5-fold cross validation. AAC, amino acid frequencies; DPC, dipeptide composition; Basic, basic properties of proteins. (C) The performance comparison between the best regression models on different feature sets with and without OGT as an additional feature.

The enzyme T_{opt} prediction model is not only useful to estimate the T_{opt} of yeast enzymes to parameterize the etcGEM (Figure 8), but also has a wide application in enzyme engineering. The best model was applied to annotate two enzyme databases: BRENDA and CAZy (Lombard et al., 2014) which is a database with carbohydrate-active enzymes. 6.5 million and 0.9 million enzymes respectively from BRENDA and CAZy were successfully mapped to microorganisms with known OGT and their T_{opt}s were predicted by the best ML model developed in this thesis. Those annotations are particularly useful for the identification of enzymes serving as starting points for protein engineering. Furthermore, we incorporated those annotated datasets into the previously developed command line-based tool Tome (Figure 7C) to ensure easy access to those data. It can

be used for the identification of enzyme functional homologues with different estimated T_{opt}, one can either simply specify an EC number or CAZy family ID and temperature range of interest to get all enzyme sequences matching the criteria. Alternatively, the sequence of an enzyme of interest can be provided in fasta format. The algorithm will then perform a protein BLAST (Camacho et al., 2009) and an additional output file will be generated containing only homologous enzymes within the specified temperature range.

Protein melting temperature (T_m) There have been many theoretical or semi-theoretical models developed for the prediction of protein melting temperatures. Those methods are mainly based on either only protein length (Sawle and Ghosh, 2011) or protein 3D structures (Murphy and Gill, 1991; Oobatake and Ooi, 1993). Those methods suffer from either low-accuracy due to over-simplified assumptions (e.g. protein length-based approaches) or the lack of high-quality structures for most enzymes. Take *S. cerevisiae* as an example, even with homology modelling, only around 50% of the enzymes have at least one structure with a length coverage higher than 95% according to SWISS MODEL (2018-01-16) (Bienert et al., 2017). In addition, there is no strong correlation between the length of the proteins and their experimentally measured melting temperatures (Leuenberger et al., 2017). Thereby, I tested if we can develop a predictive ML model for the prediction of protein T_m. An excellent dataset for this is from Leuenberger *et al* (Leuenberger et al., 2017), where melting temperatures (T_m) of 3,557 proteins from *E. coli* (730 proteins), *S. cerevisiae* (707), *Thermus thermophilus* (1,083), and human cells (1,037) were measured via a proteomics approach (Figure 10A). Next, we extracted 2,618 features from primary protein sequences, and predicted secondary structure, relative solvent accessibility and disordered regions (Details in the legend of Figure 10). The performance of several regression models was tested via 5-fold cross-validations. As shown in Figure 10B, all models tested on organism-specific datasets performed very poorly with an R^2 score close to zero, even on the combined dataset with proteins from three mesophiles. Surprisingly, the best model trained on the proteins from all four organisms achieved an R^2 score of 0.62. Then we asked if this model was truly better than ones trained on organism-specific datasets with a near-zero R^2. Importantly, an R^2 score of 0 means that the model is equivalent to a null model which uses the mean value of target variables as the prediction and doesn't depend on any other features, whereas a negative/positive R^2 score means that the model is worse/better than the null model. Therefore, during the 5-fold cross validation of random forest on the "All'' dataset, test R^2 scores on proteins from individual organisms as well as with all mesophiles were calculated (Figure 10C). The results suggested that, even though this model performed well on proteins from all four organisms, it performed worse than an organism-specific null model.

In this case, for etcGEM, the best option is to use a yeast null model which uses the mean T_m of all existing yeast proteins (51.9 °C) as the prediction for ones with unknown T_ms.

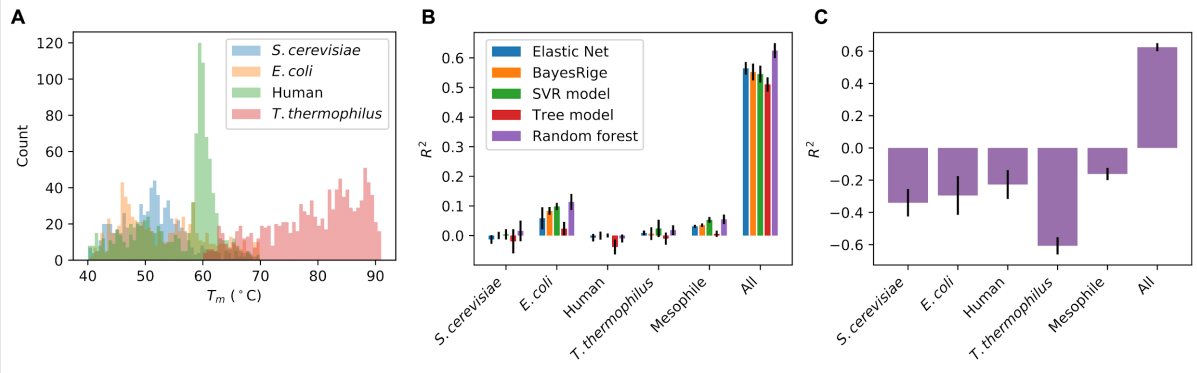

Figure 10. Development of ML models for the prediction of protein melting temperatures (T_m). (A) Distribution of melting temperatures from four different organisms. (B) The performance of five different regression models on different datasets in a 5-fold cross validation. (C) The performance of the best model trained on 'All' (random forest), compared with individual null models. Features used in this application: primary sequence and physicochemical features were extracted with propy (Cao et al., 2013), which is similar to ones extracted by iFeature which was published after those analyses. Protein sequence length, isoelectric point (pI) and molecular weight were extracted with BioPython (Cock et al., 2009). A 2-class and a 20-class relative solvent accessibility were predicted with ACCpro and ACCpro2 (Magnan and Baldi, 2014) for each protein, respectively. 162 features were calculated based on the predicted results. Protein disorder regions were predicted by DisEMBL (Linding et al., 2003). 9 features were further calculated based on the predicted results for each protein. A 3-class and an 8-class protein secondary structures were predicted by SSpro and SSpro8 (Magnan and Baldi, 2014) for each protein, respectively. Each protein sequence was thereby converted to two secondary structure sequences. For 3-class secondary structures, k-mer (k = 1, 2, 3, 4, 5) features were extracted. For 8-class secondary structures, k-mer (k = 1, 2, 3) features were extracted. ***Note: those results are not included in any papers or manuscripts since they are negative.***

Enzyme $\Delta C_p^{\ddagger}$ $\Delta C_p^{\ddagger}$ is the heat capacity difference between the ground state and the transition state in the catalytic process of enzymatic reactions (Hobbs et al., 2017; van der Kamp et al., 2018). It has been shown as the evolutionary driver for thermal adaptation of enzyme catalysis (Nguyen et al., 2017). While it remains challenging to directly measure the value of $\Delta C_p^{\ddagger}$ by experiments, it is usually obtained by fitting the Eq 9 to experimentally determined k_{cat} values at different temperatures. For enzymes in yeast etcGEM, those experimental data are missing or hidden in the literature. Thereby, an average value of -6.3 kJ/mol/K was estimated by fitting the Eq 9 to yeast specific growth rate at various temperatures (Hobbs et al., 2017). This value was then applied for all enzymes in the yeast etcGEM.

ML for uncertainty quantification and reduction

So far, I have discussed how to estimate the missing values for parameters in the etcGEM via data-driven ML approaches. The yeast etcGEM equipped with those parameters should then be validated by simulating the temperature dependence of yeast physiology to see if the results are consistent with experimental data. Thereby, three datasets were collected for the model validation: (i) the maximal specific growth rate in aerobic batch cultivations (Caspeta and Nielsen, 2015), (ii) anaerobic batch cultivations (Zakhartsev et al., 2015), and (iii) fluxes of carbon dioxide (CO_2), ethanol and glucose in chemostat cultivations (Postmus et al., 2008), at various temperatures. It showed that etcGEM can only accurately reproduce the data when the temperature is lower than about 30 °C, while it failed at the high temperature range (>30°C) (Figure 11). As illustrated in Figures 8B-C, this may due to that at lower temperatures the temperature dependence of enzyme

k_{cat} values is the major determining factor, while at higher temperatures there are additional complicate factors such as the protein denaturation and increased energy for maintenance (Zakhartsev et al., 2015). Particularly, the metabolic shift as shown in Figure 11C happens within about 2 degrees (36-38 °C) and accurate prediction of this metabolic flux shift may require very precise enzyme parameter values.

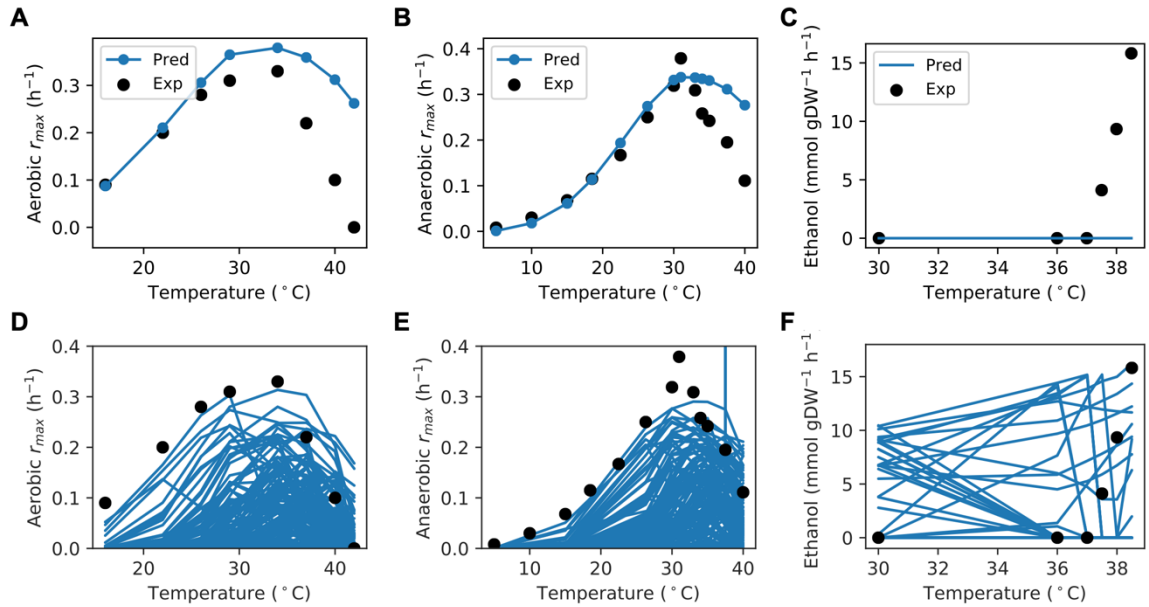

Figure 11 Simulated results of specific growth rate at (a) aerobic, (b) anaerobic batch cultivations and (c) ethanol flux at chemostat cultivation with etcGEM at different temperatures. Pred, predicted results by etcGEM. Exp, experimental data. (DEF) Simulated results when randomly sampling parameters (128 times) from pre-defined distributions that describe the uncertainties in the parameter values. r_{max}, maximal specific growth rate, corresponding to the y-axis label in A and B.

Then how accurate are our parameter values? Enzyme T_{opt} values were predicted by the ML model, which showed an R^2 score of around 0.5 in a 5-fold cross validation. It corresponds to a root mean squared error of 13.0 °C in the predicted values. Experimentally determined enzyme T_ms have an experimental error of around 3.4 °C while the values estimated by the yeast null model have a standard variance of 5.9 °C. Using the same $\Delta C_p^{\ddagger}$ value for all enzymes is a very simplified assumption. How do those uncertainties in the model parameters affect model predictions and to what extent? I thereby tested the performance of the etcGEM by first assuming normal distributions for those parameters to describe those uncertainties and then randomly sample many sets of parameter values. For$\Delta C_p^{\ddagger}$ of which uncertainty is known but its value should be in general negative (van der Kamp et al., 2018), a standard variance of 2.0 kJ/mol/K was assumed as it covers a broad range of $\Delta C_p^{\ddagger}$ and with a very low possibility of getting a positive value. The simulated results showed very big variations in the model predictions in all three cases (Figures 11D-E). It means that those uncertainties in the model parameter values largely destroyed the prediction power and reliability of the etcGEM.

