MAURÍCIO ALEXANDER DE MOURA FERREIRA

MACHINE LEARNING PREDICTION OF PROTEIN ABUNDANCE BY CODON USAGE METRICS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Microbiologia Agrícola, para obtenção do título de *Magister Scientiae*.

Orientador: Wendel Batista da Silveira
Coorientadora: Sabrina de Azevedo Silveira

VIÇOSA – MINAS GERAIS

2020

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
F383m
2020

Ferreira, Maurício Alexander de Moura, 1995-
Machine learning prediction of protein abundance by codon usage metrics / Maurício Alexander de Moura Ferreira. – Viçosa, MG, 2020.
70 f. : il. (algumas color.) ; 29 cm.

Inclui apêndice.
Orientador: Wendel Batista da Silveira.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Proteínas. 2. Códons. 3. Modelos matemáticos. 4. Engenharia Metabólica. I. Universidade Federal de Viçosa. Departamento de Microbiologia. Programa de Pós-Graduação em Microbiologia Agrícola. II. Título.

CDD 22. ed. 572.645

MAURÍCIO ALEXANDER DE MOURA FERREIRA

MACHINE LEARNING PREDICTION OF PROTEIN ABUNDANCE BY CODON USAGE METRICS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Microbiologia Agrícola, para obtenção do título de *Magister Scientiae*.

APROVADA: 27 de julho de 2020.

Assentimento:

Maurício Alexander de Moura Ferreira
Autor

Wendel Batista da Silveira
Orientador

Aos meus pais, Gersimar e Norma.

AGRADECIMENTOS

À Universidade Federal de Viçosa, pela excelência em ensino, pesquisa e extensão.

Ao Programa de Pós-Graduação em Microbiologia Agrícola e ao Departamento de Microbiologia, pela oportunidade de realização do curso de mestrado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), pelo auxílio (Código de Financiamento 001).

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela concessão da bolsa.

À Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), pelo auxílio financeiro.

Ao professor Wendel Batista da Silveira, pela orientação.

À professora Sabrina de Azevedo Silveira, pela coorientação.

À doutora Rafaela Zandonade Ventorim e a Eduardo Luís Menezes de Almeida, pela contribuição ao trabalho.

Ao professor Leonardo Lopes Bhering, por gentilmente ceder acesso ao servidor em seu laboratório.

Aos colegas de laboratório, por discussões e comentários.

BIOGRAFIA

Maurício Alexander de Moura Ferreira, natural de Belo Horizonte, é bacharel em Ciências Biológicas pela Universidade Federal do Espírito Santo. Iniciou a graduação em 2014, e ao longo do curso foi bolsista de iniciação científica em projetos de engenharia genética de plantas para controle de patógenos. Se formou em 2018, e neste mesmo ano ingressou no mestrado pelo Programa de Pós-Graduação em Microbiologia Agrícola, da Universidade Federal de Viçosa. Durante o mestrado, foi membro do Laboratório de Fisiologia de Microrganismos, sob orientação do professor Wendel Batista da Silveira, onde trabalhou com biologia sistêmica e computacional de leveduras, utilizando modelos metabólicos em escala genômica e algoritmos de *machine learning* visando aplicações em engenharia metabólica e bioprocessos.

“Qual é o segredo da vida? – perguntei.

– Esqueci – disse Sandra.

– Proteína – declarou o barman. – Descobriram alguma coisa sobre proteína.

– Sim – disse Sandra. – Isso mesmo...”

(excerto de “Cama de Gato”, obra de Kurt Vonnegut).

ABSTRACT

FERREIRA, Maurício Alexander de Moura, M.Sc., Universidade Federal de Viçosa, July, 2020. **Machine Learning Prediction of Protein Abundance by Codon Usage Metrics**. Adviser: Wendel Batista da Silveira. Co-adviser: Sabrina de Azevedo Silveira.

Proteins are responsible for most physiological processes in the cell, and their abundance provides crucial information for systems biology research. Protein abundance is determined by a number of factors, such as mRNA abundance, translation efficiency, protein turnover rate, and codon usage bias. New frameworks of genome-scale metabolic models have been recently developed to simulate phenotypes taking into account protein abundance data along with enzyme kinetics. However, these models still have the limitation of dataset availability, which impairs their reconstruction. This is due to limitations in absolute protein quantification methods, such as mass spectrometry. Moreover, absolute protein quantification has been mostly limited to model species, such as *Escherichia coli* and *Saccharomyces cerevisiae*, which hinders system biology endeavours in non-model species. Codon usage bias directly affects translation dynamics, which in turn affects protein levels, and many metrics for codon usage have been developed in order to clarify this phenomenon. In this study, it was evaluated the effect of codon usage bias of genes in protein abundance. Notably, many differences regarding codon usage patterns between genes coding for highly abundant proteins and genes coding for less abundant proteins were observed. Based on these differences, various codon metrics coupled with machine learning algorithms were applied to predict the absolute abundance of proteins used by *S. cerevisiae*. The machine learning models predicted protein abundances from codon usage metrics with remarkable accuracy. Upon integration of the predicted protein abundance in enzyme-constrained genome-scale metabolic models, the simulated phenotypes closely matched experimental data, which demonstrates that the built predictive models are valuable tools for systems metabolic engineering approaches

Keywords: Codon usage bias. Metabolic modelling. Metabolic engineering.

RESUMO

FERREIRA, Maurício Alexander de Moura, M.Sc., Universidade Federal de Viçosa, julho de 2020. **Predição da abundância de proteínas por métricas de uso de códons utilizando aprendizado de máquina**. Orientador: Wendel Batista da Silveira. Coorientadora: Sabrina de Azevedo Silveira.

Proteínas são as principais moléculas responsáveis por processos fisiológicos na célula, e o conhecimento de suas abundâncias é crucial para pesquisas em biologia sistêmica do metabolismo. A abundância de proteínas é determinada por uma série de fatores, como abundância de mRNA, eficiência da tradução, taxas de *turnover* de proteínas e viés no uso de códons. Recentemente, novas plataformas para simulação de fenótipos têm sido desenvolvidas, integrando dados de eficiência catalítica e abundância de proteína a modelos metabólicos em escala genômica. Entretanto, o uso destes modelos é dificultado pela indisponibilidade de dados de abundância de muitas proteínas, principalmente por limitações analíticas como espectrometria de massas. Além disso, muitos dos esforços em proteômica quantitativa têm sido focados em espécies-modelo, como *Saccharomyces cerevisiae* e *Escherichia coli*, o que limita abordagens sistêmicas em espécies não modelos. O viés no uso de códons é um fenômeno que afeta diretamente a dinâmica da tradução, o que impacta na quantidade de proteína presente na célula. Dessa forma, muitas métricas foram desenvolvidas para explicar matematicamente esse fenômeno. Neste trabalho, foram exploradas as diferenças no uso de códons entre sequências codificantes de proteínas de *S. cerevisiae* de alta e baixa abundância. Estes dados foram utilizados para o treinamento de algoritmos de aprendizado de máquina, com o intuito de gerar modelos capazes de predizer a abundância de proteínas. As abundâncias preditas foram então integradas em modelos metabólicos em escala genômica, e os fenótipos simulados apresentaram boa correspondência com valores experimentais. Isso demonstra que estes modelos preditivos são ferramentas valiosas para a biologia sistêmica do metabolismo e para a engenharia metabólica.

Palavras-chave: Uso de códons. Modelagem metabólica. Engenharia metabólica.

SUMÁRIO

INTRODUCTION

Proteins are responsible for many physiological processes and their abundance is a crucial information for systems biology research. The total amount of protein in a cell is referred to as the protein pool, and its efficient allocation controls metabolic flux. Protein abundance relies on factors such as mRNA abundance, translation efficiency, protein turnover rate, and codon usage bias. Taking into account the impact of enzyme amounts on metabolic flux, protein abundances along with enzyme kinetics have recently integrated into genome-scale metabolic models (GEMs), which are an important tool for understanding the physiological processes at the systems level. Moreover, they contribute to large-scale engineering of the metabolism. GEMs can predict metabolic flux through a gene-protein-reaction association of all metabolic pathways in an organism. Thus, it is used for simulating the growth of an organism, for predicting the titre of metabolites, for predicting the metabolic impact of insertion and deletion of genes and for simulating the addition of entire biosynthetic pathways for the production of a desired biotechnological product.

GEMs are reconstructed using metabolic pathway databases and genome information. However, the predictive capacity of conventional GEMs is limited as they do not take into account either enzyme kinetics or enzyme abundance. Based on this, new frameworks have been recently developed to overcome the drawbacks related to the conventional GEMs. Importantly, models of metabolism and macromolecular expression (ME models) and GEMs with enzymatic constraints and omics data (GECKO models) have displayed a better capacity to predict metabolic flux and resource allocation. Nevertheless, these models still have the limitation of dataset availability, which impairs the reconstruction of ME-models and GECKO models. Moreover, absolute protein abundance quantification, as determined by mass spectrometry, still has limitations in capturing the entire protein pool of an organism.

It is known that codon usage bias plays an important role in determining protein abundance. Considering the degeneracy of the genetic code and the difference in abundance between tRNA molecules, a specific choice of codons is able to tune the rate in which a certain protein is produced. Metrics for codon usage have been developed in order to clarify this phenomenon, such as the codon adaptation index (CAI) and tRNA adaptation index (tAI). It is important to point out that the majority of

the metrics have often been well correlated to experimentally detected RNA and protein levels.

Machine learning has been a useful tool for systems biology which has been used to predict many quantitative properties of macromolecules. Regression analysis is often employed in the form of supervised machine learning algorithms, which are trained to predict a numerical value. Thus, machine learning can be used to predict protein abundance from codon usage metrics. The model organism *Saccharomyces cerevisiae* is an ideal starting point, due to its large availability of proteomics and physiological data. The predicted abundances are expected to be incorporated in enzyme-constrained GEMs, in order to test the predictions and undertake phenotype simulations.

This dissertation is organized in two chapters. In the first chapter, a literature review presents the main concepts addressed in this work, such as protein abundance, systems biology of metabolism and codon usage bias. In the second chapter, organized as a research manuscript, it is reported the usage of machine learning to predict protein abundance from codon usage metrics. The machine learning models predicted protein abundances from codon usage metrics with remarkable accuracy, and the simulated phenotypes closely matched experimental data, which demonstrates that these predictive models are valuable tools for systems metabolic engineering approaches.

CHAPTER 1 – LITERATURE REVIEW

1.1. Protein abundance in the context of systems biology of metabolism

Proteins are the primary molecules of cellular function and therefore play an important role in physiological processes and metabolic pathways through the allocation of the cellular proteome (HUI et al., 2015). As such, the determination of the number of protein molecules per cell is a valuable information for modelling the behaviour of biological systems, which is the goal of systems biology (LERMAN et al., 2012). Protein abundance is primarily determined by a combination of factors such as mRNA abundance, translation efficiency, protein turnover rate, and codon usage bias (SHAH et al., 2013). The development of high-throughput sequencing and large-scale studies of macromolecules has led to an intensive generation of data. This has provoked a paradigm shift where the focus is no longer on understanding individual gene functions, but on large-scale signalling, regulatory and structural metabolic pathways inside the cell (IDEKER; LAUFFENBURGER, 2003)

In order to tackle this ever-increasing complexity, computational and mathematical models have been used (VIDAL, 2009). There are two approaches for modelling biological systems, the bottom-up and the top-down (NIELSEN, 2017). In the first, a network of reactions is reconstructed, resulting in a structured database that describes the studied system. The second approach involves the integrative analysis of omics data, mainly by statistical methods (PALSSON, 2015).

Metabolism can be studied by systems biology through bottom-up approaches such as kinetic models and genome-scale metabolic models (GEMs). Kinetic models are restricted to a small number of pathways and make use of differential equations to describe biochemical networks (RESAT; PETZOLD; PETTIGREW, 2009). On the other hand, GEMs aim to reconstruct the entire metabolic network of an organism and employ mathematical optimization techniques such as linear programming in order to predict phenotypes (ORTH; THIELE; PALSSON, 2010). GEMs include the stoichiometry of each reaction in the network, along with cofactor usage, such as NAD, FAD, and Mg^{2+}. Each reaction is also connected to a gene and thus allows to establish a direct correlation between the genotype and the phenotype (NIELSEN, 2017).

Conventional GEMs are reconstructed by using annotated genomes, from which gene usage information is extracted, and by using metabolic pathway databases such

as KEGG, MetaCyc and BRENDA, in which information about metabolites, reactions and enzymes are retrieved (THIELE; PALSSON, 2010). This information is transformed into a mathematical model by representing all reactions and metabolites in a matrix S, called the stoichiometric matrix (Figure 1), where each row is a metabolite, and each column is a reaction. Information about nutrient exchange with the environment and biomass production are also added to the matrix in the form of reactions. A second element, the column vector v, represents the metabolic flux through all of the reactions. By multiplying the rows of S with v, the fluxes for each reaction can be calculated by a system of linear equations (PRICE et al., 2003) based on the assumption of a steady-state metabolism, in which the product must be equal to zero.

$$S = \begin{array}{c} \\ M_1 \\ \vdots \\ M_m \end{array} \begin{array}{c} \begin{array}{ccc} v_1 & \cdots & v_n \end{array} \\ \begin{bmatrix} S_{11} & \cdots & S_{1n} \\ \vdots & \ddots & \vdots \\ S_{m1} & \cdots & S_{mn} \end{bmatrix} \end{array} \longrightarrow \sum_{j=1}^{n} s_{ij} v_j = 0$$

Figure 1: Representation of the stoichiometric matrix. Metabolites are represented by rows, and reactions are represented by columns. The vector $v = v_1 \ldots v_n$ represents the metabolic flux through all reactions. The product of S and v equals zero under the assumption of steady-state. Adapted from Sánchez et al. (2017).

A problem with these equations is that there are fewer metabolites than reactions, which means the solution space is too large for any meaningful solution (ORTH; THIELE; PALSSON, 2010). However, metabolic fluxes can still be calculated given a set of constraints and a defined cellular objective, such as biomass fixation or production of a specific metabolite. This set of constraints represents simulation conditions such as upper and lower bounds for metabolic flux, irreversibility of reactions, and nutrient exchange restrictions while excluding biologically unfeasible solutions (MASSAIU et al., 2019). The system of linear equations is solved by linear programming, and this approach is called flux balance analysis (FBA) (GIANCHANDANI; CHAVALI; PAPIN, 2010).

Conventional GEMs assume that the metabolite production is limited only by the specific substrate uptake rate. However, the metabolic flux is also dependent on enzyme concentration and catalytic efficiency, which are not considered in these conventional GEMs. Thus, many approaches have been proposed to incorporate enzyme constraints into GEMs. One of them is based on models of metabolism and macromolecular expression (ME-models), which uses data from the entire protein biosynthetic machinery (O'BRIEN et al., 2013). Another approach, termed GECKO, incorporates quantitative proteomics and enzyme kinetics (SÁNCHEZ et al., 2017). With this approach, it was possible to simulate the Crabtree effect in *S. cerevisiae* growing in a glucose-limited chemostat with increasing growth rates, which was not possible with conventional models. The integration of both absolute proteomic measurements and enzyme kinetics is incorporated in the stoichiometric matrix, which allows the new models to be compatible with existing modelling tools, such as the COBRA Toolbox (HEIRENDT et al., 2019; VLASSIS; PACHECO; SAUTER, 2014) and RAVEN Toolbox (AGREN et al., 2013; WANG et al., 2018). A new constraint is added to FBA which happens as the following equation defines:

$$v_i \leq k^{ij}{}_{cat} \cdot [E_i] \qquad (2)$$

where v_i is the metabolic flux of reaction R_j (in mmol gDWh^{-1}), $[E_i]$ is the intracellular concentration of enzyme E_i, and $k^{ij}{}_{cat}$ is the enzyme turnover rate of enzyme E_i, catalysing reaction R_j.

Considering the aforementioned context, the contribution of protein abundance to bottom-up systems biology of metabolism can be summarized as an additional constraint that restricts the solution space of a metabolic model (SÁNCHEZ et al., 2017), and with this emerges the need for accurate high-throughput quantification of proteins (LAWLESS et al., 2016).

1.2. Genome-scale metabolic models applications into metabolic engineering

Genome-scale metabolic models have opened unprecedented opportunities in the fields of metabolic engineering and synthetic biology (CHOI et al., 2019). Strain engineering projects no longer have to rely on the empirical trial-and-error experiences of conventional genetic engineering. Instead, a rational design approach can be employed by studying and engineering metabolic flux *in silico* (PETZOLD et al., 2015).

GEMs have been reconstructed for many different organisms and cell types, including bacteria, archaea, fungi, algae, and mammalian cell lines (GU et al., 2019). For industrial relevant microorganisms such as *Saccharomyces cerevisiae*, GEMs have usually been used for predicting titre, accumulation rate, the yield of specific metabolites, and growth under different conditions (NIELSEN, 2017).

