A mechanism-aware and multiomic machine learning pipeline characterizes yeast cell growth

Christopher Culley[a,b], **Supreeta Vijayakumar**[b], **Guido Zampieri**[b], **and Claudio Angione**[b,c,*]

[a]School of Electronics and Computer Science, University of Southampton, UK; [b]Department of Computer Science and Information Systems, Teesside University, UK; [c]Healthcare Innovation Centre, Teesside University, UK

This manuscript was compiled on July 17, 2020

Metabolic modeling and machine learning are key components in the emerging next generation of systems and synthetic biology tools, targeting the genotype-phenotype-environment relationship. Rather than being used in isolation, it is becoming clear that their value is maximized when they are combined. However, the potential of integrating these two frameworks for omic data augmentation and integration is largely unexplored.

We propose, rigorously assess, and compare machine learning-based data integration techniques, combining gene expression profiles with computationally generated metabolic flux data to predict yeast cell growth. To this end, we create strain-specific metabolic models for 1143 *Saccharomyces cerevisiae* mutants and we test 27 machine learning methods, incorporating state-of-the-art feature selection and multiview learning approaches. We propose a multiview neural network using fluxomic and transcriptomic data, showing that the former increases the predictive accuracy of the latter, and reveals functional patterns that are not directly deducible from gene expression alone. We test the proposed neural network on further 86 strains generated in a different experiment, therefore verifying its robustness to a new independent dataset. Finally, we show that introducing mechanistic flux features improves the predictions also for knockout strains whose genes were not modeled in the metabolic reconstruction.

Our results thus demonstrate that fusing experimental cues with *in silico* models, based on known biochemistry, can contribute with disjoint information towards biologically-informed and interpretable machine learning. Overall, this study provides new tools for understanding and manipulating complex phenotypes, increasing both the prediction accuracy and the extent of discernible mechanistic biological insights.

Systems biology | Metabolic modeling | Machine learning | Flux balance analysis | Multimodal learning

The analysis of complex, high-dimensional biological data from heterogeneous sources is currently one of the main bottlenecks in molecular biology. Such data is generated by a range of high-throughput devices that target specific biomolecules or biological processes, and is collectively known as *omic* data. Representative examples are the global genetic composition of an organism - the genome - and the overall activation level of its genes at a certain time - the transcriptome.

Popular technologies permit the monitoring of various phenomena on a genetic and epigenetic level. However, in several applications, information on genes may have limited relevance to the task at hand, describing only a part of the processes taking place in biological organisms. Metabolic data are closer to the cellular phenotype but, despite recent innovations in omic technologies, sampling metabolic activity on a large scale is still challenging (1). Machine learning provides tools to identify and exploit patterns within this metabolic information, which can aid in our understanding of the underlying biological mechanisms (2). In this context, the heterogeneity of omic data has fostered the development and application of multimodal learning methods (3).

Machine learning techniques generally ignore previous biological knowledge in driving the pattern analysis, limiting the trustworthiness and interpretability of any obtained model. To fill these gaps, constraint-based modeling (CBM) can be used to simulate steady-state metabolism on a cellular scale. Metabolic flux profiles generated *in silico* have been previously used to inform specific machine learning models (4–9), in some cases providing predictive advantages, as recently reviewed (10). However, an integrative approach that fully exploits the multimodal learning potential to integrate such models with experimental omics, and therefore able to incorporate mechanistic biological knowledge in the learning process, is still lacking.

In this work, we propose a novel multimodal learning framework that leverages both transcriptomic data and strain-specific metabolic models to predict phenotypic traits of interest. We use this framework to predict the cellular growth for 1143 strains of *Saccharomyces cerevisiae*, one of the main eukaryotic platforms in basic research as well as in biotechnol-

Significance Statement

Linking genotype and phenotype is a fundamental problem in biology, key to several biomedical and biotechnological applications. Cell growth is a central phenotypic trait, resulting from interactions between environment, gene regulation, and metabolism, yet its functional bases are still not completely understood.

We propose and test a machine learning approach that integrates large-scale gene expression profiles and mechanistic metabolic models, for characterizing cell growth and understanding its driving mechanisms in *Saccharomyces cerevisiae*. At its core, a novel multimodal learning method merges experimentally- and model-generated data.

We show that our approach can leverage the advantages of both machine learning and metabolic modeling, revealing unknown interactions between biological domains, incorporating mechanistic knowledge, and therefore overcoming black-box limitations of conventional data-driven approaches.

Author contributions: Conceptualization, G.Z. and C.A.; Methodology, C.C., G.Z and C.A.; Software, C.C.; Investigation, C.C. and S.V.; Writing - Original Draft, C.C., S.V. and G.Z.; Writing – Review & Editing, C.C., S.V., G.Z. and C.A.; Visualization, C.C., S.V., G.Z. and C.A.; Project Administration, G.Z. and C.A.; Supervision & Funding Acquisition, C.A.