In order to reduce those uncertainties in the model parameters, Bayesian statistical learning (Yau and Campbell, 2019) provides an excellent solution. It is a probabilistic framework that has been

successfully applied for quantifying and reducing uncertainties in various fields including deep learning (Kingma and Welling, 2013), ordinary differential equations (Girolami, 2008) and biochemical kinetic models (Miskovic et al., 2019). The approach uses experimental observations (D) to update *Prior* distributions ($P(\theta)$) of model parameters to *Posterior* ones ($P(\theta|D)$). In our case (Figure 12A), the problem can be formulated as: given a *generative model (M)* (etcGEM in this study) corresponding to a set of parameters θ and a set of measurements D(physiology data), with Bayes' theorem the *Prior* distribution of parameters $P(\theta)$ can be updated to a *Posterior* distribution $P(\theta|D)$ through

$$P(\theta|D) = \frac{P(D|\theta)\times P(\theta)}{P(D)} \quad [11]$$

$P(\theta|D)$ is thereby a less uncertain description of the real θ. Since the likelihood term $P(D|\theta)$ is, in most applications, computationally expensive or even infeasible to obtain, the likelihood-free inference methods like Approximate Bayesian Computation (ABC) (Sunnåker et al., 2013) can then be used. The concept of ABC is as follows: Given an observed dataset D and a model specified by $\hat{\theta}$ sampled from the *Prior* distribution $P(\theta)$, if the distance between simulated data $\hat{D}$ and observed D is less than a given threshold ϵ, then this $\hat{\theta}$ is accepted as the one sampled from $P\left(\rho\left(D,\hat{D}\right)<\epsilon\right)$. $P\left(\rho\left(D,\hat{D}\right)<\epsilon\right)$ is often used to approximate the *Posterior* $P(\theta|D)$ when ϵ is sufficiently small. In case of high-dimensional parameter space and/or when the $P(\theta)$ is very different from $P(\theta|D)$, the acceptance rate would be very low and thus this approach becomes computationally expensive to generate a population of $\hat{\theta}$ from $P\left(\rho\left(D,\hat{D}\right)<\epsilon\right)$. Thereby in this section, a sequential Monte Carlo based ABC approach (SMC-ABC) was designed to generate a population of $\hat{\theta}$ sampled from $P\left(\rho\left(D,\hat{D}\right)<\epsilon\right)$ by starting with an larger ϵ and then gradually reducing it to the final smaller one (Check **Paper III** for more details). We refer to the model equipped with $\hat{\theta}$ sampled from $P(\theta)$ or $P(\theta|D)$ as a *Prior* or *Posterior* etcGEM, respectively.

By applying this Bayesian approach to update the $P(\theta)$ of parameters in etcGEM by using three datasets shown in Figure 11 as D. At each iteration, 100 *Posterior* models were obtained and the minimal R^2 score between simulated data $\hat{D}$and experiment $D$was monitored (Figure 12B). In the end, each of the final 100 *Posterior* models has a R^2 score of at least 0.9, which thereby can accurately reproduce the experimental data (Figure 12C-D). In this case, each of *Posterior* models is forced to be able to reproduce those experimental data which may lead to the risk of an overfitted model. While this risk is unavoidable and difficult to be detected, its consequences are minimized by two factors: (1) unlike ML models, etcGEM is a theory-based model with biologically reasonable parameter values in the *Posterior*; (2) the interpretation and prediction were made by ensembling all results from all 100 *Posterior* models, which is a common strategy in ML that can take advantage of overfitted individual models (Peter Sollich, 1996).

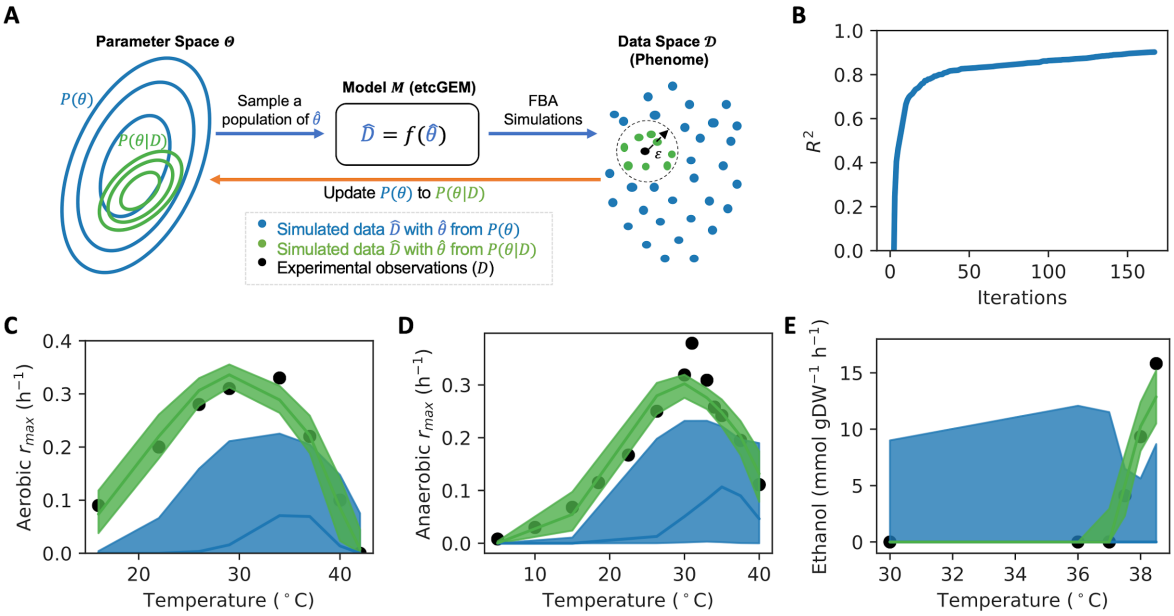

Figure 12 Uncertainty reduction and quantification with Bayesian statistical learning approach. (A) Overview of Bayesian genome scale metabolic modelling. (B) Minimal coefficient of determination score (R^2) of *Posterior* models during SMC-ABC approach. (CD) Simulated (C) aerobic and (D) anaerobic growth rates in batch cultivations at various temperatures with *Prior* and *Posterior* etcGEMs. (E) Simulated ethanol secretion flux in chemostat at various temperatures. In (CDE), lines indicate median values and shaded areas indicate regions between the 5-th and 95-th percentiles.

ML for analyzing simulation results

Direct comparison between *Prior* and the final *Posterior* distributions revealed that in all three parameter categories (T_m, T_{opt} and $\Delta C_p^{\ddagger}$), a reduced variance in the updated parameters was more likely than a change in mean values (Figure 13A). Actually, during the iterations in the SMC-ABC approach (Figure 12B), 21,504 models were generated and simulated in total. Analysis of those models would give more hints about the important factors for the performance of etcGEM. First, the unsupervised learning algorithm principle component analysis (PCA) was applied (Figure 13B). Although the first two components only explain less than 2% of the total variance in the parameter sets of those 21,504 models, a clear trend of how the *Priori* distributions were gradually updated to distinct *Posterior* distributions can still be observed (Figure 13B). Then to identify the important parameters that drove the improvement of etcGEM in the SMC-ABC approach, a supervised ML approach that enables the scoring of feature importance can be applied. A random forest model was optimized and trained by taking 2,292 parameters of those etcGEMs as input and the R^2 scores between simulated and experimental data obtained in the SMC-ABC approach were used as response values. The feature importance scores were then extracted from the obtained model. It revealed that out of all three parameter types, the largest contribution to the improved *Posterior* etcGEM performance during the Bayesian approach was from enzyme T_{opt}s (Figure 13C). A list of the top 20 most important parameters as well as their gene names can be identified (Figure 13D). This means that a more accurate estimation of those parameters is critical for the

performance of yeast etcGEM (correlation), while it doesn't mean that those genes are important for the thermotolerance of yeast metabolism (causality).

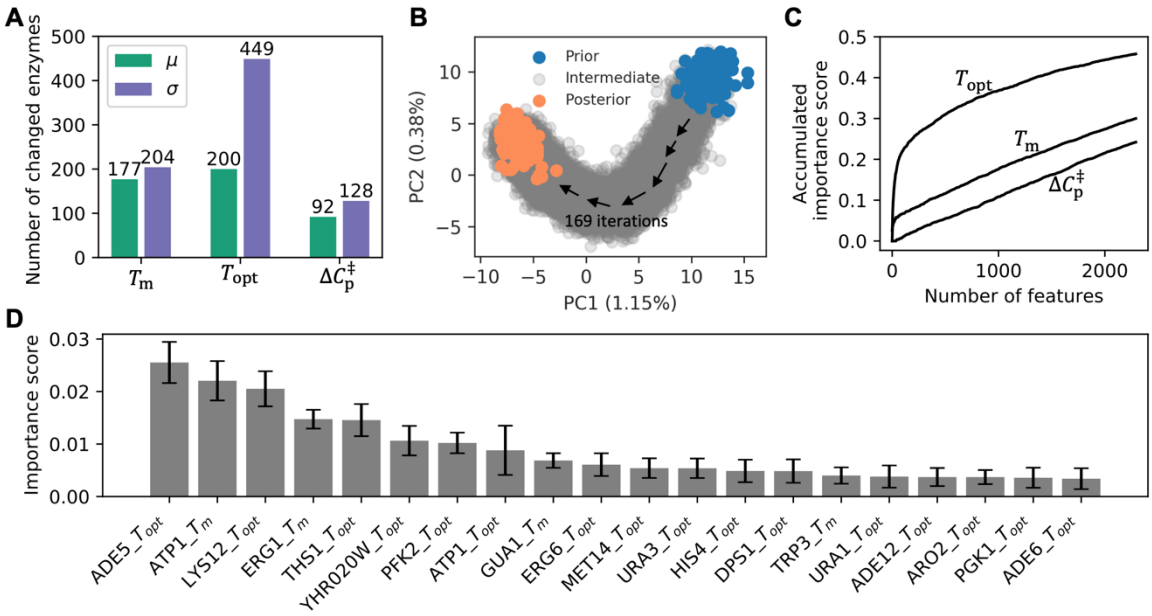

Figure 13. (A) The number of enzymes with a significantly changed mean and variance in T_m, T_{opt} and $\Delta C_p^‡$ between *Prior* and *Posterior*. (B) PCA on 21,504 parameter sets ($\hat{\theta}$) sampled in the SMC-ABC. (C) The accumulated importance score from the random forest model for each of the three parameter categories. (D) The top 20 parameters with the highest importance score were shown.

Making predictions with uncertainties quantified

Although the uncertainties in some parameters of etcGEM have been reduced via Bayesian statistical learning approach (Figure 13A), there are still large uncertainties in the parameters in the updated etcGEM (Figures 14ABC). An intuitive example is the enzyme squalene epoxidase ERG1 (Figures 14DEF): variances of all three parameters were significantly reduced in the *Posterior*, but they are still relatively large. Thereby, it's important to also quantify the uncertainties when using *Posterior* models for predictions or interpretations. In this section, predicting the most growth-limiting enzymes at superoptimal temperatures will be discussed.

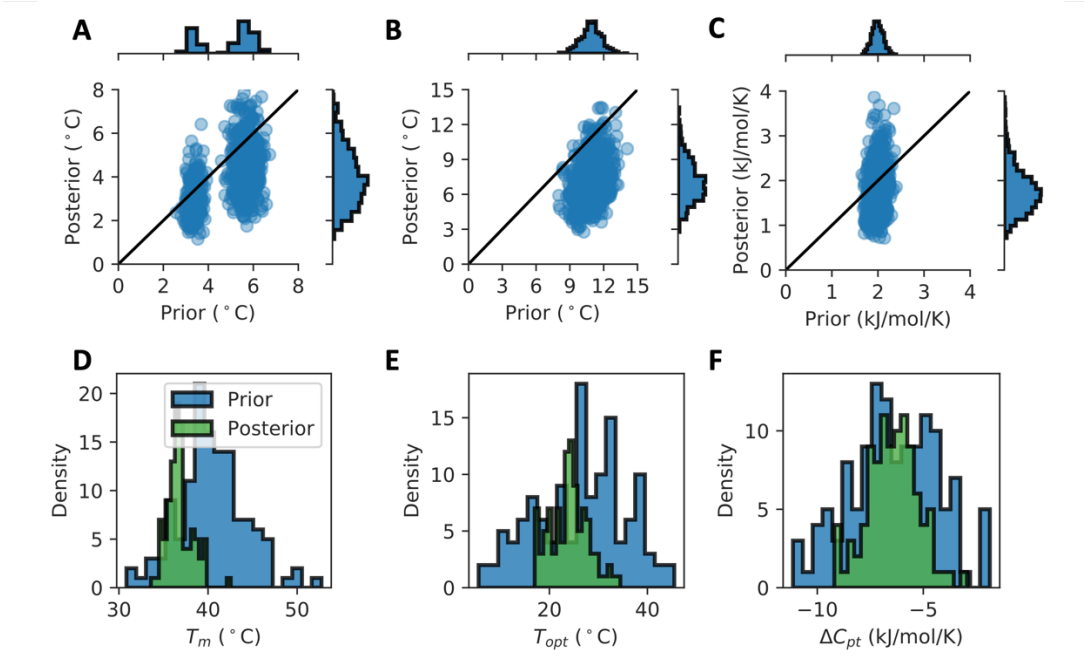

Figure 14. Large uncertainties in the *Posterior* etcGEM. (ABC) Parity plots of updated standard variance of enzyme (A) T_ms, (B) T_{opt}s, and (C) $\Delta C_p^{\ddagger}$s. (DEF) *Prior* and *Posterior* distributions of those three parameters of enzyme ERG1.

Flux sensitivity analysis was performed to calculate a coefficient for each enzyme which describes to what extent the change in enzyme activities affects the cell growth at a given temperature. Among all 764 enzymes in the model, the ERG1 showed coefficients an order of magnitude higher than others at 40 °C and 42 °C (Figure 15A), suggesting that it is the most flux-controlling enzyme at high temperatures. Further simulations showed that the removal of temperature constraints on ERG1 (making it temperature insensitive) increased the specific growth rate at both temperatures (Figure 15B). To experimentally validate the effect of ERG1 on cell growth at those two temperatures, a homolog of *ERG1* from an thermotolerant yeast *Kluyveromyces marxianus* (KmERG1), which can survive at temperatures higher than 40 °C (Lane and Morrissey, 2010) was introduced to replace the wide-type ERG1 in *S. cerevisiae*. The experimental results showed that the strain with KmERG1 indeed showed significantly better growth than the wild type at 40°C after 2 generations of adaptation, which proved our predictions. However, no significant growth difference was detected at the lethal temperature 42 °C, indicating that the *Posterior* models are still to be further updated in the future. A bunch of *Posterior* models did predict that there is a very small growth rate at 42 °C for the strain with temperature insensitive ERG1 (Figure 15B), suggesting that further update of current etcGEM with those new experimental data is feasible in the future.