One recent endeavour was the model-driven design of *Yarrowia lipolytica*, an oleaginous yeast, for the production of dicarboxylic acids (DCAs) (MISHRA et al., 2018). In this work, the authors expanded on existing models by adding reactions that entail the oxidation of fatty acids into DCAs. Next, they identified potential targets for metabolic engineering, such as the genes coding for malate dehydrogenase, malic enzyme and glutamate dehydrogenase. The simulated overexpression resulted in a 22% increase in the production of dodecanedioic acid, a long-chain DCA.

Another metabolic engineering effort was the enhanced isoprenoid production by *S. cerevisiae* (MEADOWS et al., 2016). The farnesene biosynthesis pathway was improved through the addition of heterologous reactions to the yeast GEM. The engineered strain required fewer moles of glucose per mol of farnesene (4.76 mol to 3.50 mol), while theoretical productivity improved from 2.71 g/L h^{-1} to 11.0 g/L h^{-1}.

Chinese hamster ovary (CHO) cell GEMs have also been used for metabolic engineering purposes. A GEM of this cell line was adapted and tested for the production of monoclonal antibodies in an industrial fed-batch process (CALMELS et al., 2019). This was achieved by the incorporation and correction of many reactions, such as those associated with oxidative stress, lipid metabolism, and amino acid import, and by setting the objective function to antibody production. Potential bottlenecks that impact cell performance were identified, such as amino acid utilization and ammonia accumulation.

1.3. Quantitative proteomics and systems biology of metabolism

Recent advancements of mass spectrometry (MS) and quantitative proteomics have allowed the quantification of thousands of proteins for many organisms (VITRINEL et al., 2019). MS-based absolute quantification methods often rely on spectral counting, ion intensity, stable isotope labelling, and even the use of standardized references (OTTO; BECHER; SCHMIDT, 2014). Absolute measurement methods have been used in microbes to quantify over 2300 proteins of *Escherichia coli*

(SCHMIDT et al., 2016), 1079 proteins of *Bacillus subtilis* (MUNTEL et al., 2014), and over 5000 proteins of *S. cerevisiae* (HO; BARYSHNIKOVA; BROWN, 2018).

It is still challenging to accurately measure an entire proteome by direct methods. The effectiveness of the ionization step can significantly differ between different proteins (ZIADY; KINTER, 2009). The digestion of proteins into analysable peptides is also subjected to variations in efficiency. Moreover, resulting quantifications can be further affected by loss and/or degradation of samples. The variation in physicochemical properties for different proteins also severely influences the resulting signal intensities and ionization efficiencies (OTTO; BECHER; SCHMIDT, 2014). These factors make MS an intrinsically not quantitative method that requires significant effort to generate protein quantifications (PAPPIREDDI; MARTIN; WÜHR, 2019). High-cost of reagents and equipment are also a drawback (SWIATLY et al., 2018).

Absolute protein quantification has been mostly limited to model species, such as *E. coli* and *S. cerevisiae*, which hinders system biology endeavours in non-model species (WILLIAMS et al., 2011). GEMs which account for protein abundance have been reconstructed for a limited number of species. A recent review of ME-models reported only 4 ME-models (YANG et al., 2018), while 108 stoichiometry-only GEMs are available on the BiGG repository (KING et al., 2015). Likewise, GECKO models have only been reconstructed for *S. cerevisiae* (LU et al., 2019; SÁNCHEZ et al., 2017) and *Bacillus subtilis* (MASSAIU et al., 2019). The integration of omics data to GEMs, especially protein abundance, can be useful to improve simulations. For example, the *S. cerevisiae* iMM904 model, which is integrated with proteomic measurements and solved by Linear Bound Flux Balance Analysis (LBFBA), matched more closely experimental fluxomics data than the iMM904 model without proteomics data solved by Parsimonious Flux Balance Analysis (pFBA) (TIAN; REED, 2018).

1.4. Codon usage bias and evolutionary patterns of selection

The elucidation of the genetic code has revealed that 61 codons are responsible for encoding the 20 proteinogenic amino acids, while three codons are reserved for translation interruption. As such, many codons code for the same amino acid and therefore, the genetic code is degenerate. Although synonymous codons are able to code for the same amino acid, comparative analysis of protein-coding sequences has revealed that a “preference” in codon usage is present, where certain codons are used

more frequently than others. This phenomenon was termed codon usage bias (SHARP; LI, 1986). Furthermore, it was found that synonymous codons are not functionally redundant. Considering the stochastic nature of cognate tRNA recognition in the ribosome A-site, a codon can be defined as optimal or non-optimal depending on whether it has a higher or lower chance of cognate tRNA binding, respectively (HANSON; COLLER, 2017). Codons can also be described as frequent or rare, depending on how often a certain codon appears in a coding sequence (SHARP; LI, 1987). Based on this description, codons can be classified as frequent and optimal (FreO), frequent and non-optimal (FreNO), rare and optimal (RareO), and rare and non-optimal (RareNO). This has implications in many processes such as protein folding (ANGOV, 2011; BOGUMIL et al., 2012), co-translational protein processing (KRAMER et al., 2009), translation elongation rate (NOVOA; RIBAS DE POUPLANA, 2012), translation efficiency (HANSON; COLLER, 2017), and recombinant protein production (BUHR et al., 2016).

Codon usage bias is correlated with protein and mRNA levels (ZHOU et al., 2016). There are many methods of measuring codon usage bias: the codon adaptation index (CAI) (SHARP; LI, 1987), the frequency of optimal codons (Fop) (IKEMURA, 1981), the measure of expression E (KARLIN; MRÁZEK, 2000), gene codon bias (GCB) (MERKL, 2003), and MILC-based expression level predictor, MELP (SUPEK; VLAHOVIČEK, 2005). As analysed by Supek & Vlahoviček (2005), for *E. coli* grown in rich media, the Pearson's correlation coefficient for log-transformed MELP and protein abundance was over 0.7. The correlation coefficient fluctuates depending on growth conditions, but codon usage metrics are still very useful for estimating protein and mRNA abundances (JEACOCK; FARIA; HORN, 2018).

The distribution of codons regarding optimality and frequency in a protein-coding sequence is not stochastic, that is, it follows an evolution-selected distribution given their individual contributions to protein biosynthesis (VILLADA; BRUSTOLINI; SILVEIRA, 2017). For instance, 5'- and 3'- extremities of a coding sequence have a strong selection against uniformity, in sharp contrast with more central regions. The pattern of codon composition also impacts protein structures, as certain secondary structures, such as coil regions, have an enrichment of RareNO codons that is not detected in other types of structures. The 5'- extremity is also enriched with RareNO codons, with an average decoding rate that is compatible with ramp theory (VERMA et al., 2019). Furthermore, enzymes from central metabolic pathways are highly abundant

and present strong codon usage bias (i.e., defined pattern of codon usage). Proteins from stress response pathways, on the other hand, are less abundant and have weaker codon usage bias (i.e., codons are uniformly employed) (HANSON; COLLER, 2017; QUAX et al., 2015).

There are selective pressures regarding codon usage that act on the resource allocation for protein biosynthesis (SEWARD; KELLY, 2016, 2018), translation efficiency (GINGOLD; PILPEL, 2011), and translation accuracy (AKASHI, 1994). The resource allocation for protein biosynthesis governs the synthesis of RNA molecules, and polypeptides, as well as other molecules involved in the translational process. One way to reduce the overall biosynthetic cost of a protein without altering its amino acid sequence is to reduce the expenditure on RNA synthesis or increase the translation efficiency of existing RNA molecules (SEWARD; KELLY, 2016). Genome-scale analysis of 1,320 bacterial genomes performed by Seward & Kelly (2018) revealed that genes subjected to strong selection for reducing biosynthetic cost are also subjected to strong selection to increase the translation efficiency. Translation accuracy is also important for reducing the biosynthetic resource cost of a protein and improving translation efficiency, as inaccurate translation elongation increases the time needed to sequester a ribosome, which causes a diminished availability of ribosomes and protein misfolding (YANNAI; KATZ; HERSHBERG, 2018).

1.5. Machine learning applied to biological problems

Machine learning is a subarea of computer science and statistics that studies and employs algorithms for making predictions based on data, without requiring prior explicit instructions. Predictions are done through experience, where they are able to "learn" and generalise given input data (ZAMPIERI et al., 2019). Machine learning methods can be divided in two different approaches (Figure 2), according to the type of data used and the expected output (CUPERLOVIC-CULF, 2018). The first approach is called supervised learning, where the algorithms are trained with labelled data to learn the relation between example inputs and example outputs to infer new outputs based on new inputs. It includes classification algorithms, where the predicted output is a discrete value, and regression algorithms, where the outputs are continuous values. The second approach is unsupervised learning, where algorithms are used to identify patterns in the data without any knowledge of labels, classifications or categories.

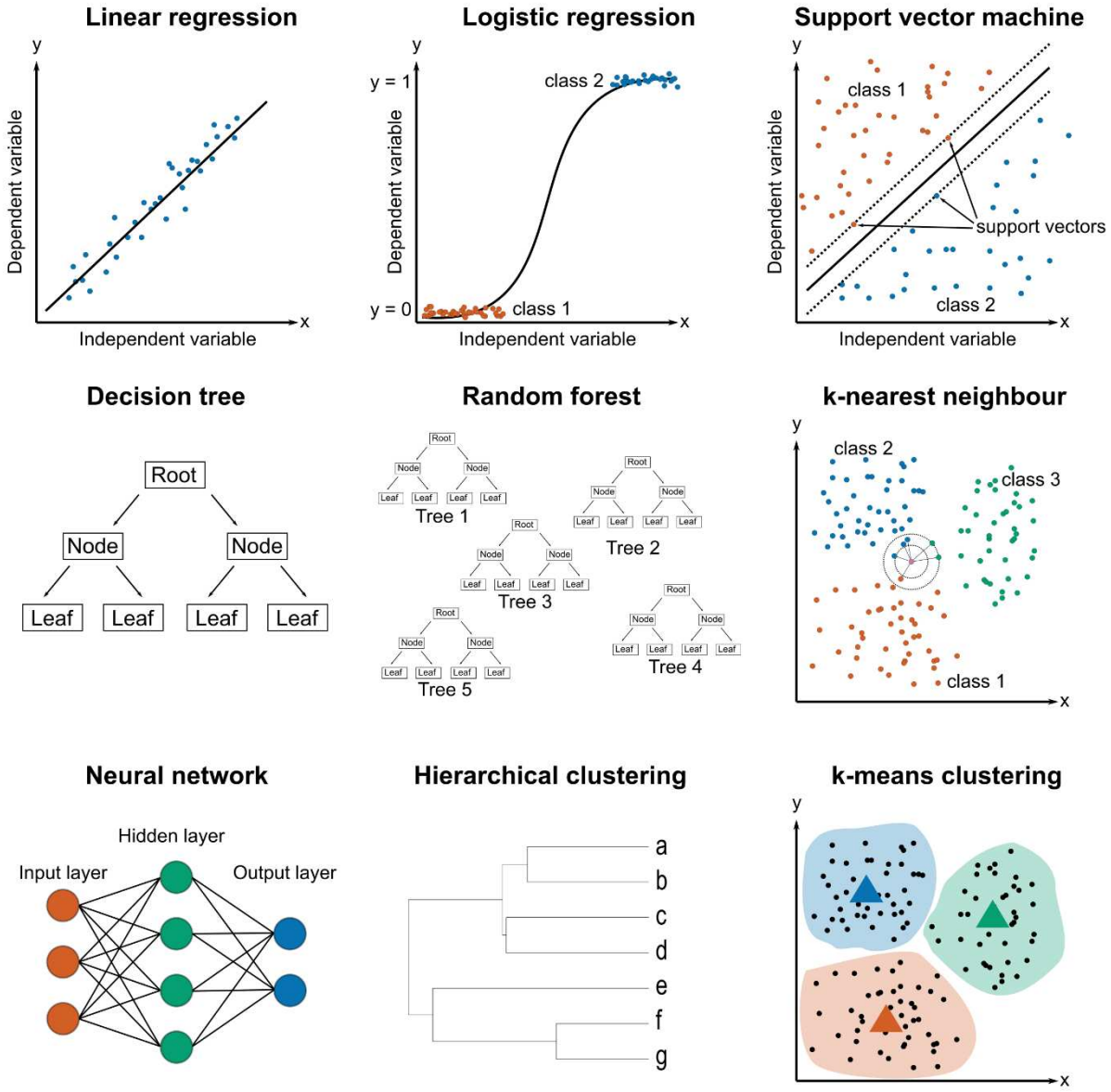


Figure 2: Overview of commonly used machine learning algorithms. The algorithms Linear Regression, Logistic Regression, Support Vector Machines, Decision Tree, Random Forest, k-Nearest Neighbour and Neural Network are examples of supervised learning algorithms, which can be employed to both classification and regression problems. Hierarchical Clustering, k-Means Clustering, and also Neural Network are examples of unsupervised learning algorithms, which are applied to clustering analysis and pattern recognition.

Machine learning has been a useful tool for systems biology, with applications such as the prediction of enzyme turnover numbers (HECKMANN et al., 2018), prediction of metabolomics time-series data from proteomics time-series data (COSTELLO; MARTIN, 2018), automated metabolic model ensemble-driven elimination of uncertainty with statistical learning (MEDLOCK; PAPIN, 2020), and

characterisation and reduction of uncertainty in kinetic models (ANDREOZZI; MISKOVIC; HATZIMANIKATIS, 2016). Thus, we hypothesise that machine learning algorithms can predict protein abundances by using codon usage metrics as features.

There have been many attempts to predict protein abundance values, usually for imputation of abundance values of missing proteins (LI et al., 2011; MEHDI et al., 2014; NIE et al., 2007; TORRES-GARCÍA et al., 2009). Mehdi et al. (2014) utilized a Bayesian network that depends primarily on protein abundance and mRNA properties, such as interaction with proteins, expression, and folding energy. An earlier effort by Nie et al. (2006) proposed a zero-inflated Poisson-based model that employed microarray data to predict protein abundance of *Desulfovibrio vulgaris*. Torres-García et al. (2009) and Li et al. (2010) improved previous work by predicting the abundance of *D. vulgaris* proteins using gradient boosted trees (GBT) and neural networks, respectively.

1.6. Conclusions and perspectives

GEMs are useful tools for systems metabolic engineering. Many new methods to enhance their prediction capacities have been developed and many more are to be expected in the future. The GECKO approach, which relies on the integration of enzymes kinetics and protein abundance, has been remarkably successful in predicting the Crabtree effect in *S. cerevisiae* and in replicating metabolic engineering endeavours. However, as we mentioned in this review, data sets of absolute protein abundance are few and far between, mainly for non-model species. Machine learning methods have many applications in systems metabolic engineering, especially for data imputation and predicting quantitative properties of macromolecules. As quantitative proteomics methods continue to evolve, it is expected that more data will be available. Until then, we believe machine learning can be used to fill this gap, especially using biological information that correlate well with protein abundance, such as codon usage bias.

1.7 References

AGREN, R. et al. The RAVEN Toolbox and Its Use for Generating a Genome-scale

Metabolic Model for Penicillium chrysogenum. **PLoS Computational Biology**, v. 9, n. 3, p. e1002980, 21 mar. 2013.

AKASHI, H. Synonymous Codon Usage. **Genetics Society of America**, v. 136, p. 927–935, 1994.

ANDREOZZI, S.; MISKOVIC, L.; HATZIMANIKATIS, V. ISCHRUNK - In Silico Approach to Characterization and Reduction of Uncertainty in the Kinetic Models of Genome-scale Metabolic Networks. **Metabolic Engineering**, v. 33, p. 158–168, 2016.

ANGOV, E. Codon usage: Nature's roadmap to expression and folding of proteins. **Biotechnology Journal**, v. 6, n. 6, p. 650–659, 2011.

BOGUMIL, D. et al. Chaperones divide yeast proteins into classes of expression level and evolutionary rate. **Genome Biology and Evolution**, v. 4, n. 5, p. 618–625, 2012.

BUHR, F. et al. Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. **Molecular Cell**, v. 61, n. 3, p. 341–351, 2016.

CALMELS, C. et al. Application of a curated genome-scale metabolic model of CHO DG44 to an industrial fed-batch process. **Metabolic Engineering**, v. 51, n. June 2018, p. 9–19, 2019.

CHOI, K. R. et al. Systems Metabolic Engineering Strategies: Integrating Systems and Synthetic Biology with Metabolic Engineering. **Trends in Biotechnology**, v. 37, n. 8, p. 817–837, 2019.

COSTELLO, Z.; MARTIN, H. G. A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. **npj Systems Biology and Applications**, v. 4, n. 1, p. 1–14, 2018.

CUPERLOVIC-CULF, M. Machine learning methods for analysis of metabolic data and metabolic pathway modeling. **Metabolites**, v. 8, n. 1, 2018.