In contrast to the conventional GEM approach that uses a single model to make predictions, Bayesian GEM uses many GEMs and takes all the predictions together into consideration. Thereby, the variation in the predicted results of all those models gives an estimate of the uncertainty in the prediction. This is important information for the decision making for the experimental validation, since the validation is usually time- and resource-consuming.

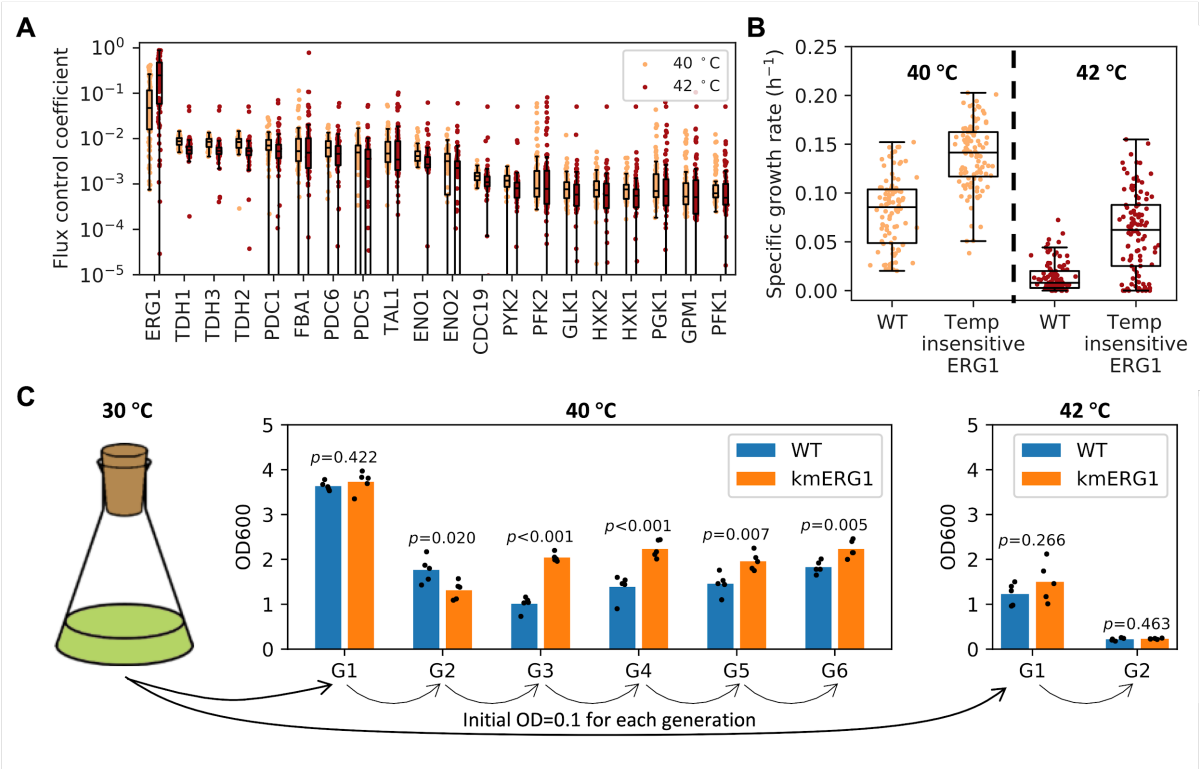

Figure 15. **Predicting growth rate-limiting enzymes.** (A) 20 enzymes with the highest flux sensitivity coefficients at 40 °C and 42 °C. Each dot represents the prediction from one *Posterior* etcGEMs. (B) Predicted maximal specific growth rate of wide-type yeast and the one without any temperature constraints on ERG1 enzyme at 40 °C and 42 °C. (C) The effect of KmERG1 expression on thermotolerance of *S. cerevisiae*. The strains were cultivated at 40 °C or 42 °C for many generations to reach the steady state of growth.

Challenges in the development of regression models in systems biology

Regression is one of the most commonly used supervised ML approaches not only in systems biology such as estimation of parameters in theory-based models (e.g. etcGEM) and prediction of quantitative traits from organism genomes, but also in many other closely related biological fields including metabolic engineering (Zhang et al.; Zhou et al., 2018), protein engineering (Capriotti et al., 2005b; Romero et al., 2013) and medicine (Ammad-ud-din et al., 2017; Barretina et al., 2012; Tan, 2016). Several characteristics shared by most biological datasets challenge the application of ML applications: (1) the number of biological samples is usually small since they are time- and resource-consuming to generate and/or it's unethical to perform lots of experiments such as animal tests; (2) the quality of training samples are low due to the noise in samples; (3) complex mechanisms underlying the modelling task make it difficult to engineer relevant features or to train a predictive model. Taking the enzyme T_{opt} as an example, its experimental determination requires expression and purification of proteins and activity measurement in temperature-controlled conditions, each of those steps needs to be carefully designed and optimized. This makes it challenging to measure the T_{opt}s of hundreds or thousands of enzymes within a reasonable time. So far, the major data source for enzyme T_{opt} has been the published papers in the past decades. Databases like BRENDA have made great contributions by collecting those enzyme information from the published literature (Jeske et al., 2019). In the BRENDA (version 2018), there were about 5,600 unique enzymes that were found with associated T_{opt} records **(Paper V)**. This provides a solid basis for the development of ML models. However, the quality of those T_{opt} values is another concern. We noticed that many T_{opt} values are not real temperature optimum since they were obtained by directly using room temperature or OGT of its host organism (for instance 37 °C, Figure 9A) without optimization. Lastly, the mechanism behind how an enzyme T_{opt} is determined by its primary protein sequence remains unclear. In this case, directly engineering relevant features based on biological knowledge remains challenging for the development of a predictive T_{opt} prediction model.

In this chapter, I firstly evaluated the effect of the presence of noise in response values on the development of regression models (**Paper IV**). Then I showed an example of applying deep transfer learning to resolve the challenges of small datasets and lack of biological knowledge-based features (**Paper V**).

The theoretical upper bound for the performance of regression models

In biological datasets, the sample labels are typically real numbers generated through experimental measurements that are inextricably associated with noise and errors (Bruggeman and Teusink, 2018; Harris and Smith, 2009; Tsimring, 2014), thus intuitively a regression model that can perfectly predict those values cannot be achieved. There should exist an upper bound for the performance of the ML model we can expect. Knowing this upper bound would give hints about whether the maximal performance has been reached on a particular dataset, or whether further model improvement is possible.

There have been some attempts made to estimate this upper bound. Given a set of samples with experimentally determined labels $\{y_{obs,i}\}$ and corresponding unknown real labels $\{y_i\}$, Assume a normally distributed experimental noise term $\varepsilon_{y,i} \sim N(0, \sigma_{y,i})$ for all samples: $y_{obs,i} = y_i + \varepsilon_{y,i}$ ($y_i \in R$). Fariselli and coworkers (Benevenuta and Fariselli, 2019; Montanucci et al., 2019) assumed the best possible model is $y = x$ in which x are the values collected from another set of experiments conducted at identical conditions. Under this assumption, the expectation of the upper bound for mean squared error (*MSE*) is $2\overline{\sigma_y^2}$ and coefficient of determination (R^2) is $\frac{\sigma_{obs}^2 - 2\overline{\sigma_y^2}}{\sigma_{obs}^2}$, where $\overline{\sigma_y^2}$ is the average variance of all sample noise and σ_{obs}^2 is the variance of the observed values. We refer this expected upper bound for R^2 as $\langle R^2 \rangle_{FP}$ and the expected lower bound for *MSE* as $\langle MSE \rangle_{FP}$ hereafter.

Here I proposed a different assumption about the best model performance we can expect: the best model performance can be achieved when (1) a complete set of features is known as $x_i \in R^k$ for each sample; (2) the real function $y = f(x)$ that can accurately calculate the real value of label y_i from the complete set of features x_i is obtained. With this assumption, the R^2 of the model $f(x)$ is given by

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_{obs,i} - \hat{y}_{obs,i})^2}{\sum_{i=1}^{m}(y_{obs,i} - \bar{y}_{obs})^2} = 1 - \frac{\sum_{i=1}^{m}(y_{obs,i} - f(x_i))^2}{\sum_{i=1}^{m}(y_{obs,i} - \bar{y}_{obs})^2} \quad [12]$$

where m is the number of samples. Although it is not possible to obtain an exact value from the above equation, since the real values $f(x_i)$ are unknown, we can instead obtain the expectation of R^2. Since $f(x_i) = y_i$, $y_{obs,i} - f(x_i) = y_{obs,i} - y_i = \varepsilon_{y,i}$, the expectation is then given by

$$\langle R^2 \rangle = 1 - \langle \frac{\sum_{i=1}^{m} \epsilon_{y,i}^2}{\sum_{i=1}^{m}(y_{obs,i} - \bar{y}_{obs})^2} \rangle = 1 - \sum_{i=1}^{m} \langle \frac{\epsilon_{y,i}^2}{\sum_{j=1}^{m}(y_{obs,i} - \bar{y}_{obs})^2} \rangle \quad [13]$$

Since $\epsilon_{y,i}$ is normally distributed with a zero-mean and variance of $\sigma_{y,i}^2$, then $\frac{\varepsilon_{y,i}}{\sigma_{y,i}}$ follows a standard normal distribution. Thereby $(\frac{\varepsilon_{y,i}}{\sigma_{y,i}})^2$ follows a chi-squared distribution with a degree of 1 ($\chi^2(1)$). The numerator becomes $\epsilon_{y,i}^2 = \sigma_{y,i}^2 \frac{\epsilon_{y,i}^2}{\sigma_{y,i}^2} \sim \sigma_{y,i}^2 \cdot \chi^2(1)$. We assume that the variance of the observed values $y_{obs,i}$ is normally distributed with a variance of σ_{obs}^2, then

$$\sum_{j=1}^{m}(y_{obs,i} - \bar{y}_{obs})^2 \sim \sigma_{obs}^2 \cdot \chi^2(m-1) \quad [14]$$

The ratio between two chi-squared distributions is an *F* distribution multiplied by the ratio between their degrees of freedom, thereby

$$\langle R^2 \rangle = 1 - \sum_{i=1}^{m} \frac{\sigma_{y,i}^2}{\sigma_{obs}^2} \langle \frac{\chi^2(1)}{\chi^2(m-1)} \rangle = 1 - \sum_{i=1}^{m} \frac{\sigma_{y,i}^2}{\sigma_{obs}^2} \frac{1}{m-1} \langle F(1, m-1) \rangle \quad [15]$$

Since $\langle F(1, m-1) \rangle = \frac{m-1}{m-3}$, then

$$\langle R^2 \rangle = 1 - \frac{1}{m-3} \sum_{i=1}^{m} \frac{\sigma_{y,i}^2}{\sigma_{obs}^2} = 1 - \frac{m}{m-3} \frac{\overline{\sigma_y^2}}{\sigma_{obs}^2} \quad [16]$$

in which $\overline{\sigma_y^2} = \frac{1}{m}\sum_{i=1}^{m}\sigma_{y,i}^2$. As the number of examples in ML is usually very large (m >> 1), we can approximate the final equation for upper bound estimation as

$$\langle R^2 \rangle \approx 1 - \frac{\overline{\sigma_y^2}}{\sigma_{obs}^2} = \frac{\sigma_{obs}^2 - \overline{\sigma_y^2}}{\sigma_{obs}^2} \quad [17]$$

With the similar approach, we can obtain the expectation for MSE as

$$\langle MSE \rangle = \frac{1}{m}\sum_{i=1}^{m}\sigma_{y,i}^2 = \overline{\sigma_y^2} \quad [18]$$

We refer this expected upper bound for R^2 as $\langle R^2 \rangle_{LG}$ and the expected lower bound for MSE as $\langle MSE \rangle_{LG}$.

Obviously, $\langle MSE \rangle_{LG}$ is half of $\langle MSE \rangle_{FP}$ and $\langle R^2 \rangle_{LG}$ is larger than $\langle R^2 \rangle_{FP}$. I then performed Monte Carlo simulations to directly compare $\langle R^2 \rangle_{FP}$ and $\langle R^2 \rangle_{LG}$ when applying them to estimate the upper bound of regression models. Briefly, a random dataset $\{x_i, y_{obs,i}\}$ was generated from a known real function $f(x)$ with added experimental noise $\sigma_{y,i}$. For this dataset $\langle R^2 \rangle_{FP}$ and $\langle R^2 \rangle_{LG}$ were calculated, and then the $R^2$of a support vector machine regression model trained on the data was calculated via a 2-fold cross validation approach (R_{ML}^2). This process was repeated for 1000 iterations. The simulations illustrated two key points (Figure 16). First, the simulations show that $R_{ML}^2$is higher than $\langle R^2 \rangle_{FP}$, which is contrary to the expectation if $\langle R^2 \rangle_{FP}$ is a true upper bound (Benevenuta and Fariselli, 2019; Montanucci et al., 2019). Second, R_{ML}^2 is smaller than but close to $\langle R^2 \rangle_{LG}$, which confirms that $\langle R^2 \rangle_{LG}$ gives a good estimation of the model performance upper bound. This shows that $\langle R^2 \rangle_{LG}$ gives a more accurate estimation of the upper bound for the performance of ML models than $\langle R^2 \rangle_{FP}$.

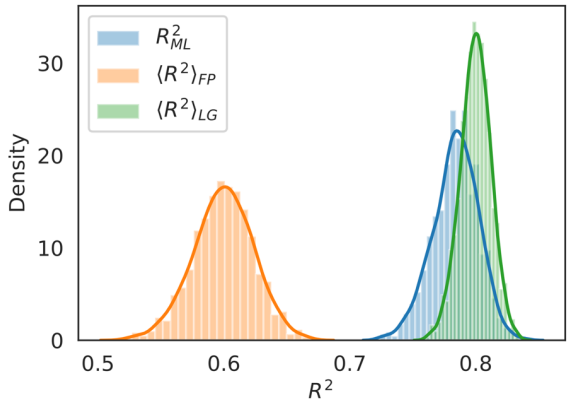

Figure 16. Monte Carlo simulation on the upper bound of R^2. A linear real function $y = 2x + 1$ was used.

$\langle R^2 \rangle_{LG}$ solely depends on two properties of the dataset: (i) the true variance of the observed response values (σ_{obs}^2) and (ii) the average variance of experimental noise of all samples ($\overline{\sigma_y^2}$). In practice, σ_{obs}^2 and $\overline{\sigma_y^2}$ are unknown and have to be approximated from the dataset. $\sigma_{y,i}$ can be

approximated with the standard error (SE) of n replicates, which represent the standard error of the mean, and σ^2_{obs} can be approximated as the variance of the target values (Figure 17).

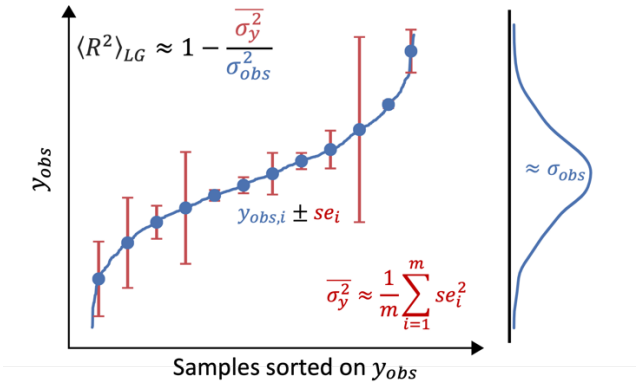

Figure 17. Schematic diagram depicting the estimation of the upper bound of model performance $\langle R^2 \rangle_{LG}$ based on experimental label noise. Data shown were randomly generated, se_i denotes standard error of sample i.

In practice, we are not only facing the challenges of noisy response values, but also noisy features and the incomplete set of features. However, for the latter two conditions it's challenging to derive a simple equation as for $\langle R^2 \rangle_{LG}$. Thereby, I performed Monte Carlo simulations to show how the noise in features and the completeness of feature sets affect the performance of regression models (Figure 18). In the first simulation (Figure 18A), different noise levels were introduced into the only feature x, R^2_{ML} of a support vector machine regression model was calculated via a 2-fold cross validation approach. It shows that the noise in features has a dramatic effect on the performance of ML models. When the noise in features is huge (e.g. $\sigma_x^2 = 1.0$ in Figure 18A), $\langle R^2 \rangle_{LG}$ is not a realistic objective for the performance of ML models since it cannot be achieved unless the noise in features is reduced to 0. In the second simulation, the Monte Carlo simulations were used to address two questions: (1) how the completeness of the feature set affects the model performance? (2) if we can improve the model performance by removing the samples with the largest noise? A linear function with 10 noise-free features was used. The response values have different levels of noise. It shows that model performance generally improved as noisy samples were removed. However, an interesting observation is that the degree to which the models improve upon removal of noisy samples depends on how many features were used to train them. For instance, if only a small fraction of relevant features were used (2/10 in Figure 18B), the removal of the noisiest samples did not improve model performance. In contrast, when the majority of the relevant features were known (8/10 and 10/10 in Figure 18B), the removal of noisy samples significantly improved the model performance. These results indicate that when R^2_{ML} is very far from the $\langle R^2 \rangle_{LG}$ upper bound, model performance can be readily improved by obtaining additional or more relevant features, as opposed to performing data cleaning to reduce sample noise. However, removal of noisy samples is at a cost of reducing training samples.

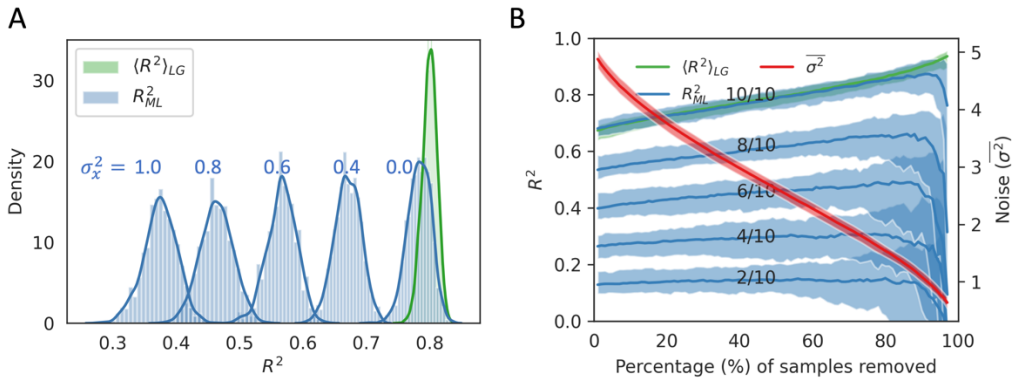

Figure 18 Monte Carlo simulations on (A) R^2 by assuming different levels of feature noise (a linear real function $y = 2x + 1$was used); and (B) data cleaning via gradually removing the samples with the largest $\sigma_{y,i}$. n/10 indicate that n features out of a complete set of 10 features are used to train and validate the model. Noise values are given as the average variance of all samples ($\overline{\sigma_y^2}$). A linear real function $f(x) = \sum_{i=1}^{10} x_i$ was used.

Above conclusion can be illustrated through the prediction of enzyme T_{opt}. Two datasets with different levels of noise in response values were generated. A first raw dataset comprising the T_{opt} of 5,343 individual enzymes was collected from the BRENDA database (Jeske et al., 2019). Using enzymes for which T_{opt} values had been measured in multiple experiments the experimental noise $\overline{\sigma_y^2}$ was estimated as $(7.84\ °C)^2$ and $\sigma_{obs}^2$was $(16.32\ °C)^2$ in this dataset. Given these values for $\overline{\sigma_y^2}$ and σ_{obs}^2 the corresponding $\langle R^2 \rangle_{LG}$ upper bound was 0.77. As mentioned earlier, some enzyme T_{opt} present in BRENDA is not real enzyme temperature optimum. Those records were marked with "assay at" in the "comment" field in the database. After removal of those values which were deemed less likely to represent true catalytic optima, a second dataset containing 1,902 enzymes was obtained. With the same approach, the experimental noise $\overline{\sigma_y^2}$ was estimated as $(7.22\ °C)^2$ and the calculated $\langle R^2 \rangle_{LG}$ as 0.85. A comprehensive feature set containing 5,494 features belonging to 20 subsets (the same ones as used in Figure 9C) was extracted based on the protein sequences. Previously we have shown that OGT provides additional information to the sequence-based features for the prediction of enzyme T_{opt} (Figure 9). Thereby including OGT into a sequence-based feature set improved the completeness of the feature set. Performance of five regression algorithms on each of 20 feature subsets were obtained via a 5-fold cross validation approach (Figure 19). I found that for each of these feature subsets reducing the noise in T_{opt} only improved model performance when OGT was included as an additional feature. This was consistent with results revealed in Monte Carlo simulations (Figure 18B), showing that noise reduction is only beneficial with more complete feature sets. The best model achieved a R^2_{ML} of 0.61, which is around 71% of $\langle R^2 \rangle_{LG}$, indicating that further improvement is still possible.

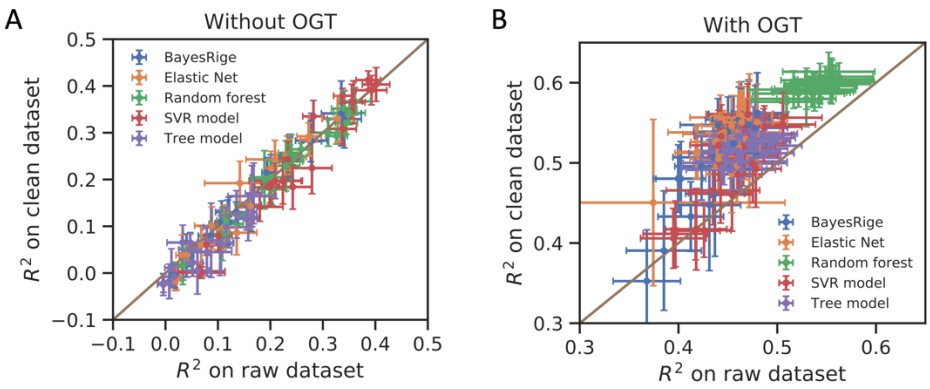

Figure 19. Comparison of model performance on raw and clean dataset (A) with; and (B) without OGT as one of the features. Each data point represents one of five regression algorithms trained on one of 20 subsets of features. Error bars show the standard deviation of R^2 scores obtained in 5-fold cross validation.

In real-world applications, it may not be feasible to obtain the experimental noise in the response values, as they are not reported or difficult to extract from literature. An example for this is the yeast quantitative traits measured in a high-throughput way (Peter et al., 2018). In this paper, the growth profiles of 971 sequenced *S. cerevisiae* isolates under 35 stress conditions had been measured, while the noise associated with those traits are not available. Predicting yeast phenotypes directly from genomes has been a very challenging task (Bae et al., 2016; Crossa et al., 2010; Jelier et al., 2011; Makowsky et al., 2011; Morota et al., 2014). In this case, how can we apply $\langle R^2 \rangle_{LG}$ for those datasets?

We noticed that since those traits are measured in the same lab with the same method, it's reasonable to assume that they have the same level of experimental noise $\overline{\sigma_y^2}$. Since $\langle R^2 \rangle_{LG}$ is an upper bound estimate $R^2_{ML} \leq \langle R^2 \rangle_{LG}$ holds true. From this we obtain that $\overline{\sigma_y^2} \leq (1 - R^2_{ML}) \times \sigma^2_{obs}$. For multiple datasets with the same level of experimental noise, $\overline{\sigma_y^2} \leq min(\{(1 - R^2_{ML,i}) \times \sigma^2_{obs,i} | \, i = 1, .., s\})$, in which s is the number of the datasets. In this way it is possible to estimate the maximal level of the experimental noise based on the ML results, and then further use it to obtain the minimal value of $\langle R^2 \rangle_{LG}$ (referred to as $\langle R^2 \rangle_{LG,min}$). In this special case, $\langle R^2 \rangle_{LG}$ could be any value between $\langle R^2 \rangle_{LG,min}$ and 1.0. $\langle R^2 \rangle_{LG,min}$ would be useful when $\langle R^2 \rangle_{LG,min}$ approaches 1.0 and one can use it to check if there is still room to further improve R^2_{ML} for some datasets. $\langle R^2 \rangle_{LG,min}$ was then applied to the yeast quantitative traits dataset and showed that for the prediction of most traits further improvement is still possible (Figure 20).

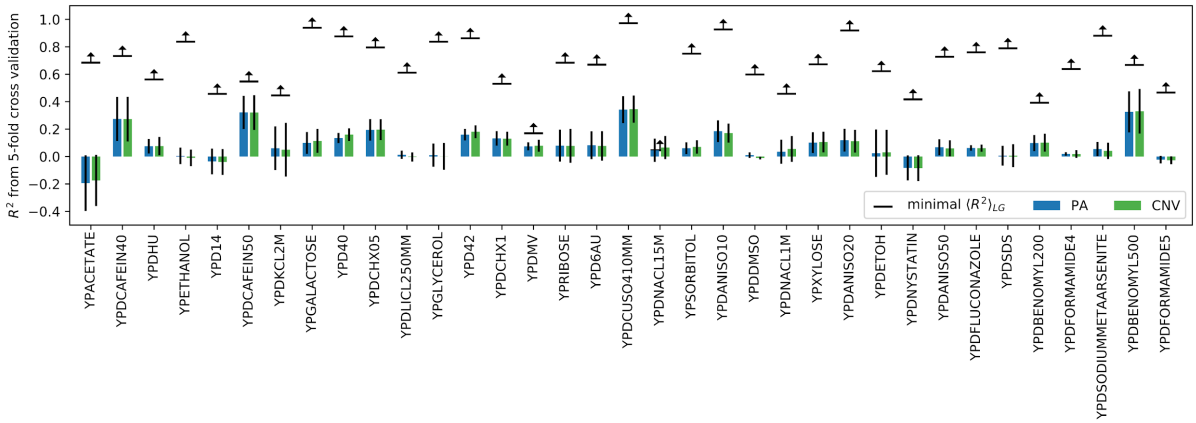

Figure 20. Prediction of 34 quantitative traits of *S. cerevisiae* from its pan-genome composition (Figure 6). The gene presence/absence (P/A) and copy number variations (CNV) in the pan-genome were used as input features.