GIANCHANDANI, E. P.; CHAVALI, A. K.; PAPIN, J. A. The application of flux balance analysis in systems biology. **Wiley Interdisciplinary Reviews: Systems Biology and Medicine**, v. 2, n. 3, p. 372–382, 2010.

GINGOLD, H.; PILPEL, Y. Determinants of translation efficiency and accuracy. **Molecular Systems Biology**, v. 7, n. 481, p. 1–13, 2011.

GU, C. et al. Current status and applications of genome-scale metabolic models. **Genome Biology**, v. 20, n. 1, p. 121, 13 dez. 2019.

HANSON, G.; COLLER, J. Codon optimality, bias and usage in translation and mRNA decay. **Nature Reviews Molecular Cell Biology**, 2017.

HECKMANN, D. et al. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. **Nature Communications**, v. 9, n. 1, p. 5252, 7 dez. 2018.

HEIRENDT, L. et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. **Nature Protocols**, v. 14, n. 3, p. 639–702, 2019.

HO, B.; BARYSHNIKOVA, A.; BROWN, G. W. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. **Cell Systems**, v. 6, n. 2, p. 192- 205.e3, 2018.

HUI, S. et al. Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. **Molecular Systems Biology**, v. 11, n. 2, p. 784, 2015.

IDEKER, T.; LAUFFENBURGER, D. Building with a scaffold: Emerging strategies for high- to low-level cellular modeling. **Trends in Biotechnology**, v. 21, n. 6, p. 255–262, 2003.

IKEMURA, T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. **Journal of Molecular Biology**, v. 151, n. 3, p. 389–409, 25 set. 1981.

JEACOCK, L.; FARIA, J.; HORN, D. Codon usage bias controls mRNA and protein abundance in trypanosomatids. **eLife**, v. 7, p. 1–20, 2018.

KARLIN, S.; MRÁZEK, J. Predicted highly expressed genes of diverse prokaryotic genomes. **Journal of bacteriology**, v. 182, n. 18, p. 5238–50, set. 2000.

KING, Z. A. et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. **Nucleic Acids Research**, v. 44, n. D1, p. D515–D522, 17 out. 2015.

KRAMER, G. et al. The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. **Nature Structural and Molecular Biology**, v. 16, n. 6, p. 589–597, 2009.

LAWLESS, C. et al. Direct and absolute quantification of over 1800 yeast proteins via selected reaction monitoring. **Molecular and Cellular Proteomics**, v. 15, n. 4, p. 1309–1322, 2016.

LERMAN, J. A. et al. In silico method for modelling metabolism and gene product expression at genome scale. **Nature Communications**, v. 3, n. May, 2012.

LI, F. et al. Prediction and Characterization of Missing Proteomic Data in *Desulfovibrio vulgaris*. **Comparative and Functional Genomics**, v. 2011, p. 780973, 2011.

LU, H. et al. A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. **Nature Communications**, v. 10, n. 1, 2019.

MASSAIU, I. et al. Integration of enzymatic data in Bacillus subtilis genome-scale metabolic model improves phenotype predictions and enables in silico design of poly-γ-glutamic acid production strains. **Microbial Cell Factories**, v. 18, n. 1, p. 1–20, 2019.

MEADOWS, A. L. et al. Rewriting yeast central carbon metabolism for industrial isoprenoid production. **Nature**, v. 537, n. 7622, p. 694–697, 2016.

MEDLOCK, G. L.; PAPIN, J. A. Guiding the Refinement of Biochemical Knowledgebases with Ensembles of Metabolic Networks and Machine Learning. **Cell Systems**, v. 10, n. 1, p. 109- 119.e3, 2020.

MEHDI, A. M. et al. Predicting the dynamics of protein abundance. **Molecular and Cellular Proteomics**, v. 13, n. 5, p. 1330–1340, 1 maio 2014.

MERKL, R. A Survey of Codon and Amino Acid Frequency Bias in Microbial Genomes Focusing on Translational Efficiency. **Journal of Molecular Evolution**, v. 57, n. 4, p.

453–466, 1 out. 2003.

MISHRA, P. et al. Genome-scale model-driven strain design for dicarboxylic acid production in Yarrowia lipolytica. **BMC Systems Biology**, v. 12, n. Suppl 2, 2018.

MUNTEL, J. et al. Comprehensive Absolute Quantification of the Cytosolic Proteome of Bacillus Subtilis by Data Independent, Parallel Fragmentation in Liquid Chromatography/Mass Spectrometry (LC/MSE). **Molecular and Cellular Proteomics**, v. 13, n. 4, p. 1008–1019, 2014.

NIE, L. et al. **Integrative analysis of transcriptomic and proteomic data: Challenges, solutions and applicationsCritical Reviews in Biotechnology**Taylor & Francis, , abr. 2007. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/07388550701334212>. Acesso em: 12 jul. 2020

NIELSEN, J. Systems Biology of Metabolism. **Annual Review of Biochemistry**, v. 86, n. 1, p. 245–275, 2017.

NOVOA, E. M.; RIBAS DE POUPLANA, L. Speeding with control: Codon usage, tRNAs, and ribosomes. **Trends in Genetics**, v. 28, n. 11, p. 574–581, 2012.

O'BRIEN, E. J. et al. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. **Molecular Systems Biology**, v. 9, n. 693, 2013.

ORTH, J. D.; THIELE, I.; PALSSON, B. Ø. What is flux balance analysis? **Nature Biotechnology**, v. 28, n. 3, p. 245–248, 1 mar. 2010.

OTTO, A.; BECHER, D.; SCHMIDT, F. Quantitative proteomics in the field of microbiology. **Proteomics**, v. 14, n. 4–5, p. 547–565, 2014.

PALSSON, B. Ø. **Systems Biology: Constraint-based Reconstruction and Analysis**. 1. ed. Cambridge: Cambridge University Press, 2015.

PAPPIREDDI, N.; MARTIN, L.; WÜHR, M. A Review on Quantitative Multiplexed Proteomics. **ChemBioChem**, v. 20, n. 10, p. 1210–1224, 2019.

PETZOLD, C. J. et al. Analytics for metabolic engineering. **Frontiers in Bioengineering and Biotechnology**, v. 3, n. September, p. 1–11, 2015.

QUAX, T. E. F. et al. Codon Bias as a Means to Fine-Tune Gene Expression. **Molecular Cell**, v. 59, n. 2, p. 149–161, 2015.

RESAT, H.; PETZOLD, L.; PETTIGREW, M. F. Kinetic Modeling of Biological Systems. In: IRETON, R. et al. (Eds.). . **Computational Systems Biology**. [s.l: s.n.]. v. 541p. 181–210.

SÁNCHEZ, B. J. et al. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. **Molecular Systems Biology**, v. 13, n. 8, p. 935, 2017.

SCHMIDT, A. et al. The quantitative and condition-dependent Escherichia coli proteome. **Nature Biotechnology**, v. 34, n. 1, p. 104–110, 2016.

SEWARD, E. A.; KELLY, S. Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. **Genome Biology**, v. 17, n. 1, p. 226, 15 dez. 2016.

SEWARD, E. A.; KELLY, S. Selection-driven cost-efficiency optimization of transcripts modulates gene evolutionary rate in bacteria. **Genome biology**, v. 19, n. 1, p. 102, 2018.

SHAH, P. et al. Rate-limiting steps in yeast protein translation. **Cell**, v. 153, n. 7, p. 1589, 2013.

SHARP, P. M.; LI, W.-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. **Nucleic Acids Research**, v. 15, n. 3, p. 1281–1295, 11 fev. 1987.

SHARP, P. M.; LI, W. H. An evolutionary perspective on synonymous codon usage in unicellular organisms. **Journal of Molecular Evolution**, v. 24, n. 1–2, p. 28–38, 1986.

SUPEK, F.; VLAHOVIČEK, K. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. **BMC Bioinformatics**, v. 6, n. 1, p. 182, 19 jul. 2005.

SWIATLY, A. et al. Mass spectrometry-based proteomics techniques and their application in ovarian cancer research. **Journal of Ovarian Research**, v. 11, n. 1, p. 1–13, 2018.

THIELE, I.; PALSSON, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. **Nature protocols**, v. 5, n. 1, p. 93–121, jan. 2010.

TIAN, M.; REED, J. L. Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis. **Bioinformatics**, v. 34, n. 22, p. 3882–3888, 2018.

TORRES-GARCÍA, W. et al. Integrative analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: a non-linear model to predict abundance of undetected proteins. **Bioinformatics**, v. 25, n. 15, p. 1905–1914, 15 maio 2009.

VERMA, M. et al. A short translational ramp determines the efficiency of protein synthesis. **Nature Communications**, v. 10, n. 1, p. 1–15, 2019.

VIDAL, M. A unifying view of 21st century systems biology. **FEBS Letters**, v. 583, n. 24, p. 3891–3894, 2009.

VILLADA, J. C.; BRUSTOLINI, O. J. B.; SILVEIRA, W. B. DA. Integrated analysis of individual codon contribution to protein biosynthesis reveals a new approach to improving the basis of rational gene design. **DNA Research**, v. 0, n. 0, p. 1–16, 1 ago. 2017.

VITRINEL, B. et al. Exploiting Interdata Relationships in Next-generation Proteomics Analysis. **Molecular & Cellular Proteomics**, v. 18, n. 8 suppl 1, p. S5–S14, 2019.

VLASSIS, N.; PACHECO, M. P.; SAUTER, T. Fast Reconstruction of Compact Context-Specific Metabolic Network Models. **PLoS Computational Biology**, v. 10, n. 1, p. 1003424, jan. 2014.

WANG, H. et al. RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. **PLOS Computational Biology**, v. 14, n. 10, p. e1006541, 18 out. 2018.

WILLIAMS, T. D. et al. Towards a system level understanding of non-model organisms

sampled from the environment: A network biology approach. **PLoS Computational Biology**, v. 7, n. 8, 2011.

YANG, L. et al. Modeling the multi-scale mechanisms of macromolecular resource allocation. **Current Opinion in Microbiology**, v. 45, p. 8–15, 2018.

YANNAI, A.; KATZ, S.; HERSHBERG, R. The codon usage of lowly expressed genes is subject to natural selection. **Genome Biology and Evolution**, v. 10, n. 5, p. 1237–1246, 2018.

ZAMPIERI, G. et al. Machine and deep learning meet genome-scale metabolic modeling. **PLOS Computational Biology**, v. 15, n. 7, p. e1007084, 2019.

ZHOU, Z. et al. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. **Proceedings of the National Academy of Sciences**, v. 113, n. 41, p. E6117–E6125, 2016.

ZIADY, A. G.; KINTER, M. Protein sequencing with tandem mass spectrometry. **Methods in molecular biology (Clifton, N.J.)**, v. 544, n. September, p. 325–341, 2009.

CHAPTER 2 – RESEARCH ARTICLE

Research article written following the Nucleic Acids Research journal author guidelines.

Protein Abundance Prediction Through Machine Learning Methods

Mauricio Ferreira[1], Rafaela Ventorim[1], Eduardo Almeida[1], Sabrina Silveira[2], Wendel Silveira[1,*]

[1] Department of Microbiology, Universidade Federal de Viçosa, Viçosa, Minas Gerais, 36570-900, Brazil
[2] Department of Computer Science, Universidade Federal de Viçosa, Viçosa, Minas Gerais, 36570-900, Brazil

* To whom correspondence should be addressed. Tel: +55 31 3612-2431; Email: wendel.silveira@ufv.br

Present Address: Wendel Silveira, Department of Microbiology, Universidade Federal de Viçosa, Viçosa, Minas Gerais, 36570-900, Brazil

ABSTRACT

Proteins are responsible for most physiological processes, and their abundance provides crucial information for systems biology research. However, absolute protein quantification, as determined by mass spectrometry, still has limitations in capturing the protein pool. Protein abundance is impacted by translation kinetics, which rely on features of codons. In this study, we evaluated the effect of codon usage bias of genes on protein abundance. Notably, we observed differences regarding codon usage patterns between genes coding for highly abundant proteins and genes coding for less abundant proteins. Analysis of synonymous codon usage and evolutionary selection showed a clear split between the two groups. Our machine learning models predicted protein abundances from codon usage metrics with remarkable accuracy, achieving R^2 values higher than previously reported in the literature. Upon integration of the predicted protein abundance in enzyme-constrained genome-scale metabolic models, the simulated phenotypes closely matched experimental data, which demonstrates that our predictive models are valuable tools for systems metabolic engineering approaches.

INTRODUCTION

Proteins are the primary molecules of cellular function. The efficient allocation of the cellular proteome is responsible for controlling metabolic flux and many physiological processes (1). The absolute abundance of each protein is valuable information for genome-scale metabolic reconstructions, as it improves the modelling of metabolic flux and protein allocation (2). Current mass spectrometry technology and quantitative proteomics analysis have allowed the quantification of thousands of proteins in different organisms (3). However, a large portion of proteins are still undetected (4) due to variation in their physicochemical properties, signal intensities, and ionisation efficiencies (5). The high-cost of reagents and equipment is another drawback (6). Absolute quantification has been mostly limited to model species, which hinders systems biology endeavours in non-model species (7). Genome-scale metabolic models (GEMs), which account for protein abundance, have been reconstructed for a limited number of species. A recent review of models of metabolism and macromolecular expression (ME-models) reported only 4 ME-models (8), while 108 stoichiometry-only GEMs (M-model) are available on the BiGG repository (9). Likewise, GEMs with enzymatic constraints using kinetics and omics data (GECKO models) have only been reconstructed for *Saccharomyces cerevisiae* (10, 11) and *Bacillus*

subtilis (12). The integration of omics data to GEMs, especially protein abundance, can be useful to improve simulations. For example, the *S. cerevisiae* iMM904 model, which is integrated with proteomic measurements and solved by Linear Bound Flux Balance Analysis (LBFBA), matched more closely experimental fluxomics data than the iMM904 model without proteomics data solved by Parsimonious Flux Balance Analysis (pFBA) (13). This finding highlights the importance of further increasing the number of GEMs integrated with protein abundance.

The abundance of proteins is primarily determined by a combination of factors, such as mRNA abundance, translation efficiency, protein turnover rate, and codon usage bias (CUB) (14). CUB, which is a phenomenon in which certain synonymous codons are employed more frequently than other codons (15), is positively correlated with protein abundance (16). Codons can be optimal or non-optimal, depending on their average decoding time (the time needed for a ribosome to read it) (17), and frequent and rare, depending on how often a certain codon appears in a coding sequence (CDS) (18). Based on this description, codons can be classified as frequent and optimal (FreO), frequent and non-optimal (FreNO), rare and optimal (RareO), and rare and non-optimal (RareNO). The distribution of codons regarding optimality and frequency in a protein-coding sequence is not stochastic, that is, it follows an evolution-selected distribution given their individual contributions to protein biosynthesis (19). For instance, 5'- and 3'- extremities of a CDS have a strong selection against uniformity, in sharp contrast with more central regions. The pattern of codon composition also impacts protein structures, as certain secondary structures, such as coil regions, have an enrichment of RareNO codons that is not detected in other types of structures. The 5'- extremity is also enriched with RareNO codons, with an average decoding rate that is compatible with ramp theory (20). Furthermore, enzymes from central metabolic pathways are highly abundant and present strong codon usage bias (i.e., defined pattern of codon usage). Proteins from stress response pathways, on the other hand, are less abundant and have weaker codon usage bias (i.e., codons are uniformly employed) (17, 21).

As protein abundance is generally conserved across diverse phylogenetic taxa (22), this poses the question of whether absolute protein abundance can be mathematically predicted based on existing data. Based on the association between protein abundance and codon usage bias, it is reasonable to consider that metrics of codon usage bias are a potential source of information that can be applied to predict the abundance of proteins of non-model species or proteins undetected by mass spectrometry.

Machine learning has been a useful tool for systems biology, with applications such as the prediction of enzyme turnover numbers (23), prediction of metabolomics time-series data from proteomics time-series data (24), automated metabolic model ensemble-driven elimination of uncertainty with statistical learning (25), and characterisation and reduction of uncertainty in kinetic models (26). Thus, we hypothesise that machine learning algorithms can predict protein abundances by using codon usage metrics as features.

In this study, we explore the evolutionary constraints that shape the codon usage bias of the *Saccharomyces cerevisiae* genome in the context of protein abundance by comparing highly abundant proteins (HAP) and lowly abundant proteins (LAP). We then apply supervised learning algorithms to a series of codon usage metrics calculated from protein-coding sequences with known protein abundance to predict the abundance of proteins employed by enzyme-constrained genome-scale metabolic models

(ecGEMs), including previously undetected proteins. The integration of these predictions in ecGEMs enables phenotype simulations that match those performed with experimentally measured protein abundances.