At last, a major concern for the application of $\langle R^2 \rangle_{LG}$ is that $\langle R^2 \rangle_{LG}$ was derived based on the assumption of normally distributed response values, while in the real-world datasets, this assumption may not hold. With Monte Carlo simulations, we show that it seems to be applicable for non-normal distributions (Figure 21). Future theoretical analysis is required to test the applicability of $\langle R^2 \rangle_{LG}$ on non-normal distributions.

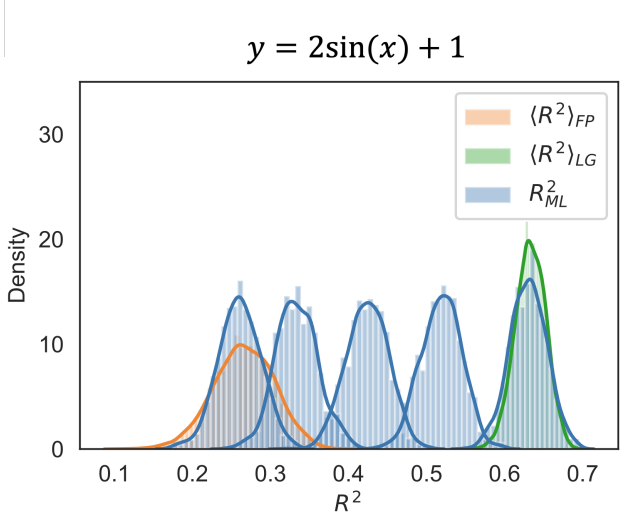

Figure 21. R^2 by assuming different levels of feature noise for the dataset generated by a nonlinear function $y = 2sin(x) + 1 + \varepsilon$. Since x amd ε are normally distributed, the nonlinear transformation gives a non-normally distributed $\{y_{obs,i}\}$.

Deep transfer learning for small biological datasets with limited features

In the last section, we have discussed in addition to the noise in sample features and labels, the completeness of feature sets plays a critical role in the development of ML models. Extracting features for biological samples remains challenging due to the poor understanding of mechanisms in certain tasks. For instance, predicting enzyme T_{opt}, as already shown in Figure 9 and details in **Paper IV**, using a very comprehensive feature set with 5,494 features was found no better than just using compositions of 20 amino acids. Using OGT as an additional feature greatly improved the performance of ML models, however it limits the application scope of the final model only to those native enzymes from microorganisms with known OGT. Ideally, we prefer an accurate model that directly takes the enzyme sequence as input and output the T_{opt} value. While this is challenging for classical models like support vector machines or random forests which rely on human-designed features, end-to-end deep neural networks (DNN) can directly take the enzyme sequences as input and output the T_{opt} value (Figure 22). With DNN, an enzyme sequence is encoded with a binary matrix with a size of $20 \times L$ (one-hot encoding), in which L is the length of the protein sequence. The whole neural network can be seen as the combination of two parts: layers for feature extraction followed by layers for classification or regression (Tang et al., 2019). This thereby is a perfect approach to get rid of the domain-knowledge based feature engineering, however it requires a large number of training samples to train a DNN, which unfortunately is not possible with most of the biological datasets.

Deep transfer learning provides a promising solution to the insufficient training data (Tan et al., 2018). The main concept of transfer learning is to transfer the knowledge gained while solving one problem (source task) to a **different** but **related** problem (target task) (Weiss et al., 2016). It takes a pre-trained DNN on a large dataset and repurposes it to another task which has only a small number of training samples available. An intuitive real-world example for transfer learning is that it's much easier to teach a person who already has a driving license for small cars to drive a truck than one who doesn't have such experience at all. In deep transfer learning, this is done by fine-tuning part of all the pre-trained DNN layers (e.g. the classifier/regressor part in Figure 22) instead of training it from scratch (randomly initialized weights). In this section, I will use examples of predicting enzyme T_{opt} and protein T_m to demonstrate its power in biological applications (**Paper V**).

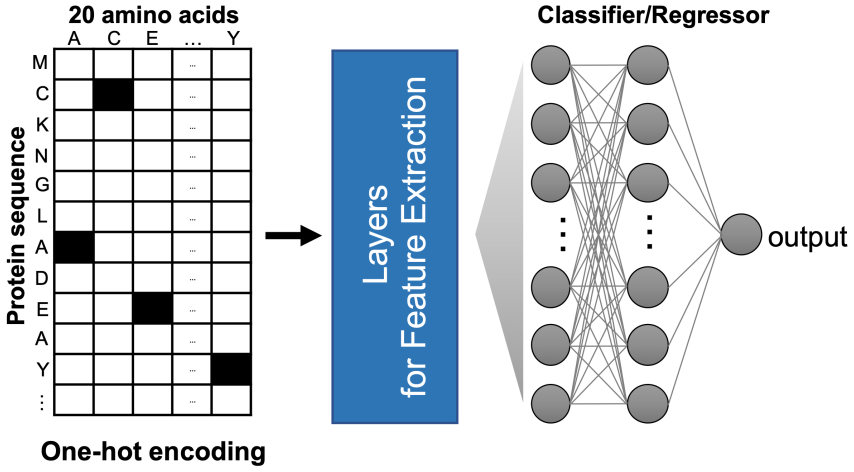

Figure 22. Illustration of a deep neural network that takes one-hot encoded protein sequences as input. Black and white boxes represent 1 and 0 in one-hot encoding, respectively.

In the case of predicting temperature-related protein properties (T_m and T_{opt}), how to choose a related source task is the key. In **Papers II&IV** and elsewhere (Engqvist, 2018), OGT has been found to be closely related to the enzyme T_{opt}. Particularly in **Paper II**, about 6.5 million enzymes in BRENDA were annotated with OGT values (Figure 23A), even after removal of low quality or similar sequences there are still 3.0 million enzymes in the dataset (Figure 23B). This provides an ideal dataset to train a deep neural network in modelling how variations in environmental temperatures affect the evolution of enzyme sequences. The resulting model would capture the sequence determinants for enzymatic thermal adaptability, which can then be utilized in the tasks of predicting other related thermal properties of enzymes like T_{opt}. To demonstrate this proposal, a residual neural network (ResNet) (He et al., 2016) was optimized and trained (Figure 23C). The model achieves a *Pearson*'s correlation coefficient of 0.77 and R^2 score of 0.59 on the hold out OGT test dataset (Figure 23D).

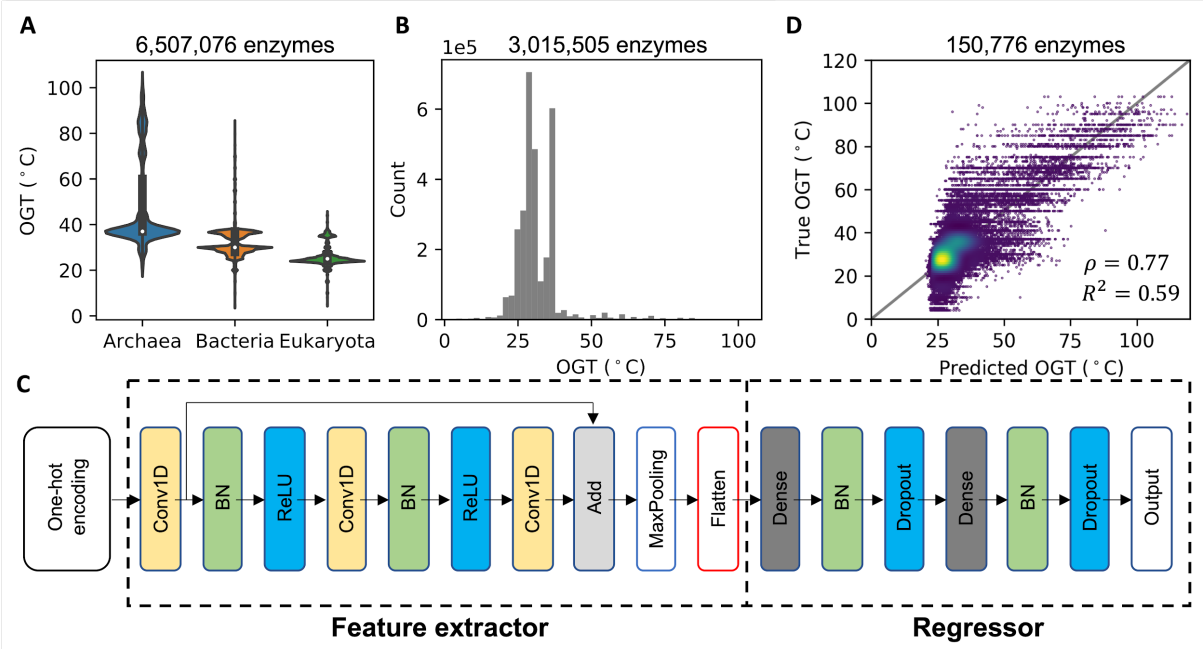

Figure 23. Pre-training with enzyme sequences labeled with OGT. (A) Violin plot of distributions of OGT-annotated enzymes from microorganisms belonging to three domains. (B) Distribution of OGT values after filtering out low-quality and redundant sequences. (C) The ResNet architecture used. (D) The comparison between predicted and true OGT values of enzymes in the hold out test dataset. ρ denotes *Pearson*'s correlation coefficient and R^2 denotes the coefficient of determination.

With this pre-trained model, I then tested its application in predicting enzyme T_{opt} and protein T_m via a transfer learning approach. There are 1,902 samples in enzyme T_{opt} dataset collected from BRENDA (**Paper IV**) and 2,506 samples in protein T_m dataset collected from literature (Leuenberger et al., 2017). Two transfer-learning approaches were tested: (1) only fine-tune the weights in the regressor part (FrozenCNN); (2) fine-tune the whole neural network (TuneAll). The results showed that transfer learning approaches outperform both (1) classical regression models with biological knowledge-based features (iFeatures, 5,494 features) or a deep learning-based feature set (UniRep, 5,700 features), and (2) the deep learning model trained from scratch (Figure

24). The pre-trained model itself showed no prediction power for enzyme T_{opt} and protein T_m before any fine-tuning (FrozenAll in Figure 24).

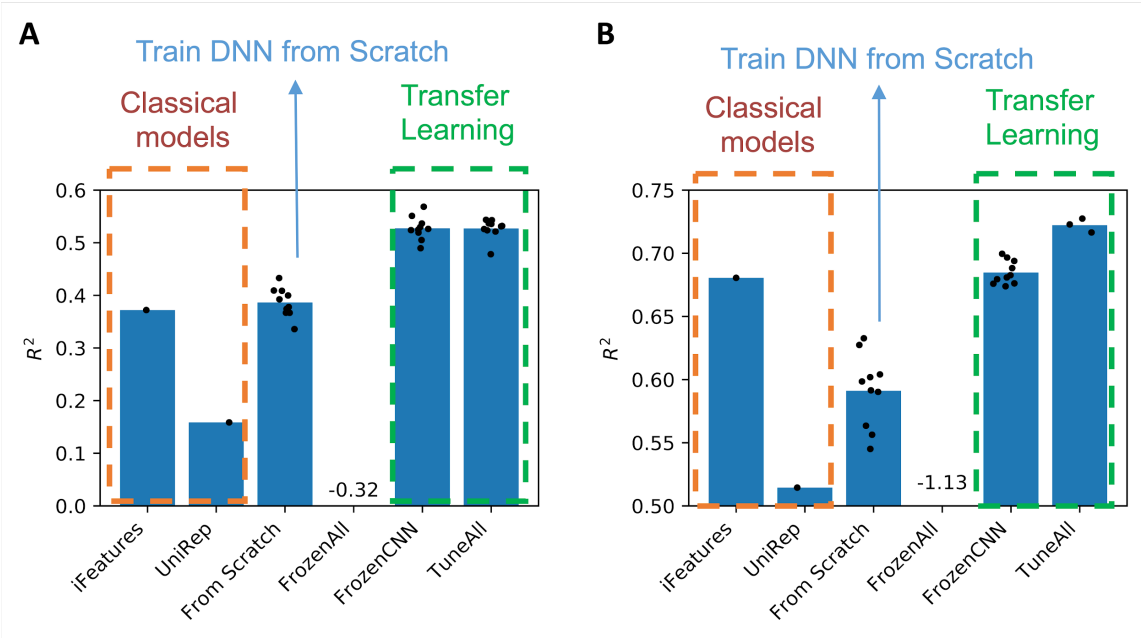

Figure 24. R^2 score of ML models on hold out test datasets. (A) Enzyme T_{opt} dataset; (B) Protein T_m dataset.

In the above examples, the deep transfer learning approach addressed limitations of training samples and feature engineering in biological datasets. The progress in these two applications is achieved from the utilization of the pre-trained OGT model, since OGT can be seen as a low-quality estimation of enzyme T_{opt} and protein T_m, given the fact that a protein should be at least functional and mostly remains in the native state under the OGT of its host organism. Deep transfer learning holds a great application potential in systems biology. For example of predicting RNA secondary structure, Singh J *et al. (Singh et al., 2019)* used a big dataset with more than 10,000 RNAs whose structures were obtained via comparative analysis to pre-train a deep neural network and then fine-tuned with a small dataset with less than 200 high-resolution RNA structures. The resulting model outperforms previous approaches. Similar approaches can, in principle, also applied to predicting enzyme turnover number (k_{cat}), for example empirically estimated less-accurate k_{cat} such as ones from the approach used in GECKO (Sánchez et al., 2017) can be used to pre-train a DNN and then further fine-tuned the resulting model with available experimentally determined values. Another interesting application is that Stumpf P. S. *et al.* applied transfer learning to transfer a DNN model trained on mouse dataset to human dataset (Stumpf et al., 2019). These examples clearly demonstrated that deep transfer learning is a promising approach in resolving the challenges in sample size and feature engineering of biological datasets.