MATERIAL AND METHODS

Data collection

The protein abundance values were obtained from a unified data set of *Saccharomyces cerevisiae* proteome quantifications, which were compiled by Ho et al. (27). This dataset has absolute abundance values that exceed 5,000 proteins; it is composed of absolute and relative measurements from 21 quantitative proteomics analyses. For all analyses, we applied the median absolute abundance values after filtering for GFP autofluorescence. Additionally, we also employed median values for measurements performed using either the minimal medium or the YPD medium.

We retrieved the CDS associated with each protein reported in the proteomics study from Ensembl Fungi (28), using the BioMart tool (29). We obtained other information, such as amino acid sequences, size and molecular weight from UniProt (30). For the tRNA analysis, we recovered the *S. cerevisiae* tRNA gene copy numbers from GtRNAdb (31, 32).

Characterisation of codon usage

To characterise codon usage bias in the *S. cerevisiae* genome, we assessed how evolutionary constraints affect codon distribution and protein biosynthesis regarding protein abundance. Of the 5,388 proteins with reported absolute abundance values, we chose the top 10% of proteins (538 proteins) with the highest overall median abundances and the top 10% proteins with the lowest overall median abundances. Proteins whose number of molecules per cell was less than 100 and were detected by only one proteomics study were not considered. We analysed the attributes of codons, such as frequency, optimality, and positional dependency, for the CDSs of proteins retrieved from the previously mentioned data set.

First, we analysed the codon positional dependency using the CodG package (19) to assess how different positions in the sequences are under evolutionary selective pressure. We employ a binning scheme, as described by Villada et al. (19), where codons in a sequence are divided in a binning scheme relative to the CDS length. Each CDS is divided into 10 parts, which includes information about codon quantity per tenth part. We generated a matrix of 59 rows, which includes all codons, except start, stop and tryptophan (since it has only one codon), and 10 columns that correspond to the 10 parts. Using this matrix, we tested the codon distribution uniformity by calculating the χ^2 value of each codon, as described by Hockenberry et al. (33), using the following equation:

$$\chi^2 = \sum_{i=1}^{n} Z^2 = \sum_{i=1}^{n} \frac{(O-E)^2}{\sigma^2} \quad (1)$$

where O is the observed count per bin in the group of sequences that are being tested, E is the expected value, and σ is the standard deviation of the codon counts per bin obtained from 200 simulated groups

of sequences, where each amino acid sequence of each CDS in each group (highly or lowly abundant proteins) was conserved but the codons were randomised. The significance of each codon is reported by a p-value of 1.7E-4 after a Bonferroni correction for 59 tests, where the p-value = 0.01, and contrasts them with the χ^2 distributions, which have n - 1 degrees of freedom. For the codon frequency, we calculated the Relative Synonymous Codon Usage (RSCU) (18) using the following equation:

$$RSCU = \frac{O_{ij}}{\left[\sum_j^{Ni} O_{ij}\right] \times \frac{1}{Ni}} \qquad (2)$$

where O_{ij} is the frequency of the j-th codon for the i-th amino acid, and Ni is the total number of synonymous codons for the i-th amino acid. According to Nasrullah et al. (34), codons with RSCU ≥ 1 are considered frequent codons, and codons with RSCU < 1 are considered rare. We also performed principal component analysis (PCA) to identify the distance or relatedness of RSCU values using Orange3 data mining software (35). Next, we calculated the selection on transcript translational efficiency (St) and selection on transcript biosynthetic cost (Sc), as described by Seward & Kelly (36, 37). Subsequently, we performed the Akashi Test (38) to calculate the selection on translation accuracy using the software Seforta (39).

Feature compilation for machine learning

To evaluate whether machine learning could capture any underlying pattern and predict absolute abundance values using codon usage metrics as features, we compiled a set of features to build predictive machine learning models. We selected codon usage metrics that are calculated individually for each gene. For each CDS, we calculated various codon usage metrics (Table S1) by using EMBOSS (40), CodonW (41), CAIcal (42), coRdon (43), stAIcalc (44), and scripts included in the original manuscripts, such as CodonMuSe (37) and iCUB (45). For codon usage metrics that require highly expressed genes as a reference, we employed the CDSs of the top 10% proteins with the highest overall median abundance.

Data-set construction

We compiled three separate datasets for training (Figure 1). The first data set uses the median absolute abundances of all 21 quantitative proteomics analyses described by Ho et al. (27) and has a total of 5,388 instances. The second training data set employs the median absolute abundances of experiments that quantified proteins of *S. cerevisiae* yeast growing in minimal media. Eleven quantification experiments were taken into account (abbreviated by Ho et al. (27) as PENG, LAW, LAHT, DGD, THAK, TKA, BRE, DEN, MAZ, CHO, and YOF), which generated 5,187 instances. The third training data set is obtained from experiments in which *S. cerevisiae* is grown in YPD medium (LU, KUL, LEE2, NAG, PIC, WEB, NEW, LEE, DAV, GHA); 5,114 instances are generated. The codon usage metrics listed in Table S1 yielded a total of 91 features, including individual gene codon usage metrics and nucleotide composition numbers (Table S2 and Table S3). After compiling the training data sets, we log-transformed the protein abundance values for the three data sets to favour a normal distribution.

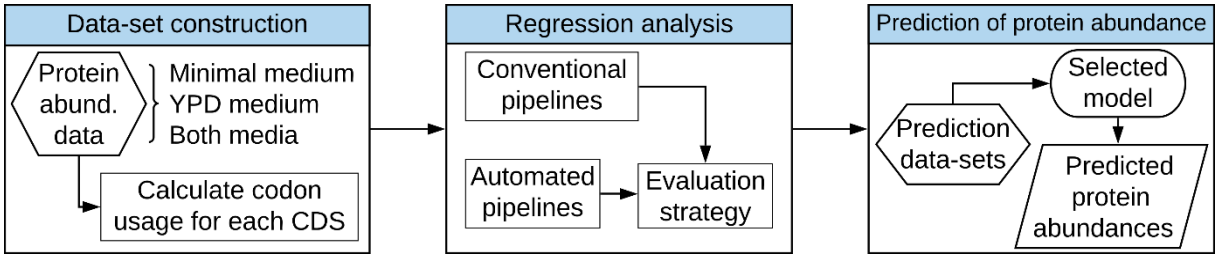


Figure 1. Machine learning applied to codon usage bias for the prediction of protein abundance. We calculated a series of codon usage metrics and nucleotide composition numbers. Three separate data sets were compiled and employed for the training: the first data set contains protein measurements from yeast cultivation in the minimal medium, the second data set consists of protein measurements from yeast cultivation in the YPD medium, and the third data set combines protein measurements for both culture media. The hexagonal boxes denote the input data; the rectangular boxes indicate the intermediate steps; the ellipsoids represent the trained regression models; and the parallelogram denotes the output from these models.

Regression model training

We explored different machine learning libraries to build a predictive model, namely, Scikit-Learn (46), H2O (47), and XGBoost (48). Since the objective was to predict the absolute abundance of proteins (numerical value), we employed all available algorithms with support for regression problems in the previously mentioned libraries, such as linear models, ensemble models, and neural networks. We also applied automated machine learning pipelines from H2O, TPOT (49), and GAMA (50). The H2O automated machine learning tool trains and cross-validates pre-configured algorithms included in the library to select the best algorithm. TPOT and GAMA utilize genetic algorithms to explore many different possible pipelines using Scikit-Learn algorithms to identify the best pipeline. TPOT also exports the code of its predicted pipeline. We describe the best algorithm for each training data set in the Supplementary Appendix. All the source code was written in Python 3.7 (51).

For manually configured pipelines, we randomly split the data set into two subsets and use 75% for training and 25% for testing. To determine the hyperparameters of each regressor, we applied a randomized approach, where each parameter is sampled from a set of possible parameter values. We selected the values that produced models with the best evaluation metrics when we employed the test data set as input. For the automated ML pipelines, we ran H2O for a maximum runtime of 6 hours with an unlimited number of tested models. We ran TPOT with 1,000 generations and a population size of 250. For GAMA, we employed an asynchronous evolutionary optimization algorithm with a maximum runtime of 3,600 seconds. We trained all automated pipelines with 10-fold cross-validation, where the data set is partitioned into 10 equally sized subsets. One of these subsets is randomly chosen as the test data set; this process is repeated until each subset has been utilized for a test exactly once.

Model evaluation and selection

After training the models, we evaluated their performance with a separate data set using data that had not been selected for training. We predicted the absolute abundance values of the enzymes

integrated in the ecYeast7 ecGEM and compared them to the median overall abundance data from Ho et al. (27). We checked the median absolute deviation (MAD) and the coefficient of determination (R^2). The MAD is defined as the median value of all absolute differences between the predicted values and the real values. We chose this metric since the absolute abundance of proteins employed for training are expressed as a median and it is robust to outliers. This metric can be calculated according to the following equation:

$$MAD(y, \hat{y}) = median(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|) \quad (3)$$

where $\hat{y}_i$ is the predicted value of the i-th sample and y_i is the known protein abundance value. When evaluating the predictive models, we searched for models with the smallest possible MAD. R^2 is a measure that represents the variance explained by independent values in the regression model; it is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

where y_i is the known protein abundance value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the observed data. We searched for models with an R^2 equal to or greater than 0.75. Models with a high MAD and low R^2 were re-trained with adjusted hyperparameters until they converged. We selected the model for each data set with the highest R^2 and lowest MAD for the next step.

Integration of predicted abundances in the enzyme-constrained metabolic model ecYeast8

After selecting the best predictive models, we decided to predict the abundance of all enzymes employed by the ecYeast8 ecGEM, which has 968 enzymes integrated with enzyme turnover values (*kcat*) but no integration of protein abundance values. Codon usage metrics were calculated for the CDS of all 968 enzymes. We compiled three prediction data sets in accordance to the medium culture utilized for protein quantifications: minimal medium, YPD medium, and a combination of these two media. The difference between each data set is the reference highly expressed genes applied to calculate the codon usage metrics, as described in the feature compilation section.

To validate the predictions obtained by machine learning, we integrated the predicted protein abundances in the ecYeast8 model. We applied the GECKO Toolbox (10) to set the protein abundances as the upper bound for reactions that use enzymes. Considering that the predicted abundance values are expressed as the number of molecules per cell and the ecGEM requires protein abundance values in millimoles per gram of cell dry weight (mmol/g_{DW}), the values were converted to mmol/g_{DW} by assuming a total cellular protein mass of 0.448 gram of protein per gram of cell dry weight (g/g_{DW}), a cellular density of 1.5 x 10^7 cells per litre (cells/L), and a total biomass of 3 grams of cells dry weight per litre (g_{DW}/L). These values were obtained from (10), (52), and (53), respectively. We re-fitted the parameters *f*, *GAM* and *sigma* using the batch model before constraining the enzymes. After integrating the protein abundances, the upper bound of the enzyme-constrained reactions was flexibilized to optimize growth at a dilution rate of 0.1 (h^{-1}) by setting the carbon source as glucose with an uptake rate

of 1.1 mmol/gDWh^{-1}. To check whether the unit conversion step was performed correctly, we also re-ran the enzyme abundance integration step with ecYeast7 using the quantitative proteomics data from Lahtvee et al. (54).

Model growth simulations

We attempted to replicate the results obtained by Sánchez et al. (10) and compared the simulations of the Yeast7 model with the ecYeast7 model using experimental data obtained from (55). We simulated a chemostat with a dilution rate fixed at 0.1 h^{-1}. We removed any constraints for substrate uptake and limited unmeasured enzyme mass by 0.448 g/gDW, set the *f* value to 0.4421 g/g, and used a σ value of 0.5. For the optimization, we minimized the glucose uptake rate and fixed the glucose uptake rate to the optimal value with a 0.1% flexibility, and minimized enzyme usage. Regarding the ecYeast8 model integrated with ML-predicted abundances, we simulated a chemostat with the same previously mentioned conditions, except that unmeasured enzyme mass was not limited, as the protein pool pseudo-reaction was not included in the models. We compared the results to experimental fluxes measured by ^{13}C-MFA obtained from (55). For all simulations, we checked the metabolic fluxes regarding consumption of glucose and O_2 and the production of CO_2. We have also compared the models by executing flux variability analysis (FVA). All model configuration and problem setups were performed with the RAVEN Toolbox 2 (56) in MATLAB 2017a (The MathWorks Inc., Natick, Massachusetts). Problems were solved using the Gurobi Optimizer version 8.11 (57).

RESULTS

Codon usage is markedly different between the coding sequences of highly abundant proteins and those of lowly abundant proteins

We evaluated how evolutionary constraints shape the codon usage bias of the *S. cerevisiae* genome by comparing the CDSs of HAP and LAP. We observed a noticeable contrast between the CDSs of HAP and those of LAP. Regarding the codon frequency, in the principal component analysis of RSCU values, two distinct groups of CDSs were observed. The first group is composed of mostly HAP, whilst the second group consists of mostly LAP (Figure S1). A heatmap of the 20 most and least abundant proteins showed that codons of HAP CDSs are more enriched with frequent codons and that rare codons are totally or almost totally depleted (RSCU ≈ 0). Meanwhile, CDSs of LAP have a weaker bias in codon usage, that is, there is no preference for certain synonymous codons (Figure 2).

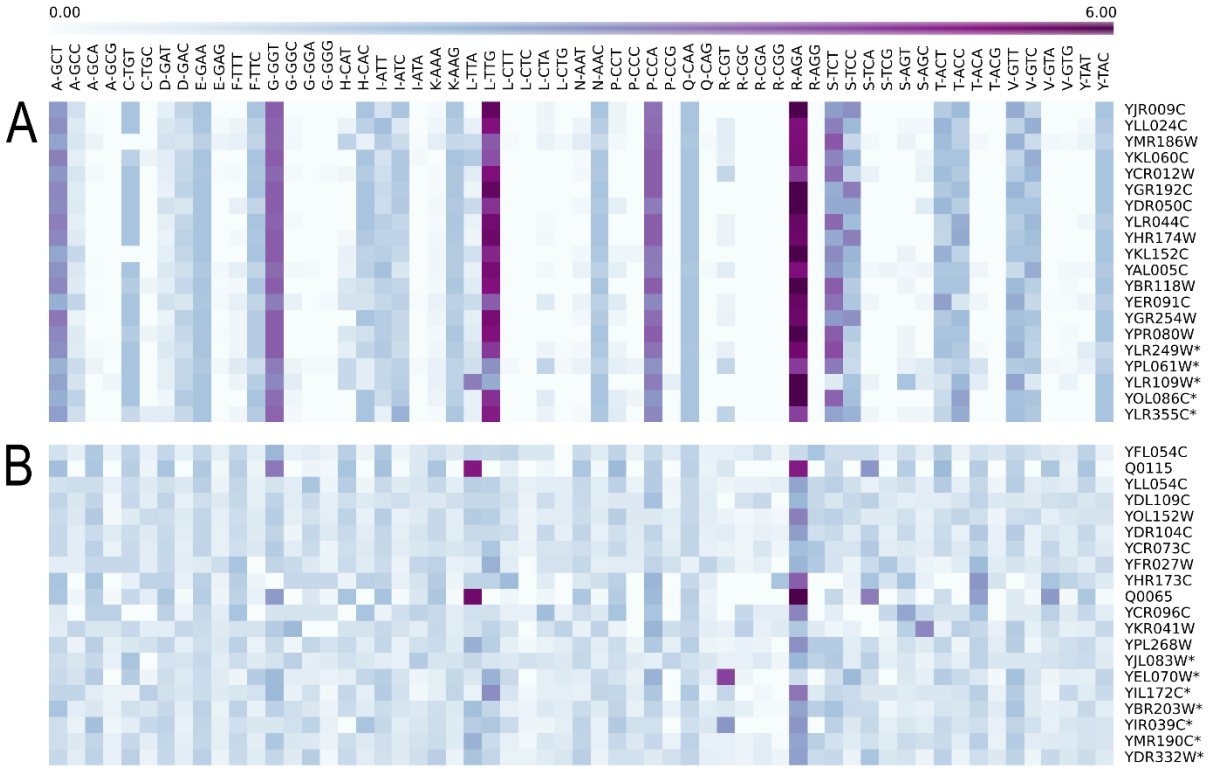


Figure 2. Heatmap of relative synonymous codon usage that illustrates the difference in codon usage in regard to frequency between the 20 most abundant and 20 least abundant proteins of *S. cerevisiae*. Columns denote the 59 codons with one or more synonymous codons. (A) Rows indicate CDSs of the 20 most abundant proteins. (B) Rows represent CDSs of the 20 least abundant proteins. CDSs of the HAP, in contrast to the CDSs of LAP, have many enriched (high RSCU) or depleted (low RSCU) codons. Sequences marked with an asterisk denote the presence of a signal peptide at the 5' extremity as detected by SignalP 5 (58).

For the top 10% proteins with the highest abundances (n = 538), we observed selection against uniformity at the 5' end in the CDSs, as indicated by the enriched codons in the first bin. Otherwise, CDS encoding for LAPs (top 10% lowest, n = 538) presented a higher uniformity of codon distribution (Figure 3 and Figure S2).