Summary and Perspectives

To answer the question of "***What is life***", systems biologists seek to use mathematical modeling to create an *in-silico* model that can simulate and predict the behaviour of lives from a molecular level. If we quote from Richard Feynman "***What I cannot create, I do not understand***", understanding life means that we can at least create a fully interpretable theory-based model to describe life. As we are now still at the very early-stage to this goal, most biological systems are poorly understood, and a fully theory-based model may not be feasible yet. Compared with the development of theories, generation of data about the systems behaviours is easier and already available for many biological systems, especially due to the accumulated data from molecular biology in the past decades and recent development of high-throughput technologies. Thereby, data-driven ML approaches can be a promising alternative or complement to the theory-based models at this early-stage. Depending on the availability of data and theory about the biological systems, there are three scenarios where ML can be directly implemented (Figure 4B). In the first scenario, where biological data is abundant, while the underlying mechanisms remain insufficient for building a theory-based model, training supervised ML models would not only yield an accurate predictor for system behavior but also suggest the most important features for further investigation. I used two examples of predicting yeast strain types from the pan-genome features (**Paper I**) and microorganism optimal growth temperatures from proteome features (**Paper II**), to showcase the applications of ML approaches. In the second scenario where there are theory-based models but lack of high-quality experimental data for parameterizing the model, ML can be implemented in many different ways. In this thesis I used the example of modelling the thermosensitivity of yeast metabolism (**Paper III**) to show that ML can be used to (1) train regression models to predict the missing parameter values in the theory-based model; (2) analyze the simulation results from theory-based models with both supervised and unsupervised ML approaches; (3) quantify and reduce the uncertainties in the model parameters and thereby improve the performance of the theory-based model. In this thesis, the third scenario, where both theory and data are abundant, is not covered since there are not many such biological systems at this stage and the ML approach may not be as irreplaceable as for the former two scenarios. Still, ML holds the application potential in such a scenario, such as training surrogate models to speed up the simulation of theory-based models and applying ML approaches to analyze the simulation results.

There are also other potential applications of ML that are not covered in this thesis. For some biological systems belonging to the second scenario, only part of the systems can be modelled with theory-based models. In such cases, ML models and theory-based models can be used in a compositional way. For example, to understand the effects of diverse metabolite supplementations on the antibiotic half-maximal inhibitory concentrations (IC_{50}) in *E. coli*, we can of course train a ML model that takes the supplementation profiles as input and predict the IC_{50} values. However, such a model is not that useful since the objective is to understand the underlying mechanisms instead of to develop a predictive model. The theory-based model GEM is already available for *E. coli* metabolism, but it cannot be directly used to predict IC_{50}, indicating that there is a knowledge gap between metabolism and IC_{50} of antibiotics. In this case, Yang J. H. *et al*. used the simulated results from GEM as features to develop a regression model to predict IC_{50}. By interpreting the resulting models, important pathways for IC_{50} were identified and then used for experimentally

validation (Yang et al., 2019). Another similar example is predicting the drug side effect (Shaked et al., 2016).

Development of ML models for biological systems is facing many challenges. First of all, noise in features and labels is very common in biological samples. In this thesis, I addressed a very fundamental problem: what is the best model performance we can expect from a dataset with noisy response values (**Paper IV**). A theoretical upper bound of $\langle R^2 \rangle_{LG}$ was derived for regression models. This adds an additional step of estimating this upper bound before the development of regression models. Knowing $\langle R^2 \rangle_{LG}$ would give us a hint about if it is worthy to develop a regression model for such a noisy dataset, or if the current model already hits this upper bound and thereby further improvement is not possible. In addition to the noise in labels, noisy features are also very common. In this thesis, the effect of noise in features was shown to have a great impact on the performance of regression models by Monte Carlo simulations. However, the theoretical analysis is still missing, and future work is required.

In addition to the noisy samples, ML is also facing the limited training samples in systems biology. This is not an issue for simple biological systems where most relevant features can be easily engineered, such as modelling a metabolic pathway with only five enzymes and using the resulting model to guide the optimization of enzyme expression levels to maximize the production of the end product (Zhou et al., 2018). In this application, around 24 samples were found sufficient to train a predictive model. However, for most other complex biological systems, a large number of training samples are required for ML approaches. For such systems, engineering sufficient relevant features based on domain-knowledge is very challenging. In this thesis, I demonstrated with the prediction of enzyme T_{opt} and protein T_{m} that deep transfer learning approaches are very promising to resolve challenges of both small dataset and insufficient features (**Paper V**). DNN can directly take the raw samples as input without manual feature engineering, as it can extract features itself during the training. Training a DNN from scratch requires much more samples than classical models. However, if we take a DNN that is pre-trained for a certain task and retrain it for a different but related task, only a small number of samples is required for the second dataset. In this thesis, I pre-trained a DNN on a dataset with around 3 million enzymes labeled with OGT values and then retrained it on T_{opt} and T_{m} datasets with only a few thousand samples. This transfer learning approach achieved the state-of-the-art performance for those two tasks.

Deep learning, as the most advanced ML approach, holds great application potential in systems biology due to its ability in modelling very complex tasks. However, the characteristics of low-interpretability and big data-requirements prevent its wide application in systems biology. In addition to the transfer learning approach that can relax the requirement of the large number of training samples, designing visible neural networks (VNN) seems to be a promising solution to the model interpretation (Yu et al., 2018). The architecture of a VNN is designed based on the biological networks. If we take modelling the genotype-phenotype relationships as an example, the traditional ML approach as presented in **Paper I** relies on a black-box model that maps the variations in genomes directly to the phenotypes. Ma J. *et al.* instead proposed a VNN (Ma et al., 2018), where the architecture was designed based on the cell's hierarchy of subsystems. The model was then trained on several million genotypes and the interpretation of the resulting model can be used to investigate the molecular mechanisms underlying the genotype-phenotype relationships.

This demonstrated the power of VNNs in systems biology. Such models can only be invented by biologists based on biological knowledge. The application of VNN in systems biology is still a developing field and requires more efforts in the future.

As showcased in both this thesis and other publications, data-driven ML approaches have a broad application in systems biology at this stage of model development. Thereby, well-structured datasets would greatly benefit the application of ML in systems biology. Although there have been standard databases like BRENDA that collect data from published literature, lots of data are still hidden in the publications. In the future, reporting experimental data in a well-structured and reusable way is becoming more and more important.

In the end, I believe that in the future ML will become a standard tool in systems biology as it naturally complements existing approaches in systems biology. It will be a great assistant for systems biologists on the way to seeking the answers for "***What is Life?***" with mathematical modelling approaches.

Acknowledgments

I would like to thank my supervisor Jens for accepting me as a part of SysBio. Thank you for your supervision and giving me the freedom to explore the research topics that I am interested in. You have always been very supportive and patient. Your comments and suggestions have always been very constructive and critical. Thank Jens and my co-supervisor Verena for letting me join the PAcMEN project, which provided many training programs and opportunities to improve myself. Thank my co-supervisor Martin for all the discussions and contributions to my project. Thank you for your patience and time for the numerous talks we had. Thank my co-supervisor Ibrahim for your help and discussions we had.

I had the honor to involve many great collaborators in my projects. Thank Jan, Boyang and Hao Wang for your valuable contributions and constructive suggestions to my projects. You have been very helpful and patient about my endless questions about both my projects. Thank Jan and Hao Wang for polishing my thesis. Thank Jun, Sandra, Aleksej, Johan Larsbrink and Kersten S. Rabe for your time and expert input to my project. Thank Sandra for your valuable contributions to my project. Thank Yating for joining my project and performing the experiments. I was so impressed by your responsibility and efficiency. Thank my collaborators Jianwen, Lizhan from George Chen's group at Tsinghua University, and William and Matthew from Nigel Scrutton's group at University of Manchester.

I also had the honor to join others' projects. Thank Quanli, Yi, Yun, Xiaowei, Jun and Hongzhong for allowing me to join your amazing projects. I have learnt a lot from you, and I am looking forward to the future collaborations.

I have been surrounded by so many great people in the last years. Thank Avlant, Benjamin and Tyler for your valuable suggestions to my projects. I have enjoyed so much about the talks we had. Thanks to my PAcMEN buddies: Christoph, Soren, Wasti, Vasil, Helen, Jonathan, Roy, Anna, Thomas, Khalil, Paul, Mu-En, Pierre, Homa, Leslie and Jose. I really enjoyed the great time we had together. Thanks to Xin Chen, Yating, Jicheng, Boyang, Lei, Xiaowei, Yanyan, Hao Luo, Yu, Quanli, Yi, Yong, Yongjin, Gaowa, Xin Wen, Xin Liu and Rasool for your company and afterwork time we had during these years. Thanks to Carl for your company and many interesting chats we had. Thanks to Haitao, Yating, Jicheng, Yu and Hao Luo for the game nights we had, which helped me get through this stressful period of thesis writing and job interviews. Thanks to the administrators, Anne-Lise, Anna, Erica, Martina, Shaq, Gunilla, Elin and Thomas for your professional assistance and support.

Last of all, my biggest thanks go to my family. Mom and Dad for always being very supportive about my choice. Brother and sister, for taking care of Mom and Dad during the years when I was away. My wife Haitao, for always being with me and cheering me up. I love you all.

Dad, may your soul rest in peace.

References

Abdi, H., and Valentin, D. (2007). Multiple correspondence analysis. Encyclopedia of Measurement and Statistics *2*, 651–657.

Alber, M., Buganza Tepole, A., Cannon, W.R., De, S., Dura-Bernal, S., Garikipati, K., Karniadakis, G., Lytton, W.W., Perdikaris, P., Petzold, L., et al. (2019). Integrating machine learning and multiscale modeling-perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. NPJ Digit Med *2*, 115.

Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G.M. (2019). Unified rational protein engineering with sequence-based deep representation learning. Nat. Methods.

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics *26*, 1340–1347.

Ammad-ud-din, M., Khan, S.A., Wennerberg, K., and Aittokallio, T. (2017). Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression. Bioinformatics *33*, i359–i368.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.

Bae, S., Choi, S., Kim, S.M., and Park, T. (2016). Prediction of Quantitative Traits Using Common Genetic Variants: Application to Body Mass Index. Genomics Inform. *14*, 149–159.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Beam, A.L., and Kohane, I.S. (2018). Big Data and Machine Learning in Health Care. JAMA *319*, 1317–1318.

Becquet, C., Blachon, S., Jeudy, B., Boulicaut, J.-F., and Gandrillon, O. (2002). Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. Genome Biol. *3*, RESEARCH0067.

Beer, D.G., Kardia, S.L.R., Huang, C.-C., Giordano, T.J., Levin, A.M., Misek, D.E., Lin, L., Chen, G., Gharib, T.G., Thomas, D.G., et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat. Med. *8*, 816–824.

Benevenuta, S., and Fariselli, P. (2019). On the Upper Bounds of the Real-Valued Predictions. Bioinform. Biol. Insights *13*, 1177932219871263.

Berthold, M.R., Feelders, A., and Krempl, G. (2020). Advances in Intelligent Data Analysis XVIII: 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27–29, 2020, Proceedings (Springer International Publishing).

Bienert, S., Waterhouse, A., de Beer, T.A.P., Tauriello, G., Studer, G., Bordoli, L., and Schwede, T. (2017). The SWISS-MODEL Repository—new features and functionality. Nucleic Acids Research *45*, D313–D319.

Bocicor, M., Czibula, G., and Czibula, I. (2011). A Reinforcement Learning Approach for Solving the Fragment Assembly Problem. In 2011 13th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, pp. 191–198.

Botstein, D., and Fink, G.R. (2011). Yeast: an experimental organism for 21st Century biology. Genetics *189*, 695–704.

Breiman, L. (1997). Arcing the edge (Technical Report 486, Statistics Department, University of California at …).

Breiman, L. (2001). Random Forests. Mach. Learn. *45*, 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and regression trees. Monterey, Calif., USA: Wadsworth.

Breitling, R. (2010). What is systems biology? Front. Physiol. *1*, 9.

Bridge, L.J., Mead, J., Frattini, E., Winfield, I., and Ladds, G. (2018). Modelling and simulation of biased agonism dynamics at a G protein-coupled receptor. J. Theor. Biol. *442*, 44–65.

Bruggeman, F.J., and Teusink, B. (2018). Living with noise: On the propagation of noise from molecules to phenotype and fitness. Current Opinion in Systems Biology *8*, 144–150.

Buchanan, C.L., Connaris, H., Danson, M.J., Reeve, C.D., and Hough, D.W. (1999). An extremely thermostable aldolase from Sulfolobus solfataricus with specificity for non-phosphorylated substrates. Biochem. J *343 Pt 3*, 563–570.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Cao, D.-S., Xu, Q.-S., and Liang, Y.-Z. (2013). propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics *29*, 960–962.

Capriotti, E., Fariselli, P., and Casadio, R. (2005a). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Research *33*, W306–W310.

Capriotti, E., Fariselli, P., and Casadio, R. (2005b). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. *33*, W306–W310.

Carbonell, P., Fichera, D., Pandit, S.B., and Faulon, J.-L. (2012). Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. BMC Syst. Biol. *6*, 10.

Carmona-Saez, P., Chagoyen, M., Rodriguez, A., Trelles, O., Carazo, J.M., and Pascual-Montano, A. (2006). Integrated analysis of gene expression by Association Rules Discovery. BMC Bioinformatics *7*, 54.