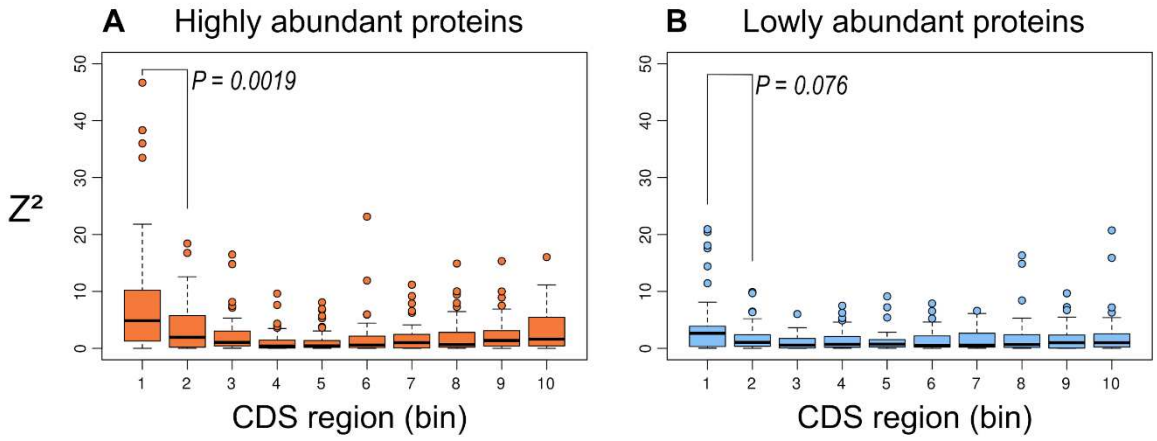


Figure 3. Evolutionary selection for position-dependent codon usage bias determined by the chi-squared test for the matrix constructed with the CodG package. Each CDS was equally divided into 10 bins to evaluate how each position contributes to overall codon usage. (A) Deviation from uniformity in the CDSs of HAPs shows bias towards 5' end. (B) CDSs of lowly abundant proteins show higher uniformity.

Since the values of translation efficiency (St) and selection for biosynthetic costs (Sc) represent how strongly natural selection acts on the translational efficiency and biosynthetic cost of codons, respectively, we calculated them for both HAP and LAP. Importantly, Sc values suggest that HAP CDSs undergo selection pressures to reduce the biosynthetic cost (Sc < 0) and maximize the translation efficiency (St > 0) (Figure 4).

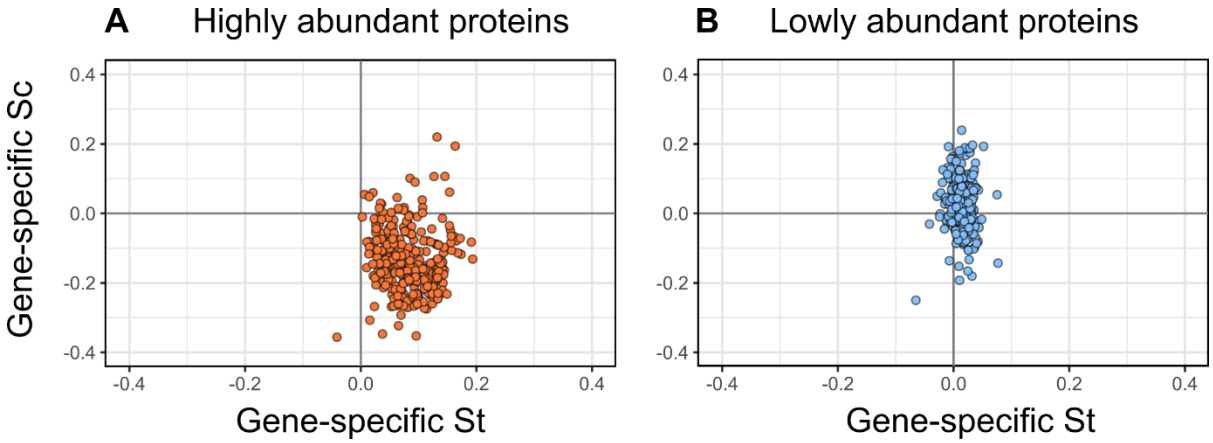


Figure 4. Codon evolutionary selection for translation efficiency (St) and biosynthetic costs (Sc). Higher St values indicate that the gene is strongly selected for translation efficiency, whilst lower Sc values indicate that the gene is strongly selected to reduce the biosynthetic cost of codons. (A) Most Sc values of the CDSs of HAPs are lower than zero, whereas most St values of the CDSs of HAPs are higher than zero, which suggests that they are selected for decreased biosynthetic costs and increased translation efficiency, respectively. (B) CDSs of LAPs did not show any tendency towards any direction.

On the other hand, the LAP sequences did not show any tendency, that is, mean St ≈ 0 and mean Sc ≈ 0. Interestingly, the Akashi test revealed that both groups are subjected to the same selection for translation accuracy (mean odds ratio ≈ 1) (Figure 5). A combination of these results indicate that the codon usage bias is different between the CDSs of LAP and those of HAP.

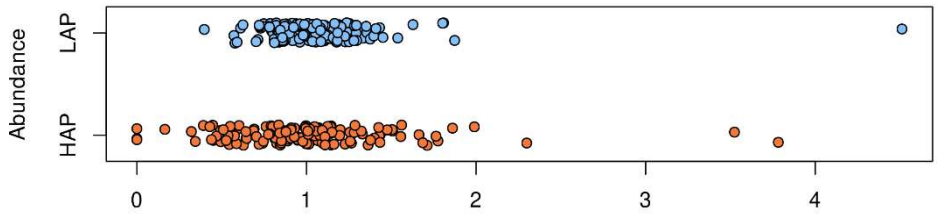


Figure 5. Codon evolutionary selection for translation accuracy. Each dot represents a CDS of HAP or LAP. The CDS of both highly and lowly abundant proteins are centred around an odds ratio of 1, which suggests that both groups are subjected to the same selection for translation accuracy. The odds ratio represents the association between a set of optimal codons and evolutionary constrained sites.

Machine learning models can predict protein abundance

Once the three data sets were compiled, we separately trained predictive models for each data set. We tested each algorithm with support for the regression problems from three different libraries (Scikit-Learn, H2O, and XGBoost) and tested the automated machine learning pipelines (TPOT, H2O, GAMA). By validating with an independent data set, which contained 456 enzymes with measured protein abundance integrated in the ecYeast7 model, we evaluated the MAD and R^2 for all tested algorithms of the three libraries. The results for the five best conventional algorithms are shown in Table 1, and those for all tested algorithms are shown in Table S4.

Regarding the automated machine learning tools, the best predicted pipeline for the all media data set was a stacked ensemble predicted by TPOT. For the minimal medium data set, the best predicted pipeline was also a stacked ensemble predicted by H2O; for the YPD medium data set, it was a gradient boosting machine from the XGBoost library predicted by H2O. We observed that the model performances for the automated pipelines were like those for conventional implemented pipelines (refer to Table 2).

Table 1. Regression evaluation metrics for the five best trained algorithms for each data set. Each algorithm was trained and evaluated by hold-out validation (75% training, 25% validation). In addition, we employed an independent data set for testing. Spearman's ρ and its associated p-value assesses the correlation between the predicted values and the median values obtained by Ho et al. (27).

Algorithm	Coefficient of determination (R^2)	Median absolute deviation (MAD)	Spearman's ρ	p-value
All abundances data set				
AdaBoost (Scikit-Learn)	**0.951**	**0.006**	**0.756**	**4.71E-101**
Random Forest (H2O)	0.899	0.175	0.950	0
Random Forest (Scikit-Learn)	0.843	0.219	0.779	1.03E-220
Bagging Meta-estimator (Scikit-Learn)	0.842	0.188	0.750	9.11E-186
Multilayer perceptron (H2O)	0.834	0.224	0.916	0
Minimal medium data set				
AdaBoost (Scikit-Learn)	**0.801**	**0.183**	**0.763**	**2.16E-100**
Random Forest (H2O)	0.782	0.234	0.905	0
Extremely Randomized Trees (Scikit-Learn)	0.774	0.277	0.777	2.62E-219
Bagging Meta-estimator (Scikit-Learn)	0.772	0.232	0.749	6.44E-188
Gradient Tree Boosting (Scikit-Learn)	0.722	0.370	0.775	1.61E-217
YPD medium data set				
Extremely Randomized Trees (Scikit-Learn)	0.843	0.227	0.740	4.27E-178
Random Forest (H2O)	0.828	0.265	0.917	0
AdaBoost (Scikit-Learn)	0.825	0.290	0.744	4.34E-181
Random Forest (Scikit-Learn)	0.816	0.282	0.743	2.06E-180
Gradient Boosting Estimator (H2O)	0.734	0.380	0.863	9.67E-282

We determined that the AdaBoost estimator from Scikit-Learn, which was implemented with the TPOT-predicted stacked ensemble as a base estimator, was the best predictive model for the data set of all protein abundances and the data set of the minimal medium abundances; it achieves R^2 values of 0.951 and 0.801, respectively (refer to Supplementary Appendix for details). For the YPD medium data set, the gradient boosted tree from the XGBoost library, which was predicted and optimized by the H2O automated tool, was the best model with an R^2 of 0.927 and the lowest MAD for the data set.

Table 2. Regression evaluation metrics for automated machine learning pipelines for each data set. Each algorithm was trained and evaluated by 10-fold cross-validation using an independent data set for testing. Spearman's ρ and its associated p-value assesses the correlation between the predicted values and the median values obtained by Ho et al. (27).

Algorithm	**Coefficient of determination (R^2)**	**Median absolute deviation (MAD)**	**Spearman's ρ**	**p-value**
All abundances data set				
GAMA	0.897	0.042	0.787	2.82E-228
H2O AutoML	0.907	0.120	0.960	0
TPOT	0.908	0.004	0.794	2.33E-235
Minimal medium data set				
GAMA	0.780	0.223	0.745	1.30E-184
H2O AutoML	0.858	0.212	0.931	0
TPOT	0.835	0.238	0.752	3.79E-187
YPD medium data set				
GAMA	0.709	0.308	0.743	6.45E-181
H2O AutoML	**0.927**	**0.021**	**0.960**	**0**
TPOT	0.671	0.355	0.751	1.18E-186

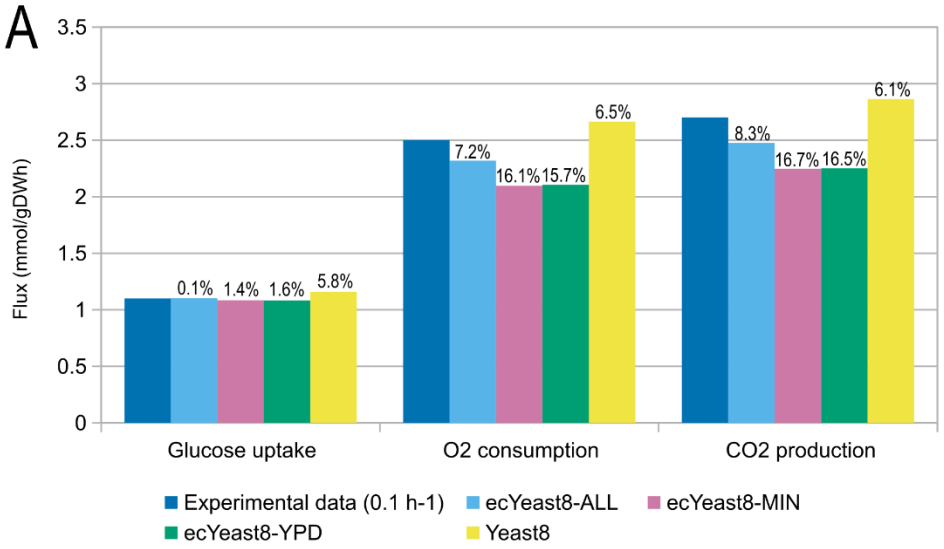
Integration of predictions into enzyme-constrained GEMs

Since we were able to predict the absolute abundance of all 968 enzymes with reasonable accuracy, we decided to incorporate them into the genome-scale metabolic model of *S. cerevisiae*, ecYeast8. We were interested in demonstrating that our predictive model could be useful for reconstructing enzyme-constrained GEMs. For this purpose, we converted the unit from nº of protein molecules/cell (absolute abundance) to mmol/g_{DW}, as required for simulating the metabolism (refer to Material and Methods). Using the GECKO toolbox described by Sánchez et al. (10), we obtained three

different models—ecYeast8-MIN, ecYeast8-YPD and ecYeast8-ALL—using the minimal medium prediction data set, YPD medium prediction data set, and predicted abundances for all the 21 protein measurements in the prediction data set, respectively. Note that the protein pool pseudo-reaction was not included in the models as we incorporated the abundance for all enzymes.

Metabolic fluxes simulated with ML-predicted enzyme abundances are similar to experimental data

The reproduction of the results using the ecYeast7 model and quantitative proteomics data (54) showed that our unit conversion step was correctly performed, as it closely approached the original predictions (Figure S3). After including the predicted protein abundances in ecYeast8, we performed a chemostat simulation to compare the performance of the three models to the conventional GEM Yeast8 and the experimental fluxes quantified by ^{13}C-MFA. When grown aerobically at a dilution rate of 0.1 h^{-1} and limited by glucose concentrations, all models predicted similar fluxes to the experimental values and the conventional model (Figure 6A). Between the three models, ecYeast8-ALL displayed the best predicted values. The predicted fluxes of both ecYeast8-MIN and ecYeast8-YPD were lower than the fluxes predicted by ecYeast8-ALL. We further compared the models by a flux variability analysis. The three ecGEMs had significant reductions in flux variability for most reactions when compared to Yeast8 (Figure 6B).



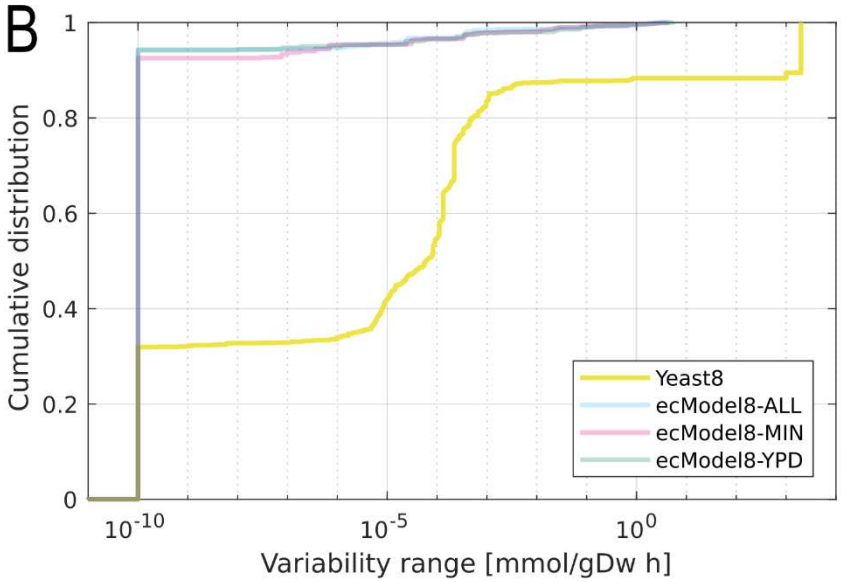


Figure 6. Comparisons between the three ecYeast8 integrated with ML-predicted protein abundances and the traditional GEM Yeast 8. A) Predictions of metabolic flux obtained by the ecYeast8 model integrated with ML-predicted protein abundances and the Yeast8 model (without enzyme constraints), compared to the experimental data. Percentage values represent the relative error when compared to experimental values. B) Flux variability cumulative distribution for Yeast8 and ecYeast8 integrated with ML-predicted protein abundances.

DISCUSSION

In this study, we explored the evolutionary constraints that affect codon usage bias of CDSs of highly and lowly abundant proteins. The difference between both groups of proteins boosted us to evaluate whether this pattern could be recognized by machine learning algorithms and employed to predict protein abundances. Machine learning algorithms remarkably predicted protein abundances from codon usage metrics. Predicted abundances of proteins were integrated in enzyme-constrained genome-scale metabolic models and successfully applied to simulate metabolic fluxes.

Although synonymous codons can code for the same amino acid, comparative analysis of protein-coding sequences has revealed a “preference” of certain codons. This bias in codon frequency is related to the concentration of tRNA molecules with complementary anticodons and their gene copy numbers (citation). We show that RSCU values separate proteins into two groups: HAP and LAP, which shows that codon usage bias varies according to protein abundance levels. The CDSs of HAP have the highest or lowest RSCU values, which underscores a strong codon bias. Otherwise, the CDSs of LAP have no discernible bias. Consistent with our results, Novoa et al. (59) observed that for a large number of amino acids, codon usage completely changes depending on the expression level.

The stochastic nature of cognate tRNA recognition in the ribosome A-site enables evolutionary forces to act by selecting codons that more closely match the abundance of tRNAs in the cell (60). This selection is likely associated with the differences observed in both translation initiation rates and translation elongation rates (20, 61). The enrichment of certain codons positions of a gene is also

determined by natural selection (19), such as the presence of clusters of RareNO codons at the 5' extremity. We tested this finding by quantifying the position-dependent codon usage bias using a binning scheme as detailed by Villada et al (19). We observed that the CDSs of HAP follows the results obtained from the genome-scale analysis, which show a deviation from uniformity most markedly in the first bin, which was not observed in the CDSs of LAP. This finding agrees with the ramp theory, which poses that a "bottleneck" at the beginning of a CDS is necessary to slow ribosomes to prevent jams and collisions (14, 20).

There are selective pressures regarding codon usage that act on the resource allocation for protein biosynthesis (36, 37), translation efficiency (62), and translation accuracy (38). The resource allocation for protein biosynthesis governs the synthesis of RNA molecules, and polypeptides, as well as other molecules involved in the translational process. One way to reduce the overall biosynthetic cost of a protein without altering its amino acid sequence is to reduce the expenditure on RNA synthesis or increase the translation efficiency of existing RNA molecules (37). Genome-scale analysis of 1,320 bacterial genomes performed by Seward & Kelly (37) revealed that genes subjected to strong selection for reducing biosynthetic cost are also subjected to strong selection to increase the translation efficiency.

In this study, we performed translation efficiency and translation accuracy analyses on CDSs of HAP and LAP and observed that the results rely on protein abundance, which was also observed by Seward & Kelly (37). HAPs experience the strongest selection for reducing the biosynthetic cost and increasing the translation efficiency. By increasing the translation efficiency, more proteins could be synthesized with less mRNA. As observed by Ho et al. (27), the function of HAPs seems to be related to processes such as ribosome biogenesis, protein biosynthesis and cell morphogenesis. The fact that these proteins undergo evolutionary selection for increased translation efficiency and decreased biosynthetic cost is consistent with their high demand by the cell at most physiological conditions to maintain the metabolic activity. On the other hand, LAPs do not seem to be affected by the selection of these parameters. These proteins are associated with processes such as DNA replication, DNA repair, RNA processing and cell-cycle regulation, which are necessary in small amounts and therefore might not require resource allocation optimization through natural selection. Therefore, the difference observed between HAP and LAP in terms of CUB and its relation to the biosynthetic cost reinforces our hypothesis that codon metrics can be useful for predicting protein abundance.

Translation accuracy is also important for reducing the biosynthetic resource cost of a protein and improving translation efficiency, as inaccurate translation elongation increases the time needed to sequester a ribosome, which causes a diminished availability of ribosomes and protein misfolding (63). By applying the Akashi test, we observed that CDSs of both HAP and LAP are subjected to the same strength of selection. This finding was also experimentally noted by Yannai et al. (63) in *Escherichia coli* genes.

As we observed that codon usage bias varies depending on the protein abundance, we were interested in evaluating whether machine learning could capture any underlying pattern and predict absolute abundance values using codon usage metrics as features. We were also interested in addressing whether proteome quantifications with different yeast culture media could impact these predictions. We trained regression algorithms with a series of codon usage metrics for three different

data sets and predicted the abundance of proteins that were previously unincorporated in the ecYeast8 model. The data sets differed regarding the type of medium culture employed in each proteomics experiment: we applied the median of the protein measurements obtained for the minimal medium, YPD medium, and combination of both media. After evaluating the regression metrics, we selected the AdaBoost estimator for the data sets of the minimal medium and combination of the minimal and YPD media. The AdaBoost algorithm employs a combination of "weak learners" for the training and predictions; it supports the integration of many other machine learning algorithms as weak learners to improve its performance (64). The TPOT run exported a stacked ensemble of several algorithms; thus, we decided to integrate the predicted pipeline into AdaBoost. We discovered that this approach outperformed all other algorithms and achieved higher R^2 scores and lower MAD values. For the YPD medium data set, however, this approach was bested by another algorithm. The selected algorithm for this medium was the gradient boosted tree from the XGBoost library, which was predicted and optimized by the H2O automated tool.

There have been many attempts to predict protein abundance values, usually for imputation of abundance values of missing proteins (65–68). Mehdi et al. (68) utilized a Bayesian network that depends primarily on protein abundance and mRNA properties, such as interaction with proteins, expression, and folding energy. Even though these studies obtained satisfactory correlation values, many of their features rely on experimental data. The advantage of our approach is that our features depend entirely on intrinsic information contained in gene sequences (i.e., codon usage metrics), which can all be determined *in silico* and still yield reasonable correlation values. An earlier effort by Nie et al. (65) proposed a zero-inflated Poisson-based model that employed microarray data to predict protein abundance of *Desulfovibrio vulgaris*. Torres-García et al. (66) and Li et al. (67) improved on the previous work by predicting the abundance of *D. vulgaris* proteins using gradient boosted trees (GBT) and neural networks, respectively. However, the R^2 for both algorithms was lower than that obtained by our study (refer to Tables 1-2). For the GBT algorithm, the R^2 varied between 0.393 and 0.582, while for the neural networks, it ranged from 0.47 to 0.68. Considering the regression metrics in these comparisons, our proposed prediction models present significant improvement over previous models.

The proposed regression models predicted protein abundances with remarkable accuracy. Thus, we evaluated whether these predictions could be integrated into GEM to simulate the metabolic phenotypes. Comparing the simulated values to the experimental values would be a way to demonstrate the application of our method for constraint-based metabolic modelling. We compared the models based on protein measurements obtained from a minimal (ecYeast8-MIN) medium, rich medium (ecYeast8-YPD) and combination of both media (ecYeast8-ALL). We did not observe differences between ecYeast8-MIN and ecYeast8-YPD. Referring to the results of quantitative proteomics studies, Ho et al. (27) showed that the same medium did not cluster after being subjected to hierarchical clustering or k-means clustering. This finding could explain the similarity between ecYeast8-MIN and ecYeast8-YPD. Notably, the ecYeast8-ALL outperformed the models ecYeast8-MIN and ecYeast8-YPD since its simulations were closer to the experimental values. Consistent with this result, the results obtained using ecYeast8-ALL were similar to the results obtained by Sánchez et al. (10), which were based on experimental protein measurements. Note that the cumulative distribution of the flux variability in the

three ecGEMs showed a variability range lower than that for Yeast8, which highlights their capacity to predict metabolic fluxes mostly by reactions constrained by enzymes. According to our results, Sánchez et al. (10) also showed that the inclusion of enzymatic constraints reduced the flux variability of simulations for the ecYeast7 model when compared to Yeast7. The same result was reported for ecYeast8 (11) and the ecGEM of *B. subtilis* ec_iYO844 (12). Importantly, our predictions also maintained a physiologically relevant solution.

Considering that species from the same domain share similar codon bias signatures (59), the use of our proposed regression models to predict the protein abundance of other species should be feasible. The extensive range of codon usage metrics used to train the machine learning algorithms could capture most patterns that underlie protein abundance in a domain. Since the models were trained using data from *S. cerevisiae*, this approach should work reasonably well for other eukaryotes. It might be possible to create similar predictive models for bacteria using absolute protein abundance data from *Escherichia coli* (69).

Our results underscore that codon usage metrics allow the prediction of protein abundances by machine learning. The observed difference between the CDSs of HAPs and the CDSs of LAPs supports this statement, as both groups sharply contrasted in the performed tests. Considering that codon usage bias is an intrinsic feature of gene sequences, all the metrics employed for compiling the data sets can be determined *in silico*, which simplifies the use of the proposed models for other species and is an advantage over previous attempts, which rely on other experimentally measured data. The machine learning models generated in our study can be a valuable tool for predicting protein abundances for yeasts that do not have large-scale quantitative proteomics available. Taking into account that the integration of protein abundances in GEMs allows improvement in phenotype simulations, our proposed regression models can be useful in system metabolic engineering approaches.

DATA AVAILABILITY

All data and code are available in the GitHub repository:
(https://github.com/LabFisUFV/protein_abundance_prediction)

Protein abundance data was obtained from Ho et al. (27) supplementary material:
(https://www.cell.com/cell-systems/fulltext/S2405-4712(17)30546-X?#supplementaryMaterial)

Quantitative proteomics data for ecYeast7 simulations was obtained from Lahtvee et al. (57) supplementary material:
(https://www.cell.com/cell-systems/fulltext/S2405-4712(17)30088-1#supplementaryMaterial)

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

We thank Professor Leonardo Lopes Bhering (Department of General Biology, UFV) for granting access to the server at his laboratory (Laboratório de Processamento Biométrico (PROBIO)/BIODATA). We are grateful to Professor Marcelo Mendes Brandão (Institute of Biology, UNICAMP) and Dr. Otávio José Bernardes Brustolini (Laboratório Nacional de Computação Científica, Petrópolis) for their critical discussion and comments on this work.

FUNDING

This work was supported by the Brazilian National Council for Scientific and Technological Development (CNPq) [grant number 148661/2018-1]; the Foundation for Research Support of the State of Minas Gerais (FAPEMIG); and the Coordination for the Improvement of Higher Education Personnel (CAPES) - Finance Code 001.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Hui,S., Silverman,J.M., Chen,S.S., Erickson,D.W., Basan,M., Wang,J., Hwa,T. and Williamson,J.R. (2015) Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. *Mol. Syst. Biol.*, **11**, 784.
2. Lerman,J.A., Hyduke,D.R., Latif,H., Portnoy,V.A., Lewis,N.E., Orth,J.D., Schrimpe-Rutledge,A.C., Smith,R.D., Adkins,J.N., Zengler,K., *et al.* (2012) In silico method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.*, **3**.
3. Vitrinel,B., Koh,H.W.L., Mujgan Kar,F., Maity,S., Rendleman,J., Choi,H. and Vogel,C. (2019) Exploiting Interdata Relationships in Next-generation Proteomics Analysis. *Mol. Cell. Proteomics*, **18**, S5–S14.
4. Pappireddi,N., Martin,L. and Wühr,M. (2019) A Review on Quantitative Multiplexed Proteomics. *ChemBioChem*, **20**, 1210–1224.
5. Otto,A., Becher,D. and Schmidt,F. (2014) Quantitative proteomics in the field of microbiology. *Proteomics*, **14**, 547–565.
6. Swiatly,A., Plewa,S., Matysiak,J. and Kokot,Z.J. (2018) Mass spectrometry-based proteomics techniques and their application in ovarian cancer research. *J. Ovarian Res.*, **11**, 1–13.
7. Williams,T.D., Turan,N., Diab,A.M., Wu,H., Mackenzie,C., Bartie,K.L., Hrydziuszko,O., Lyons,B.P., Stentiford,G.D., Herbert,J.M., *et al.* (2011) Towards a system level understanding of non-model organisms sampled from the environment: A network biology approach. *PLoS*

Comput. Biol., **7**.

8. Yang,L., Yurkovich,J.T., King,Z.A. and Palsson,B.O. (2018) Modeling the multi-scale mechanisms of macromolecular resource allocation. *Curr. Opin. Microbiol.*, **45**, 8–15.
9. King,Z.A., Lu,J., Dräger,A., Miller,P., Federowicz,S., Lerman,J.A., Ebrahim,A., Palsson,B.O. and Lewis,N.E. (2015) BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.*, **44**, D515–D522.
10. Sánchez,B.J., Zhang,C., Nilsson,A., Lahtvee,P., 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.
11. 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**.
12. Massaiu,I., Pasotti,L., Sonnenschein,N., Rama,E., Cavaletti,M., Magni,P., Calvio,C. and Herrgård,M.J. (2019) Integration of enzymatic data in Bacillus subtilis genome-scale metabolic model improves phenotype predictions and enables in silico design of poly-γ-glutamic acid production strains. *Microb. Cell Fact.*, **18**, 1–20.
13. Tian,M. and Reed,J.L. (2018) Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis. *Bioinformatics*, **34**, 3882–3888.
14. Shah,P., Ding,Y., Niemczyk,M., Kudla,G. and Plotkin,J.B. (2013) Rate-limiting steps in yeast protein translation. *Cell*, **153**, 1589.
15. Sharp,P.M. and Li,W.H. (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.*, **24**, 28–38.
16. Zhou,Z., Dang,Y., Zhou,M., Li,L., Yu,C., Fu,J., Chen,S. and Liu,Y. (2016) Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc. Natl. Acad. Sci.*, **113**, E6117–E6125.
17. Hanson,G. and Coller,J. (2017) Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol. Cell Biol.*, 10.1038/nrm.2017.91.
18. Sharp,P.M. and Li,W.-H. (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
19. Villada,J.C., Brustolini,O.J.B. and Silveira,W.B. da (2017) Integrated analysis of individual codon contribution to protein biosynthesis reveals a new approach to improving the basis of rational gene design. *DNA Res.*, **0**, 1–16.
20. Verma,M., Choi,J., Cottrell,K.A., Lavagnino,Z., Thomas,E.N., Pavlovic-Djuranovic,S., Szczesny,P., Piston,D.W., Zaher,H.S., Puglisi,J.D., *et al.* (2019) A short translational ramp determines the efficiency of protein synthesis. *Nat. Commun.*, **10**, 1–15.
21. Quax,T.E.F., Claassens,N.J., Soll,D. and van der Oost,J. (2015) Codon Bias as a Means to Fine-Tune Gene Expression. *Mol. Cell*, **59**, 149–161.
22. Laurent,J.M., Vogel,C., Kwon,T., Craig,S.A., Boutz,D.R., Huse,H.K., Nozue,K., Walia,H., Whiteley,M., Ronald,P.C., *et al.* (2010) Protein abundances are more conserved than mRNA

abundances across diverse taxa. *Proteomics*, **10**, 4209–4212.

23. 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. *Nat. Commun.*, **9**.
24. Costello,Z. and Martin,H.G. (2018) A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *npj Syst. Biol. Appl.*, **4**, 19.
25. Medlock,G.L. and Papin,J.A. (2020) Guiding the Refinement of Biochemical Knowledgebases with Ensembles of Metabolic Networks and Machine Learning. *Cell Syst.*, **10**, 109-119.e3.
26. Andreozzi,S., Miskovic,L. and Hatzimanikatis,V. (2016) ISCHRUNK - In Silico Approach to Characterization and Reduction of Uncertainty in the Kinetic Models of Genome-scale Metabolic Networks. *Metab. Eng.*, **33**, 158–168.
27. Ho,B., Baryshnikova,A. and Brown,G.W. (2018) Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Syst.*, **6**, 192-205.e3.
28. Howe,K.L., Contreras-Moreira,B., De Silva,N., Maslen,G., Akanni,W., Allen,J., Alvarez-Jarreta,J., Barba,M., Bolser,D.M., Cambell,L., *et al.* (2019) Ensembl Genomes 2020—enabling non-vertebrate genomic research. *Nucleic Acids Res.*, **48**, D689–D695.
29. Kinsella,R.J., Kähäri,A., Haider,S., Zamora,J., Proctor,G., Spudich,G., Almeida-King,J., Staines,D., Derwent,P., Kerhornou,A., *et al.* (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, **2011**.
30. Consortium,T.U. (2018) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
31. Chan,P.P. and Lowe,T.M. (2008) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.*, **37**, D93–D97.
32. Chan,P.P. and Lowe,T.M. (2015) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.*, **44**, D184–D189.
33. Hockenberry,A.J., Sirer,M.I., Amaral,L.A.N. and Jewett,M.C. (2014) Quantifying position-dependent codon usage bias. *Mol. Biol. Evol.*, **31**, 1880–1893.
34. Nasrullah,I., Butt,A.M., Tahir,S., Idrees,M. and Tong,Y. (2015) Genomic analysis of codon usage shows influence of mutation pressure, natural selection, and host features on Marburg virus evolution. *BMC Evol. Biol.*, **15**.
35. Demšar,J., Curk,T., Erjavec,A., Gorup,Č., Hočevar,T., Milutinovič,M., Možina,M., Polajnar,M., Toplak,M., Starič,A., *et al.* (2013) Orange: Data mining toolbox in python. *J. Mach. Learn. Res.*, **14**, 2349–2353.
36. Seward,E.A. and Kelly,S. (2016) Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. *Genome Biol.*, **17**, 226.
37. Seward,E.A. and Kelly,S. (2018) Selection-driven cost-efficiency optimization of transcripts modulates gene evolutionary rate in bacteria. *Genome Biol.*, **19**, 102.
38. Akashi,H. (1994) Synonymous Codon Usage. *Genet. Soc. Am.*, **136**, 927–935.
39. Camiolo,S., Melito,S., Milia,G. and Porceddu,A. (2014) Seforta, an integrated tool for detecting the signature of selection in coding sequences. *BMC Res. Notes*, **7**, 2–4.

40. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–7.
41. Peden,J.F. (2000) Analysis of Codon Usage.
42. Puigbò,P., Bravo,I.G. and Garcia-Vallve,S. (2008) CAIcal: A combined set of tools to assess codon usage adaptation. *Biol. Direct*, **3**, 38.
43. Elek,A., Kuzman,M. and Vlahovicek,K. (2019) coRdon: Codon Usage Analysis and Prediction of Gene Expressivity.
44. Sabi,R., Volvovitch Daniel,R. and Tuller,T. (2016) stAIcalc : tRNA adaptation index calculator based on species-specific weights. *Bioinformatics*, **33**, btw647.
45. Liu,S.S., Hockenberry,A.J., Jewett,M.C. and Amaral,L.A.N. (2018) A novel framework for evaluating the performance of codon usage bias metrics. *J. R. Soc. Interface*, **15**, 20170667.
46. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V., *et al.* (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
47. H2O.ai (2018) Python Interface for H2O, Python module version 3.10.0.8.
48. Chen,T. and Guestrin,C. (2016) XGBoost: A scalable tree boosting system. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, **13-17-Augu**, 785–794.
49. Olson,R.S. and Moore,J.H. (2019) TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. In Hutter,F., Kotthoff,L., Vanschoren,J. (eds), *Automated Machine Learning: Methods, Systems, Challenges*. Springer International Publishing, Cham, pp. 151–160.
50. Gijsbers,P. and Vanschoren,J. (2019) GAMA: Genetic Automated Machine learning Assistant. *J. Open Source Softw.*, **4**, 1132.
51. Van Rossum,G. and Drake,F. (2009) Python 3.
52. Yap,P.Y. and Trau,D. (2019) Direct yeast cell count at OD600. *Tip Biosyst.*
53. Li,E. and Mira de Orduña,R. (2010) A rapid method for the determination of microbial biomass by dry weight using a moisture analyser with an infrared heating source and an analytical balance. *Lett. Appl. Microbiol.*, **50**, 283–288.
54. Lahtvee,P.J., Sánchez,B.J., Smialowska,A., Kasvandik,S., Elsemman,I.E., Gatto,F. and Nielsen,J. (2017) Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. *Cell Syst.*, **4**, 495-504.e5.
55. Jouhten,P., Rintala,E., Huuskonen,A., Tamminen,A., Toivari,M., Wiebe,M., Ruohonen,L., Penttilä,M. and Maaheimo,H. (2008) Oxygen dependence of metabolic fluxes and energy generation of Saccharomyces cerevisiae CEN.PK113-1A. *BMC Syst. Biol.*, **2**.
56. Wang,H., Marcišauskas,S., Sánchez,B.J., Domenzain,I., Hermansson,D., Agren,R., Nielsen,J. and Kerkhoven,E.J. (2018) RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. *PLOS Comput. Biol.*, **14**, e1006541.
57. Gurobi Optimization,L. (2020) Gurobi Optimizer Reference Manual.
58. Almagro Armenteros,J.J., Tsirigos,K.D., Sønderby,C.K., Petersen,T.N., Winther,O., Brunak,S., von Heijne,G. and Nielsen,H. (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.*, **37**, 420–423.

59. Novoa,E.M., Jungreis,I., Jaillon,O., Kellis,M. and Leitner,T. (2019) Elucidation of Codon Usage Signatures across the Domains of Life. *Mol. Biol. Evol.*, **36**, 2328–2339.
60. Hanson,G. and Coller,J. (2018) Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol. Cell Biol.*, **19**, 20–30.
61. Novoa,E.M. and Ribas de Pouplana,L. (2012) Speeding with control: Codon usage, tRNAs, and ribosomes. *Trends Genet.*, **28**, 574–581.
62. Gingold,H. and Pilpel,Y. (2011) Determinants of translation efficiency and accuracy. *Mol. Syst. Biol.*, **7**, 1–13.
63. Yannai,A., Katz,S. and Hershberg,R. (2018) The codon usage of lowly expressed genes is subject to natural selection. *Genome Biol. Evol.*, **10**, 1237–1246.
64. Freund,Y. and Schapire,R.E. (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.*, **55**, 119–139.
65. Nie,L., Wu,G., Culley,D.E., Scholten,J.C.M. and Zhang,W. (2007) Integrative analysis of transcriptomic and proteomic data: Challenges, solutions and applications. *Crit. Rev. Biotechnol.*, **27**, 63–75.
66. Torres-García,W., Zhang,W., Runger,G.C., Johnson,R.H. and Meldrum,D.R. (2009) Integrative analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: a non-linear model to predict abundance of undetected proteins. *Bioinformatics*, **25**, 1905–1914.
67. Li,F., Nie,L., Wu,G., Qiao,J. and Zhang,W. (2011) Prediction and Characterization of Missing Proteomic Data in *Desulfovibrio vulgaris*. *Comp. Funct. Genomics*, **2011**, 780973.
68. Mehdi,A.M., Patrick,R., Bailey,T.L. and Boden,M. (2014) Predicting the dynamics of protein abundance. *Mol. Cell. Proteomics*, **13**, 1330–1340.
69. Schmidt,A., Kochanowski,K., Vedelaar,S., Ahrné,E., Volkmer,B., Callipo,L., Knoops,K., Bauer,M., Aebersold,R. and Heinemann,M. (2016) The quantitative and condition-dependent Escherichia coli proteome. *Nat. Biotechnol.*, **34**, 104–110.

TABLE AND FIGURES LEGENDS

Figure 1. Machine learning applied to codon usage bias for the prediction of protein abundance. We calculated a series of codon usage metrics and nucleotide composition numbers. Three separate data sets were compiled and employed for the training: the first data set contains protein measurements from yeast cultivation in the minimal medium, the second data set consists of protein measurements from yeast cultivation in the YPD medium, and the third data set combines protein measurements for both culture media. The hexagonal boxes denote the input data; the rectangular boxes indicate the intermediate steps; the ellipsoids represent the trained regression models; and the parallelogram denotes the output from these models.

Figure 2: Heatmap of relative synonymous codon usage that illustrates the difference in codon usage in regard to frequency between the 20 most abundant and 20 least abundant proteins of *S. cerevisiae*. Columns denote the 59 codons with one or more synonymous codons. (A) Rows indicate CDSs of the

20 most abundant proteins. (B) Rows represent CDSs of the 20 least abundant proteins. CDSs of the HAP, in contrast to the CDSs of LAP, have many enriched (high RSCU) or depleted (low RSCU) codons. Sequences marked with an asterisk denote the presence of a signal peptide at the 5' extremity as detected by SignalP 5 (58).

Figure 3: Evolutionary selection for position-dependent codon usage bias determined by the chi-squared test for the matrix constructed with the CodG package. Each CDS was equally divided into 10 bins to evaluate how each position contributes to overall codon usage. (A) Deviation from uniformity in the CDSs of HAPs shows bias towards 5' end. (B) CDSs of lowly abundant proteins show higher uniformity.

Figure 4: Codon evolutionary selection for translation efficiency (St) and biosynthetic costs (Sc). Higher St values indicate that the gene is strongly selected for translation efficiency, whilst lower Sc values indicate that the gene is strongly selected to reduce the biosynthetic cost of codons. (A) Most Sc values of the CDSs of HAPs are lower than zero, whereas most St values of the CDSs of HAPs are higher than zero, which suggests that they are selected for decreased biosynthetic costs and increased translation efficiency, respectively. (B) CDSs of LAPs did not show any tendency towards any direction.

Figure 5: Codon evolutionary selection for translation accuracy. Each dot represents a CDS of HAP or LAP. The CDS of both highly and lowly abundant proteins are centred around an odds ratio of 1, which suggests that both groups are subjected to the same selection for translation accuracy. The odds ratio represents the association between a set of optimal codons and evolutionary constrained sites.

Figure 6: Comparisons between the three ecYeast8 integrated with ML-predicted protein abundances and the traditional GEM Yeast 8. A) Predictions of metabolic flux obtained by the ecYeast8 model integrated with ML-predicted protein abundances and the Yeast8 model (without enzyme constraints), compared to the experimental data. Percentage values represent the relative error when compared to experimental values. B) Flux variability cumulative distribution for Yeast8 and ecYeast8 integrated with ML-predicted protein abundances.

Table 1: Regression evaluation metrics for the five best trained algorithms for each data set. Each algorithm was trained and evaluated by hold-out validation (75% training, 25% validation). In addition, we employed an independent data set for testing. Spearman's ρ and its associated p-value assesses the correlation between the predicted values and the median values obtained by Ho et al. (27).

Table 2: Regression evaluation metrics for automated machine learning pipelines for each data set. Each algorithm was trained and evaluated by 10-fold cross-validation using an independent data set for testing. Spearman's ρ and its associated p-value assesses the correlation between the predicted values and the median values obtained by Ho et al. (27).

GENERAL CONCLUSIONS

Our results underscore that codon usage metrics allow the prediction of protein abundances by machine learning. The observed difference between the CDSs of highly abundant proteins and the CDSs of lowly abundant proteins supports this statement, as both groups sharply contrasted in the performed tests. Considering that codon usage bias is an intrinsic feature of gene sequences, all the metrics employed for compiling the data sets can be determined *in silico*, which simplifies the use of the proposed models for other species and is an advantage over previous attempts, which rely on other experimentally measured data. The extensive range of codon usage metrics used to train the machine learning algorithms could capture most patterns that underlie protein abundance in a domain, as species from the same domain share similar codon bias signatures. Since the models were trained using data from *S. cerevisiae*, the use of our proposed regression models can be capable of predicting the protein abundance of other eukaryotes. Taking into account that the integration of protein abundances in GEMs allows improvement in phenotype simulations, our proposed regression models can be useful in system metabolic engineering approaches.

APPENDIX A – Supplementary appendix

Supplementary appendix written following the Nucleic Acids Research journal author guidelines.

SUPPLEMENTARY APPENDIX

Protein Abundance Prediction Through Machine Learning Methods

Mauricio Ferreira[1], Rafaela Ventorim[1], Eduardo Almeida[1], Sabrina Silveira[2], Wendel Silveira[1,*]

[1] Department of Microbiology, Universidade Federal de Viçosa, Viçosa, Minas Gerais, 36570-900, Brazil
[2] Department of Computer Science, Universidade Federal de Viçosa, Viçosa, Minas Gerais, 36570-900, Brazil

* To whom correspondence should be addressed. Tel: +55 31 3612-2431; Email: wendel.silveira@ufv.br

Present Address: Wendel Silveira, Department of Microbiology, Universidade Federal de Viçosa, Viçosa, Minas Gerais, 36570-900, Brazil

SUPPLEMENTARY INFORMATION

Expanded description of selected machine learning models

We discovered that the AdaBoost estimator from Scikit-Learn, which was implemented with the TPOT-predicted stacked ensemble as a base estimator, was the best predictive model for the data set of all protein abundances and the data set of minimal medium abundances. The adaptive boosting algorithm (AdaBoost) employs a combination of regressors referred to as “weak learners” to generate a larger regressor, which is better than any single regressor. This combination of weak learners works by repeatedly training with different distributions of the input data set and combining their outputs. In the case of regression, the outputs are combined by a weighted average of median (1). While the weak learners are traditionally decision trees, the Scikit-Learn library supports the integration of other machine learning algorithms as weak learners to improve its performance, with the parameter referred to as the “base estimator”. The TPOT run exported a stacked ensemble of several algorithms (refer to Table S5). Thus, we decided to integrate the predicted pipeline into AdaBoost. We noticed that it outperformed all other algorithms and achieved higher R^2 scores and lower MAD values.

For the YPD medium data set, the best predictive model was the extreme gradient boosting algorithm from the XGBoost library, which was predicted and optimized by the H2O automated tool. XGBoost integrates multiple trees into a stronger learner, such as AdaBoost and other boosting algorithms. However, XGBoost has better performance than other algorithms as it is capable of running in parallel; does not need transformation of numerical, continuous data; and minimizes overfitting by implementing regularization procedures (2).

SUPPLEMENTARY TABLES

Table S1: List of codon usage metrics employed as features for constructing the training data sets.

Codon usage metrics	**Reference**
Information theory-based codon usage bias (iCUB)	(3)
tRNA adaptation index (tAI)	(4)
Codon adaptation index (CAI)	(5)
Codon bias index (CBI)	(6)
Frequency of optimal codons (Fop)	(7)
Effective number of codons (ENC)	(8)
ENC alternative implementation (ENC')	(9)
G+C content of gene	(10)
G+C of 3rd codon position	(10)
Base composition at silent sites	(10)

Hydropathicity of protein	(10)
Aromaticity of protein	(10)
B measure of codon bias	(11)
E measure of expression	(12)
Maximum likelihood codon bias (MCB)	(13)
Measure independent of length and composition (MILC)	(14)
MILC-based expression level predictor (MELP)	(14)
Synonymous codon usage orderliness (SCUO)	(15)
Gene codon bias (GCB)	(16)
Evolutionary selection pressure on nucleotide biosynthetic cost (Sc)	(17, 18)

Evolutionary selection pressure on gene translation efficiency (St)	(17, 18)
Nucleotide composition	(19)

Table S2: List of features compiled for the training data sets using codon usage metrics calculated individually for gene and nucleotide composition numbers.

iCUB	tAI	T3s	C3s	A3s	G3s	CAI_CODONW	CBI
Fop_CODONW	Nc	GC3s	GC	L_sym	L_aa	Gravy	Aromo
CAI_EMBOSS	CAI_coRdon	MELP	E	GCB	Fop_coRdon	MILC_SELF	MILC_REF
B_SELF	B_REF	MCB_SELF	MCB_REF	ENC	ENC_prime_SELF	ENC_prime_REF	SCUO
Sc	St	A	C	T	G	%A	%C
%T	%G	%G+C	%G+A	%G+T	%A+T	%A+C	%C+T
A1	C1	T1	G1	%A1	%C1	%T1	%G1
%G1+C1	%G1+A1	%G1+T1	%A1+T1	%A1+C1	%C1+T1	A2	C2
T2	G2	%A2	%C2	%T2	%G2	%G2+C2	%G2+A2
%G2+T2	%A2+T2	%A2+C2	%C2+T2	A3	C3	T3	G3
%A3	%C3	%T3	%G3	%G3+C3	%G3+A3	%G3+T3	%A3+T3
%A3+C3	%C3+T3	%G3s+C3s					

Table S3: Example of how data sets were structured. The first column is the systematic name of all open reading frames (ORFs). The second column contains the protein abundance values for each ORF. From the third column to the last column, all columns are codon usage metrics or nucleotide composition numbers. A total of 91 columns are present in the data sets. The protein abundance values consist of the median values from several different quantitative proteomics analyses and are expressed as the number of molecules per cell.

ORF	Protein abundance	iCUB	tAI	CAI (EMBOSS)	Fop	Sc	St	...	%G3s+ C3s
Q0045	2440	20	0.263	0.584	0.599	0.05	0.04	...	11.3
Q0050	353	20	0.227	0.561	0.683	0.03	-0.01	...	9.9
Q0055	271	20	0.242	0.601	0.679	-0.01	0.02	...	13.3
Q0060	1029	20	0.199	0.587	0.810	-0.17	-0.01	...	3.7
Q0065	127	20	0.227	0.573	0.771	-0.02	0.02	...	6.2
Q0085	1076	20	0.249	0.633	0.766	0.02	-0.01	...	7.3
Q0115	183	25	0.196	0.600	0.805	-0.01	0.06	...	3.8
...	...	...	...	...	...	...	...	...	...
YPR204W	370	37	0.286	0.435	0.267	0.05	0.02	...	47.9

Table S4: Regression evaluation metrics for all tested algorithms for each data set. Each algorithm was trained and evaluated by hold-out validation (75% training, 25% validation) using an independent data set for testing. Spearman's ρ and its associated p-value assesses the correlation between the predicted values and median values obtained by Ho et al. (20).