Carrera, J., and Covert, M.W. (2015). Why Build Whole-Cell Models? Trends Cell Biol. *25*, 719–722.

Caspeta, L., and Nielsen, J. (2015). Thermotolerant Yeast Strains Adapted by Laboratory Evolution Show Trade-Off at Ancestral Temperatures and Preadaptation to Other Stresses. MBio *6*, e00431.

Cawley, G.C., and Talbot, N.L.C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. *11*, 2079–2107.

Chang, R.L., Andrews, K., Kim, D., Li, Z., Godzik, A., and Palsson, B.O. (2013). Structural systems biology evaluation of metabolic thermotolerance in Escherichia coli. Science *340*, 1220–1223.

Chen, K., Gao, Y., Mih, N., O'Brien, E.J., Yang, L., and Palsson, B.O. (2017). Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. Proc. Natl. Acad. Sci. U. S. A. *114*, 11548–11553.

Chen, Y., Li, G., and Nielsen, J. (2019). Genome-Scale Metabolic Modeling from Yeast to Human Cell Models of Complex Diseases: Latest Advances and Challenges. In Yeast Systems Biology: Methods and Protocols, S.G. Oliver, and J.I. Castrillo, eds. (New York, NY: Springer New York), pp. 329–345.

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T.T., Wang, Y., Webb, G.I., Smith, A.I., Daly, R.J., Chou, K.-C., et al. (2018). iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics *34*, 2499–2502.

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T.T., Leier, A., Revote, J., Zhu, Y., Powell, D.R., Akutsu, T., Webb, G.I., et al. (2020). iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief. Bioinform. *21*, 1047–1057.

Chiu, Y.-C., Chen, H.-I.H., Zhang, T., Zhang, S., Gorthi, A., Wang, L.-J., Huang, Y., and Chen, Y. (2019). Correction to: Predicting drug response of tumors from integrated genomic profiles by deep neural networks. BMC Med. Genomics *12*, 119.

Cios, K.J., Swiniarski, R.W., Pedrycz, W., and Kurgan, L.A. (2007). Unsupervised Learning: Association Rules. In Data Mining: A Knowledge Discovery Approach, K.J. Cios, R.W. Swiniarski, W. Pedrycz, and L.A. Kurgan, eds. (Boston, MA: Springer US), pp. 289–306.

Claesen, M., and De Moor, B. (2015). Hyperparameter Search in Machine Learning.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. Mach. Learn. *20*, 273–297.

Crawford, J., and Greene, C.S. (2020). Incorporating biological structure into machine learning models in biomedicine. Curr. Opin. Biotechnol. *63*, 126–134.

Crossa, J., Campos, G. de L., Pérez, P., Gianola, D., Burgueño, J., Araus, J.L., Makumbi, D., Singh, R.P., Dreisigacker, S., Yan, J., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics *186*, 713–724.

Daniel, R.M., and Danson, M.J. (2010). A new understanding of how temperature affects the catalytic activity of enzymes. Trends Biochem. Sci. *35*, 584–591.

Demirjian, D.C., Morís-Varas, F., and Cassidy, C.S. (2001). Enzymes from extremophiles. Curr. Opin. Chem. Biol. *5*, 144–151.

Dill, K.A., Ghosh, K., and Schmit, J.D. (2011). Physical limits of cells and proteomes. Proc. Natl. Acad. Sci. U. S. A. *108*, 17876–17882.

Doshi-Velez, F., and Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning.

Driessen, R.P.C., Sitters, G., Laurens, N., Moolenaar, G.F., Wuite, G.J.L., Goosen, N., and Dame, R.T. (2014). Effect

of temperature on the intrinsic flexibility of DNA and its interaction with architectural proteins. Biochemistry *53*, 6430–6438.

Engqvist, M.K.M. (2018). Correlating enzyme annotations with a large set of microbial growth temperatures reveals metabolic adaptations to growth at diverse temperatures. BMC Microbiol. *18*, 177.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. Nat. Med. *25*, 24–29.

Feist, A.M., and Palsson, B.O. (2010). The biomass objective function. Curr. Opin. Microbiol. *13*, 344–349.

Förster, J., Famili, I., Fu, P., Palsson, B.Ø., and Nielsen, J. (2003). Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res. *13*, 244–253.

Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B., Merlevede, A., Roncoroni, M., Voordeckers, K., Miraglia, L., et al. (2016). Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. Cell *166*, 1397–1410.e16.

Gazestani, V.H., and Lewis, N.E. (2019). From Genotype to Phenotype: Augmenting Deep Learning with Networks and Systems Biology. Curr Opin Syst Biol *15*, 68–73.

Girolami, M. (2008). Bayesian inference for differential equations. Theor. Comput. Sci. *408*, 4–16.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science *274*, 546, 563–567.

Goldstein, M., and Wooff, D. (2007). Bayes Linear Statistics: Theory and Methods (John Wiley & Sons).

Green, A.G., Swithers, K.S., Gogarten, J.F., and Gogarten, J.P. (2013). Reconstruction of ancestral 16S rRNA reveals mutation bias in the evolution of optimal growth temperature in the Thermotogae phylum. Mol. Biol. Evol. *30*, 2463–2474.

Grimaud, G.M., Mairet, F., Sciandra, A., and Bernard, O. (2017). Modeling the temperature effect on the specific growth rate of phytoplankton: a review. Rev. Environ. Sci. Technol. *16*, 625–645.

Gu, S., Holly, E., Lillicrap, T., and Levine, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 3389–3396.

Guruprasad, K., Reddy, B.V., and Pandit, M.W. (1990). Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng. *4*, 155–161.

Harris, E.F., and Smith, R.N. (2009). Accounting for measurement error: a critical but often overlooked process. Arch. Oral Biol. *54 Suppl 1*, S107–S117.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity Mappings in Deep Residual Networks. In Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, eds. (Cham: Springer International Publishing), pp. 630–645.

Heavner, B.D., and Price, N.D. (2015). Comparative Analysis of Yeast Metabolic Network Models Highlights Progress, Opportunities for Metabolic Reconstruction. PLoS Comput. Biol. *11*, e1004530.

Heckmann, D., Lloyd, C.J., Mih, N., Ha, Y., Zielinski, D.C., Haiman, Z.B., Desouki, A.A., Lercher, M.J., and Palsson, B.O. (2018). Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. Nature Communications *9*.

Henry, C.S., Broadbelt, L.J., and Hatzimanikatis, V. (2007). Thermodynamics-based metabolic flux analysis. Biophys. J. *92*, 1792–1805.

Herrgård, M.J., Lee, B.-S., Portnoy, V., and Palsson, B.Ø. (2006). Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae. Genome Res. *16*, 627–635.

Hickey, D.A., and Singer, G.A.C. (2004). Genomic and proteomic adaptations to growth at high temperature. Genome Biol. *5*, 117.

Hobbs, J.K., Jiao, W., Easter, A.D., Parker, E.J., Schipper, L.A., and Arcus, V.L. (2017). Change in Heat Capacity for Enzyme Catalysis Determines Temperature Dependence of Enzyme Catalyzed Rates. ACS Chem. Biol. *12*, 868.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. (2019). A Benchmark for Interpretability Methods in Deep Neural Networks. In Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, eds. (Curran Associates, Inc.), pp. 9737–9748.

Huang, M., Bao, J., Hallström, B.M., Petranovic, D., and Nielsen, J. (2017). Efficient protein production by yeast requires global tuning of metabolism. Nat. Commun. *8*, 1131.

Ideker, T., Galitski, T., and Hood, L. (2001). A new approach to decoding life: systems biology. Annu. Rev. Genomics Hum. Genet. *2*, 343–372.

Jamshidi, N., and Palsson, B.Ø. (2008). Formulating genome-scale kinetic models in the post-genome era. Mol. Syst. Biol. *4*, 171.

Jelier, R., Semple, J.I., Garcia-Verdugo, R., and Lehner, B. (2011). Predicting phenotypic variation in yeast from individual genome sequences. Nat. Genet. *43*, 1270–1274.

Jeske, L., Placzek, S., Schomburg, I., Chang, A., and Schomburg, D. (2019). BRENDA in 2019: a European ELIXIR core data resource. Nucleic Acids Res. *47*, D542–D549.

Kaelbling, L.P., Littman, M.L., and Moore, A.W. (1996). Reinforcement Learning: A Survey. J. Artif. Intell. Res. *4*, 237–285.

van der Kamp, M.W., Prentice, E.J., Kraakman, K.L., Connolly, M., Mulholland, A.J., and Arcus, V.L. (2018). Dynamical origins of heat capacity changes in enzyme-catalysed reactions. Nat. Commun. *9*, 1177.

Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., and Kumar, V. (2017). Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. IEEE Trans. Knowl. Data Eng. *29*, 2318–2331.

Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., Jr, Assad-Garcia, N., Glass, J.I., and Covert, M.W. (2012). A whole-cell computational model predicts phenotype from genotype. Cell *150*, 389–401.

Kato, S., Itoh, T., Yuki, M., Nagamori, M., Ohnishi, M., Uematsu, K., Suzuki, K., Takashina, T., and Ohkuma, M. (2019). Isolation and characterization of a thermophilic sulfur- and iron-reducing thaumarchaeote from a terrestrial acidic hot spring. ISME J. *13*, 2465–2474.

Kemble, H., Nghe, P., and Tenaillon, O. (2019). Recent insights into the genotype–phenotype relationship from massively parallel genetic assays. Evol. Appl. *12*, 1721–1742.

Khodayari, A., and Maranas, C.D. (2016). A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. Nat. Commun. *7*, 13806.

Khodayari, A., Zomorrodi, A.R., Liao, J.C., and Maranas, C.D. (2014). A kinetic model of Escherichia coli core metabolism satisfying multiple sets of mutant flux data. Metab. Eng. *25*, 50–62.

Kingma, D.P., and Welling, M. (2013). Auto-Encoding Variational Bayes.

Kitano, H. (2002). Systems biology: a brief overview. Science *295*, 1662–1664.

Koch, I., Nöthen, J., and Schleiff, E. (2017). Modeling the Metabolism of Arabidopsis thaliana: Application of Network Decomposition and Network Reduction in the Context of Petri Nets. Front. Genet. *8*, 85.

Kumar, S., and Nussinov, R. (2001). How do thermophilic proteins deal with heat? Cell. Mol. Life Sci. *58*, 1216–1233.

Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132.

Lane, M.M., and Morrissey, J.P. (2010). Kluyveromyces marxianus: A yeast emerging from its sister's shadow. Fungal Biology Reviews *24*, 17–26.

Lanzeni, S., Messina, E., and Archetti, F. (2008). Graph models and mathematical programming in biochemical network analysis and metabolic engineering design. Comput. Math. Appl. *55*, 970–983.

Leuenberger, P., Ganscha, S., Kahraman, A., Cappelletti, V., Boersema, P.J., von Mering, C., Claassen, M., and Picotti, P. (2017). Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. Science *355*.

Lin, C., Jain, S., Kim, H., and Bar-Joseph, Z. (2017). Using neural networks for reducing the dimensions of single-cell RNA-Seq data. Nucleic Acids Res. *45*, e156.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., and Russell, R.B. (2003). Protein disorder prediction: implications for structural proteomics. Structure *11*, 1453–1459.

Liu, J.K., O'Brien, E.J., Lerman, J.A., Zengler, K., Palsson, B.O., and Feist, A.M. (2014). Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale. BMC Syst. Biol. *8*, 110.

Lobry, J.R., and Gautier, C. (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Res. *22*, 3174–3180.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. *42*, D490–D495.

Lu, H., Li, F., Sánchez, B.J., Zhu, Z., Li, G., Domenzain, I., Marcišauskas, S., Anton, P.M., Lappa, D., Lieven, C., et al. (2019). A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nat. Commun. *10*, 3586.

Lundberg, S.M., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural

Information Processing Systems 30, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. (Curran Associates, Inc.), pp. 4765–4774.

Luxton, D.D. (2016). Chapter 1 - An Introduction to Artificial Intelligence in Behavioral and Mental Health Care. In Artificial Intelligence in Behavioral and Mental Health Care, D.D. Luxton, ed. (San Diego: Academic Press), pp. 1–26.

Ma, J., Yu, M.K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., and Ideker, T. (2018). Using deep learning to model the hierarchical structure and function of a cell. Nat. Methods *15*, 290.

Magnan, C.N., and Baldi, P. (2014). SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics *30*, 2592–2597.

Mahmud, M., Kaiser, M.S., Hussain, A., and Vassanelli, S. (2018). Applications of Deep Learning and Reinforcement Learning to Biological Data. IEEE Trans Neural Netw Learn Syst *29*, 2063–2079.

Makowsky, R., Pajewski, N.M., Klimentidis, Y.C., Vazquez, A.I., Duarte, C.W., Allison, D.B., and de los Campos, G. (2011). Beyond missing heritability: prediction of complex traits. PLoS Genet. *7*, e1002051.

Märtens, K., Hallin, J., Warringer, J., Liti, G., and Parts, L. (2016). Predicting quantitative traits from genome and phenome with near perfect accuracy. Nat. Commun. *7*, 11512.

Meng, C., Zeleznik, O.A., Thallinger, G.G., Kuster, B., Gholami, A.M., and Culhane, A.C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. Brief. Bioinform. *17*, 628–641.

Miskovic, L., Béal, J., Moret, M., and Hatzimanikatis, V. (2019). Uncertainty reduction in biochemical kinetic models: Enforcing desired model properties. PLoS Comput. Biol. *15*, e1007242.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). Foundations of machine learning (MIT press).