Algorithm	**Coefficient of determination (R^2)**	**Median absolute deviation (MAD)**	**Spearman's ρ**	**p-value**
All abundances data-set				
AdaBoost (Scikit-Learn)	0.951	0.006	0.756	4.71E-101
Random Forest (H2O)	0.899	0.175	0.950	0
Random Forest (Scikit-Learn)	0.843	0.219	0.779	1.03E-220
Bagging Meta-estimator (Scikit-Learn)	0.842	0.188	0.750	9.11E-186
Multilayer perceptron (H2O)	0.834	0.224	0.916	0
Extremely Randomized Trees (Scikit-Learn)	0.775	0.277	0.778	2.62E-219
Gradient Boosting Estimator (H2O)	0.765	0.328	0.884	0
Gradient Tree Boosting (Scikit-Learn)	0.702	0.404	0.748	5.10E-184
Ridge Regression (Scikit-Learn)	0.643	0.435	0.769	1.64E-211
Linear Regression (Scikit-Learn)	0.642	0.441	0.770	1.53E-212
Bayesian Ridge (Scikit-Learn)	0.637	0.435	0.767	1.43E-209
Theil-Sen (Scikit-Learn)	0.633	0.455	0.765	1.43E-207
Orthogonal Matching Pursuit (Scikit-Learn)	0.631	0.439	0.753	4.45E-198
Elastic Net	0.626	0.456	0.753	1.54E-197

(Scikit-Learn)				
Generalized Linear Model (H2O)	0.625	0.453	0.801	9.36E-212
Huber (Scikit-Learn)	0.611	0.467	0.751	4.01E-196
Lasso (Scikit-Learn)	0.555	0.503	0.464	1.12E-58
Nearest Neighbors (Scikit-Learn)	0.487	0.544	0.521	5.77E-76
Support Vector Regressor (Scikit-Learn)	0.414	0.100	0.298	1.51E-23
Lasso Lars (Scikit-Learn)	0.020	0.815	0.732	6.92E-182
Passive Agressive (Scikit-Learn)	-5.474	2.318	0.188	4.58E-10
Gaussian Process (Scikit-Learn)	-11.518	1.389	0.114	4.39E-05
XGBoost	-33.300	7.530	0.703	2.24E-161
Decision Tree (Scikit-Learn)	-36.518	7.880	0.595	6.80E-99
Minimal medium data-set				
AdaBoost (Scikit-Learn)	**0.801**	**0.183**	**0.763**	**2.16E-100**
Random Forest (H2O)	0.782	0.234	0.905	0
Extremely Randomized Trees (Scikit-Learn)	0.774	0.277	0.777	2.62E-219
Bagging Meta-estimator (Scikit-Learn)	0.772	0.232	0.749	6.44E-188
Gradient Tree Boosting (Scikit-Learn)	0.722	0.370	0.775	1.61E-217
Gradient Boosting Estimator (H2O)	0.716	0.329	0.856	8.69E-272

Multilayer perceptron (H2O)	0.678	0.370	0.829	7.20E-240
Random Forest (Scikit-Learn)	0.645	0.386	0.750	1.16E-195
Ridge Regression (Scikit-Learn)	0.643	0.435	0.769	1.64E-211
Linear Regression (Scikit-Learn)	0.642	0.441	0.770	1.53E-212
Bayesian Ridge (Scikit-Learn)	0.637	0.435	0.767	1.43E-209
Theil-Sen (Scikit-Learn)	0.633	0.455	0.765	1.43E-207
Orthogonal Matching Pursuit (Scikit-Learn)	0.631	0.439	0.753	4.45E-198
Elastic Net (Scikit-Learn)	0.626	0.456	0.753	1.54E-197
Generalized Linear Model (H2O)	0.625	0.412	0.803	2.84E-213
Huber (Scikit-Learn)	0.611	0.467	0.751	4.01E-196
Lasso (Scikit-Learn)	0.555	0.503	0.732	6.92E-182
Nearest Neighbors (Scikit-Learn)	0.487	0.544	0.521	5.77E-76
Passive Aggressive (Scikit-Learn)	0.384	0.731	0.710	4.63E-166
Lasso Lars (Scikit-Learn)	0.020	0.815	0.464	1.12E-58
XGBoost	-0.078	4530.561	0.778	1.12E-219
Gaussian Process (Scikit-Learn)	-9.227	0.000	0.082	0.002612580833657
Decision Tree (Scikit-Learn)	-37.463	8.007	0.616	2.32E-109
YPD medium data-set				
Extremely Randomized	0.843	0.227	0.740	4.27E-178

Trees (Scikit-Learn)				
Random Forest (H2O)	0.828	0.265	0.917	0
AdaBoost (Scikit-Learn)	0.825	0.290	0.744	4.34E-181
Random Forest (Scikit-Learn)	0.816	0.282	0.743	2.06E-180
Gradient Boosting Estimator (H2O)	0.734	0.380	0.863	9.67E-282
Multilayer perceptron (H2O)	0.699	0.341	0.872	8.13E-294
Gradient Tree Boosting (Scikit-Learn)	0.699	0.407	0.748	3.55E-184
Ridge Regression (Scikit-Learn)	0.635	0.435	0.738	1.12E-176
Orthogonal Matching Pursuit (Scikit-Learn)	0.631	0.442	0.736	6.38E-175
Linear Regression (Scikit-Learn)	0.631	0.437	0.736	6.76E-175
Elastic Net (Scikit-Learn)	0.627	0.460	0.730	9.90E-171
Bayesian Ridge (Scikit-Learn)	0.626	0.456	0.734	6.65E-174
Generalized Linear Model (H2O)	0.624	0.465	0.802	1.01E-212
Huber (Scikit-Learn)	0.611	0.453	0.718	7.31E-163
Lasso (Scikit-Learn)	0.571	0.517	0.708	2.40E-156
Decision Tree (Scikit-Learn)	0.564	0.352	0.620	1.12E-109
Nearest Neighbors (Scikit-Learn)	0.490	0.553	0.499	1.19E-65
Support Vector Regressor (Scikit-Learn)	0.399	0.367	0.286	1.18E-20

Passive Aggressive (Scikit-Learn)	-0.723	0.904	0.389	2.15E-38
Bagging Meta-estimator (Scikit-Learn)	-0.795	1.834	0.756	4.39E-190
Gaussian Process (Scikit-Learn)	-11.518	1.389	0.114	4.39E-05
XGBoost (Scikit-Learn)	-33.301	7.532	0.690	1.88E-145

Table S5: List of algorithms that are part of the stacked ensemble applied as weak learners for the AdaBoost estimator. This list was generated as part of the TPOT automated pipeline prediction. The algorithms are utilized in this order.

Algorithm	**Scikit-Learn/XGBoost function**
Cross-validated Lasso	LassoLarsCV
Linear Support Vector Regression	LinearSVR
Ridge regression with built-in cross-validation	RidgeCV
Linear Support Vector Regression	LinearSVR
Extremely Randomized Trees Regressor	ExtraTreesRegressor
Ridge regression with built-in cross-validation	RidgeCV
Linear Support Vector Regression	LinearSVR
Ridge regression with built-in cross-validation	RidgeCV
Stochastic Gradient Descent	SGDRegressor
Linear Support Vector Regression	LinearSVR
Elastic Net model with iterative fitting along a regularization path	ElasticNetCV
Extreme Gradient Boosting	XGBRegressor
Extremely Randomized Trees Regressor	ExtraTreesRegressor
Linear Support Vector Regression	LinearSVR
Random Forest Regressor	RandomForestRegressor

SUPPLEMENTARY FIGURES

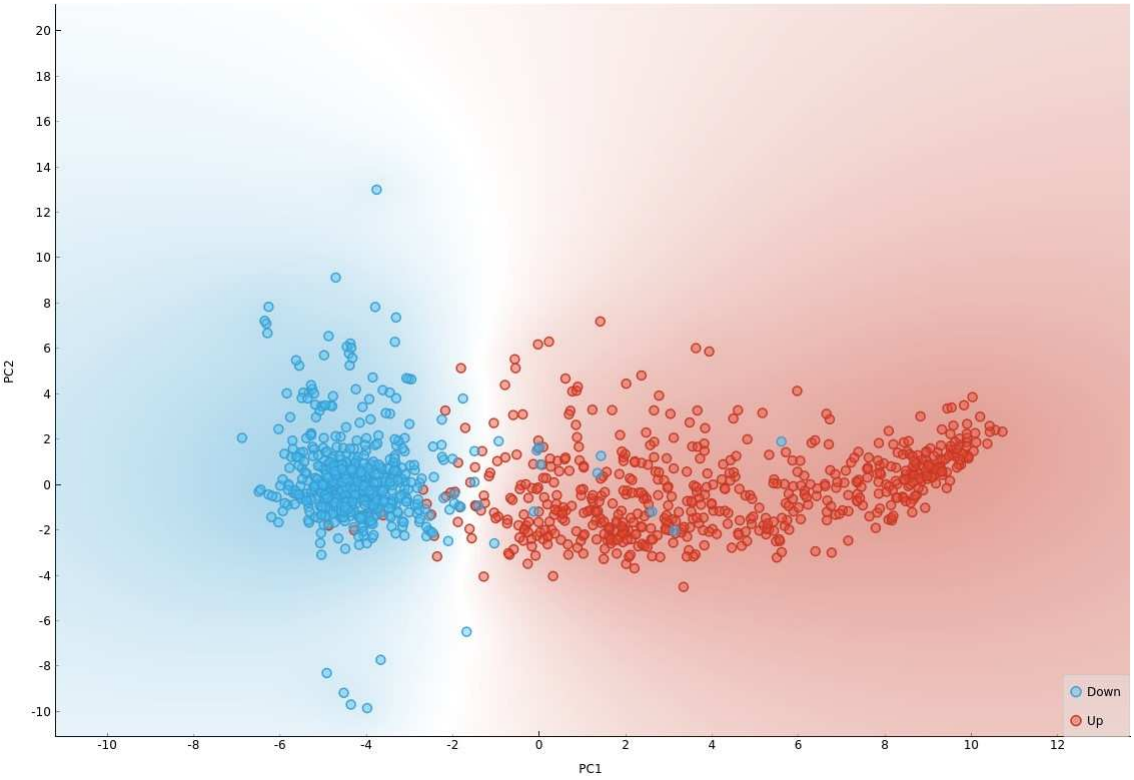


Figure S1: Principal component analysis of RSCU values calculated from CDSs of highly abundant proteins (HAP) and lowly abundant proteins (LAP). Two distinct groups of CDSs could be observed. The first group is composed of mostly HAP, and the second group is composed of mostly LAP.

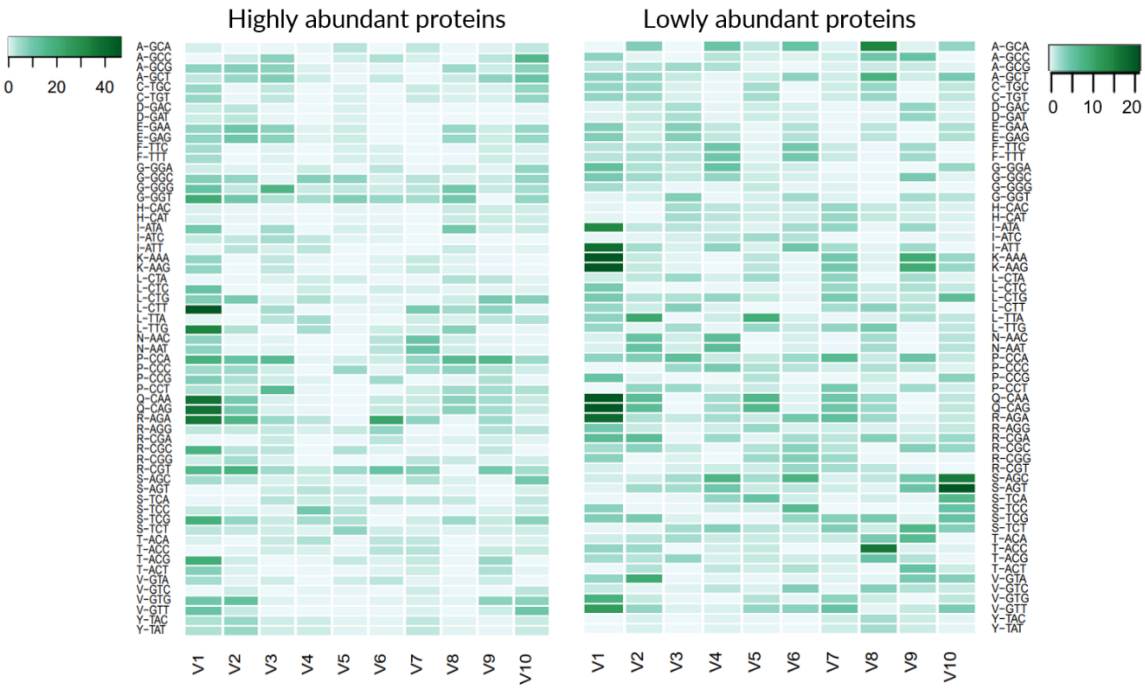


Figure S2: Evolutionary selection for position-dependent codon usage bias as determined by the chi-squared test on the matrix constructed with the CodG package. Each CDS was equally divided into 10 bins to evaluate how each position contributes to the overall codon usage. A) Deviation from uniformity in the CDSs of highly abundant proteins show bias towards 5' end. B) The CDSs of lowly abundant proteins show higher uniformity.

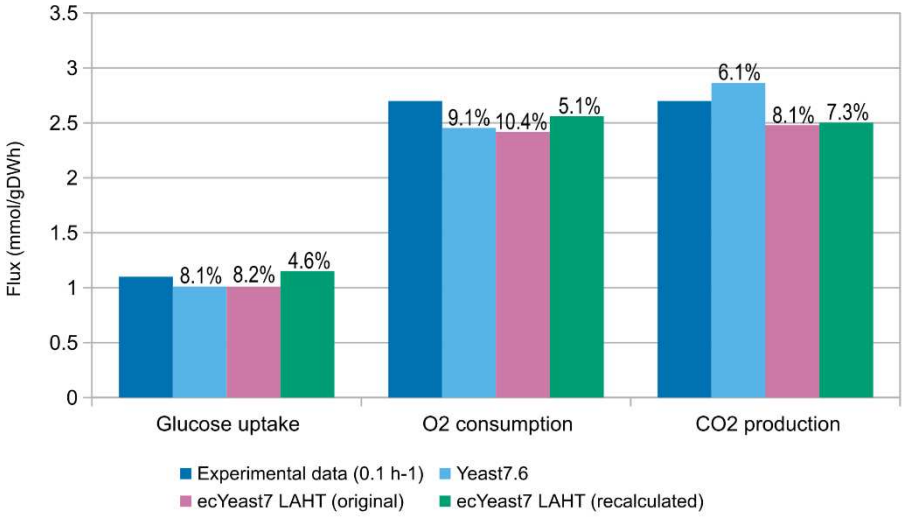


Figure S3: Predictions of metabolic flux obtained by the Yeast7 model and ecYeast7. We attempted to validate our unit conversion step by replicating the analysis performed by Sánchez et al. (2017). The “original” ecYeast7 LAHT model employs the quantitative proteomics data from Lahtvee et al. (2016). The “recalculated” ecYeast7 LAHT model applies the median absolute values from Ho et al. (2018). Percentage values represent the relative error when compared to experimental values

SUPPLEMENTARY REFERENCES

1. Freund,Y. and Schapire,R.E. (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.*, **55**, 119–139.
2. Chen,T. and Guestrin,C. (2016) XGBoost: A scalable tree boosting system. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, **13-17-Augu**, 785–794.
3. Liu,S.S., Hockenberry,A.J., Jewett,M.C. and Amaral,L.A.N. (2018) A novel framework for evaluating the performance of codon usage bias metrics. *J. R. Soc. Interface*, **15**, 20170667.
4. Reis,M. d., Savva,R. and Wernisch,L. (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.*, **32**, 5036–5044.
5. Sharp,P.M. and Li,W.-H. (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
6. Bennetzens,J.L. and Hall,B.D. (1981) Codon Selection in Yeast. *J. Biol. Chem.*, **257**, 3026–3031.
7. Ikemura,T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J. Mol. Biol.*, **151**, 389–409.
8. Wright,F. (1990) The ‘effective number of codons’ used in a gene. *Gene*, **87**, 23–29.
9. Novembre,J.A. (2002) Accounting for Background Nucleotide Composition When Measuring Codon Usage Bias. *Mol. Biol. Evol.*, **19**, 1390–1394.
10. Peden,J.F. (2000) Analysis of Codon Usage.
11. Karlin,S., Mrázek,J., Campbell,A. and Kaiser,D. (2001) Characterizations of highly expressed genes of four fast-growing bacteria. *J. Bacteriol.*, **183**, 5025–40.
12. Karlin,S. and Mrázek,J. (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *J. Bacteriol.*, **182**, 5238–50.
13. Urrutia,A.O. and Hurst,L.D. (2001) Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics*, **159**, 1191–9.
14. Supek,F. and Vlahoviček,K. (2005) Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics*, **6**, 182.
15. Wan,X.-F., Xu,D., Kleinhofs,A. and Zhou,J. (2004) Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evol. Biol.*, **4**, 19.
16. Merkl,R. (2003) A Survey of Codon and Amino Acid Frequency Bias in Microbial Genomes Focusing on Translational Efficiency. *J. Mol. Evol.*, **57**, 453–466.
17. Seward,E.A. and Kelly,S. (2016) Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. *Genome Biol.*, **17**, 226.
18. Seward,E.A. and Kelly,S. (2018) Selection-driven cost-efficiency optimization of transcripts modulates gene evolutionary rate in bacteria. *Genome Biol.*, **19**, 102.
19. Puigbò,P., Bravo,I.G. and Garcia-Vallve,S. (2008) CAIcal: A combined set of tools to assess codon usage adaptation. *Biol. Direct*, **3**, 38.
20. Ho,B., Baryshnikova,A. and Brown,G.W. (2018) Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Syst.*, **6**, 192-205.e3.