Molnar, C. (2020). Interpretable Machine Learning (Lulu.com).

Montáns, F.J., Chinesta, F., Gómez-Bombarelli, R., and Kutz, J.N. (2019). Data-driven modeling and learning in science and engineering. Comptes Rendus Mécanique *347*, 845–855.

Montanucci, L., Martelli, P.L., Ben-Tal, N., and Fariselli, P. (2019). A natural upper bound to the accuracy of predicting protein stability changes upon mutations. Bioinformatics *35*, 1513–1517.

Morota, G., Abdollahi-Arpanahi, R., Kranis, A., and Gianola, D. (2014). Genome-enabled prediction of quantitative traits in chickens using genomic annotation. BMC Genomics *15*, 109.

Murphy, K.P., and Gill, S.J. (1991). Solid model compounds and the thermodynamics of protein unfolding. J. Mol. Biol. *222*, 699–709.

Nakashima, H., Fukuchi, S., and Nishikawa, K. (2003). Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. J. Biochem. *133*, 507–513.

Neidleman, S.L. (1987). Effects of temperature on lipid unsaturation. Biotechnol. Genet. Eng. Rev. *5*, 245–268.

Nguyen, V., Wilson, C., Hoemberger, M., Stiller, J.B., Agafonov, R.V., Kutter, S., English, J., Theobald, D.L., and Kern, D. (2017). Evolutionary drivers of thermoadaptation in enzyme catalysis. Science *355*, 289–294.

Nielsen, J. (2015). BIOENGINEERING. Yeast cell factories on the horizon. Science *349*, 1050–1051.

Nielsen, J. (2017). Systems Biology of Metabolism. Annu. Rev. Biochem. *86*, 245–275.

Nielsen, J., and Jewett, M.C. (2008). Impact of systems biology on metabolic engineering of Saccharomyces cerevisiae. FEMS Yeast Res. *8*, 122–131.

O'Brien, E.J., Lerman, J.A., Chang, R.L., Hyduke, D.R., and Palsson, B.Ø. (2013). Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Mol. Syst. Biol. *9*, 693.

Oh, K.-S., and Jung, K. (2004). GPU implementation of neural networks. Pattern Recognit. *37*, 1311–1314.

Oobatake, M., and Ooi, T. (1993). Hydration and heat stability effects on protein unfolding. Prog. Biophys. Mol. Biol. *59*, 237–284.

Orth, J.D., Thiele, I., and Palsson, B.Ø. (2010). What is flux balance analysis? Nat. Biotechnol. *28*, 245–248.

Osterlund, T., Nookaew, I., and Nielsen, J. (2012). Fifteen years of large scale metabolic modeling of yeast: developments and impacts. Biotechnol. Adv. *30*, 979–988.

Patil, K.R., and Nielsen, J. (2005). Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc. Natl. Acad. Sci. U. S. A. *102*, 2685–2689.

Pearson, T.A., and Manolio, T.A. (2008). How to interpret a genome-wide association study. JAMA *299*, 1335–1344.

Peng, G.C.Y., Alber, M., Buganza Tepole, A., Cannon, W.R., De, S., Dura-Bernal, S., Garikipati, K., Karniadakis, G., Lytton, W.W., Perdikaris, P., et al. (2020). Multiscale Modeling Meets Machine Learning: What Can We Learn? Arch. Comput. Methods Eng.

Peter, I.S., Faure, E., and Davidson, E.H. (2012). Predictive computation of genomic logic processing functions in

embryonic development. Proc. Natl. Acad. Sci. U. S. A. *109*, 16434–16442.

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., et al. (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature *556*, 339–344.

Peter Sollich, A.K. (1996). Learning with ensembles: How over-fitting can be useful.

PIATETSKY-SHAPIRO, and G (1991). Discovery, Analysis, and Presentation of Strong Rules. Knowledge Discovery in Databases 229–238.

Postmus, J., Canelas, A.B., Bouwman, J., Bakker, B.M., van Gulik, W., de Mattos, M.J.T., Brul, S., and Smits, G.J. (2008). Quantitative analysis of the high temperature-induced glycolytic flux increase in Saccharomyces cerevisiae reveals dominant metabolic regulation. J. Biol. Chem. *283*, 23524–23532.

Raina, R., Madhavan, A., and Ng, A.Y. (2009). Large-scale deep unsupervised learning using graphics processors. In Proceedings of the 26th Annual International Conference on Machine Learning, (New York, NY, USA: Association for Computing Machinery), pp. 873–880.

Raj, A., and van Oudenaarden, A. (2009). Single-molecule approaches to stochastic gene expression. Annu. Rev. Biophys. *38*, 255–270.

Ralha, C.G., Schneider, H.W., Walter, M.E.M.T., and Bazzan, A.L. (2010). Reinforcement Learning Method for BioAgents. In 2010 Eleventh Brazilian Symposium on Neural Networks, pp. 109–114.

Rappé, M.S., and Giovannoni, S.J. (2003). The uncultured microbial majority. Annu. Rev. Microbiol. *57*, 369–394.

Reddy, V.N., Liebman, M.N., and Mavrovouniotis, M.L. (1996). Qualitative analysis of biochemical reaction systems. Comput. Biol. Med. *26*, 9–24.

Rhee, S.Y., Wood, V., Dolinski, K., and Draghici, S. (2008). Use and misuse of the gene ontology annotations. Nat. Rev. Genet. *9*, 509–515.

Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Robertson, A.D., and Murphy, K.P. (1997). Protein Structure and the Energetics of Protein Stability. Chem. Rev. *97*, 1251–1268.

Romero, P.A., Krause, A., and Arnold, F.H. (2013). Navigating the protein fitness landscape with Gaussian processes. Proc. Natl. Acad. Sci. U. S. A. *110*, E193–E201.

Ryu, J.Y., Kim, H.U., and Lee, S.Y. (2019). Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc. Natl. Acad. Sci. U. S. A. *116*, 13996–14001.

Saa, P.A., and Nielsen, L.K. (2017). Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks. Biotechnol. Adv. *35*, 981–1003.

Samuel, A.L. (1959). Some Studies in Machine Learning Using the Game of Checkers. IBM J. Res. Dev. *3*, 210–229.

Sánchez, B.J., Zhang, C., Nilsson, A., Lahtvee, P.-J., Kerkhoven, E.J., and Nielsen, J. (2017). Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Mol. Syst. Biol. *13*, 935.

Santosa, F., and Symes, W.W. (1986). Linear Inversion of Band-Limited Reflection Seismograms. SIAM J. Sci. and Stat. Comput. *7*, 1307–1330.

Sawle, L., and Ghosh, K. (2011). How do thermophilic proteins and proteomes withstand high temperature? Biophys. J. *101*, 217–227.

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. Neural Netw. *61*, 85–117.

Schrodinger, E. (1944). What is life? The Physical Aspect of the Living Cell (Cambridge: Cambridge University Press).

Schuster, S.C. (2008). Next-generation sequencing transforms today's biology. Nat. Methods *5*, 16–18.

Seeger, M. (2004). Gaussian processes for machine learning. Int. J. Neural Syst. *14*, 69–106.

Shahzad, K., and Loor, J.J. (2012). Application of Top-Down and Bottom-up Systems Approaches in Ruminant Physiology and Metabolism. Curr. Genomics *13*, 379–394.

Shaked, I., Oberhardt, M.A., Atias, N., Sharan, R., and Ruppin, E. (2016). Metabolic Network Prediction of Drug Side Effects. Cell Syst *2*, 209–213.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. (2016). Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving.

Shaw, W.M., Yamauchi, H., Mead, J., Gowers, G.-O.F., Bell, D.J., Öling, D., Larsson, N., Wigglesworth, M., Ladds, G., and Ellis, T. (2019). Engineering a Model Cell for Rational Tuning of GPCR Signaling. Cell *177*, 782–796.e27.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of Go without human knowledge. Nature *550*, 354–359.

Singh, J., Hanson, J., Paliwal, K., and Zhou, Y. (2019). RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nat. Commun. *10*, 5407.

Slivka, D.R., Dumke, C.L., Tucker, T.J., Cuddy, J.S., and Ruby, B. (2012). Human mRNA response to exercise and temperature. Int. J. Sports Med. *33*, 94–100.

Smallbone, K., Simeonidis, E., Swainston, N., and Mendes, P. (2010). Towards a genome-scale kinetic model of cellular metabolism. BMC Syst. Biol. *4*, 6.

Söhngen, C., Podstawka, A., Bunk, B., Gleim, D., Vetcininova, A., Reimer, L.C., Ebeling, C., Pendarovski, C., and Overmann, J. (2016). BacDive – The Bacterial Diversity Metadatabase in 2016. Nucleic Acids Research *44*, D581–D585.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. J. R. Stat. Soc. Series B Stat. Methodol. *36*, 111–147.

Stumpf, P.S., Du, D., Imanishi, H., Kunisaki, Y., Semba, Y., Noble, T., Smith, R.C.G., Rose-Zerili, M., West, J.J., Oreffo, R.O.C., et al. (2019). Mapping biology from mouse to man using transfer learning (bioRxiv).

Sunnåker, M., Busetto, A.G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. (2013). Approximate Bayesian computation. PLoS Comput. Biol. *9*, e1002803.

Tan, M. (2016). Prediction of anti-cancer drug response by kernelized multi-task learning. Artificial Intelligence in Medicine *73*, 70–77.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2018). A Survey on Deep Transfer Learning. In Artificial Neural Networks and Machine Learning – ICANN 2018, (Springer International Publishing), pp. 270–279.

Tang, B., Pan, Z., Yin, K., and Khateeb, A. (2019). Recent Advances of Deep Learning in Bioinformatics and Computational Biology. Front. Genet. *10*, 214.

Tarca, A.L., Carey, V.J., Chen, X.-W., Romero, R., and Drăghici, S. (2007). Machine learning and its applications to biology. PLoS Comput. Biol. *3*, e116.

Tettelin, H., and Masignani, V. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome." Proceedings of the.

Tin Kam Ho (1995). Random decision forests. In Proceedings of 3rd International Conference on Document Analysis and Recognition, pp. 278–282 vol.1.

Tomita, M. (2001). Whole-cell simulation: a grand challenge of the 21st century. Trends Biotechnol. *19*, 205–210.

Tsimring, L.S. (2014). Noise in biology. Rep. Prog. Phys. *77*, 026601.

Vidal, M. (2009). A unifying view of 21st century systems biology. FEBS Lett. *583*, 3891–3894.

Vieille, C., and Zeikus, G.J. (2001). Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol. Mol. Biol. Rev. *65*, 1–43.

Villadsen, J., Nielsen, J., and Lidén, G. (2011). Bioreaction Engineering Principles (Springer Science & Business Media).

Wang, Z., Wang, D., Li, C., Xu, Y., Li, H., and Bao, Z. (2018). Deep reinforcement learning of cell movement in the early stage of C.elegans embryogenesis. Bioinformatics *34*, 3169–3177.

Weiss, K., Khoshgoftaar, T.M., and Wang, D. (2016). A survey of transfer learning. Journal of Big Data *3*, 9.

Wunderlich, Z., and Mirny, L.A. (2006). Using the topology of metabolic networks to predict viability of mutant strains. Biophys. J. *91*, 2304–2311.

Yang, J.H., Wright, S.N., Hamblin, M., McCloskey, D., Alcantar, M.A., Schrübbers, L., Lopatkin, A.J., Satish, S., Nili, A., Palsson, B.O., et al. (2019). A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. Cell *177*, 1649–1661.e9.

Yau, C., and Campbell, K. (2019). Bayesian statistical learning for big data biology. Biophys. Rev. *11*, 95–102.

Yu, M.K., Ma, J., Fisher, J., Kreisberg, J.F., Raphael, B.J., and Ideker, T. (2018). Visible Machine Learning for Biomedicine. Cell *173*, 1562–1565.

Zakhartsev, M., Yang, X., Reuss, M., and Pörtner, H.O. (2015). Metabolic efficiency in yeast Saccharomyces cerevisiae in relation to temperature dependent growth and biomass yield. J. Therm. Biol. *52*, 117–129.

Zampieri, G., Vijayakumar, S., Yaneske, E., and Angione, C. (2019). Machine and deep learning meet genome-scale metabolic modeling. PLoS Comput. Biol. *15*, e1007084.

Zeldovich, K.B., Berezovsky, I.N., and Shakhnovich, E.I. (2007). Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput. Biol. *3*, e5.

Zeng, P., Zhao, Y., Qian, C., Zhang, L., Zhang, R., Gou, J., Liu, J., Liu, L., and Chen, F. (2015). Statistical analysis for genome-wide association study. J. Biomed. Res. *29*, 285–297.

Zhang, J., Petersen, S.D., Radivojevic, T., Ramirez, A., Pérez, A., Abeliuk, E., Sánchez, B.J., Costello, Z., Chen, Y., Fero, M., et al. Predictive engineering and optimization of tryptophan metabolism in yeast through a combination of mechanistic and machine learning models.

Zheng, H., Yuan, J., and Chen, L. (2017). Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation. Energies *10*, 1168.

Zhou, Y., Li, G., Dong, J., Xing, X.-H., Dai, J., and Zhang, C. (2018). MiYA, an efficient machine-learning workflow in conjunction with the YeastFab assembly strategy for combinatorial optimization of heterologous metabolic pathways in Saccharomyces cerevisiae. Metab. Eng. *47*, 294–302.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B Stat. Methodol.