The authors declare no conflict of interest.

[*] To whom correspondence should be addressed. E-mail: c.angione@tees.ac.uk

ogy and, more recently, used for characterizing the processes associated with human diseases (11).

Cellular growth and gene expression are closely related in unicellular organisms, as they co-participate in mutual regulation. On the one hand, growth is sustained by genes implicated in ribosomal and translational functions. In parallel, the expression of genes is affected by global and unspecific regulation originating from the physiological state of the cell (12). This relationship has yet to be fully understood, and therefore predicting cellular growth following genetic manipulations is still challenging. Understanding and controlling cellular growth have important applications in disease modeling, biotechnology, and for the development of efficient cell factories (13). CRISPR-Cas9-enabled genetic engineering now gives the ability to modify yeast DNA with single-nucleotide precision *in vivo* (14), achieving engineered strains that maximize a desired output. However, the identification of such strains is a complex issue (15). For instance, streamlining yeast metabolism for the production of valuable compounds often requires the deletion of multiple genes and efficient diversion of resources towards production pathways (16).

In an attempt to fully elucidate relationships between cellular growth and other processes, mathematical models have been developed, particularly in bacteria and yeast (17–19). For instance, coarse-grained models were designed to describe the global relationship between the allocation of resources toward protein synthesis and growth (20). Further, extensive models of metabolic networks are commonly used to simulate cellular metabolism under different growth conditions (21, 22). These models offer quantitative mechanistic representations of molecular processes, but often require detailed knowledge about uptake rates from the environment to achieve precise estimates.

On the other hand, accurate and flexible models connecting gene expression and cell growth can be obtained by data-driven statistical and machine learning methods. As gene expression maintains a steady-state during the log phase (23), it is possible to predict the growth rate even in cases where experimental measurements are not feasible. This is particularly relevant in the development of synthetic systems, where phenotypic traits have to be tightly controlled. Previous research has focused on building linear predictive models for yeast growth (24), and more recently machine learning for both *E. coli* and *S. cerevisiae* (25). While both studies used gene expression profiles alone, metabolic activity is also tightly bound to cell growth (26).

Our idea is that reconnecting metabolic activity to cell growth with a data-driven and multiview approach should support more accurate machine learning predictions, while incorporating biological mechanisms within the learning process. To the best of our knowledge, this is the first time that an approach of this kind is proposed. To investigate this idea, we used a compendium of 1143 single-gene knockout *S. cerevisiae* strains, with their genome-wide expression profiles as training data to build models that predict cell doubling times. We augmented the array of biological predictors by incorporating a metabolic modeling phase, wherein we use CBM transcriptomic profile integration to simulate strain-specific metabolism using parsimonious flux balance analysis (pFBA). From these simulations, we extracted reaction fluxes as additional features (fluxomic data). We then applied machine learning methods using the transcriptomic and fluxomic datasets combined across 27 data-method combinations, testing different approaches for their multiview integration. When the integration of the two omics is performed within a neural network architecture, we found a significant improvement compared to using transcriptomic data alone. Upon finding that a newly proposed model, a multimodal artificial neural network, achieves the best performance, we tested it on further 86 "unseen" strains generated in a different experiment and not used in the training phase, verifying its robustness to this new independent dataset.

Our contributions thus focus on two aspects: (i) an investigation into the viability of building predictive models using transcriptomic and fluxomic information through a comparison of machine learning, feature selection, and multiview data integration approaches; and (ii) an examination of the benefits of using metabolic modeling in building multimodal machine learning predictive models, evaluating to what extent this mechanistic data is used to drive the learning process.

Results

Our goal was to develop and evaluate a multiomic mechanism-aware pipeline for predicting *S. cerevisiae* growth rate. To this end, we developed the workflow summarized in Figure 1. In brief, we used constraint-based modeling (CBM) of metabolism to estimate the metabolic activity of each yeast mutant in the exponential growth phase, starting from their transcriptional activity. Then, we built and cross-compared 27 machine learning models of yeast growth from a combination of transcript abundance and metabolic flux information. These steps and their output are described in detail in the following.

Strain-specific metabolic modeling of yeast mutants. Genome-scale metabolic models (GSMMs) aim to capture and simulate the entire metabolic activity within a cell. Since different transcription rates lead to alterations of cell behavior, we used gene expression data to create 1229 strain-specific models that emulate the corresponding metabolism. Through these simulations, we extracted a measure of this metabolic activity in the form of reaction fluxes for each strain (fluxomic data).

In particular, we focused on a transcriptomic dataset with 1143 single deletion strains of *S. cerevisiae* (27), and a second dataset comprising 86 single and double mutants (28), for a total of 1229 strains. The former was used as the main resource for model training, optimization and testing, while the latter served as an experimentally-independent test set in the predictive modeling stage. We used a recently refined GSMM of yeast metabolism (29) in conjunction with Eq. 2 in *Materials and Methods* to build the corresponding 1229 strain-specific models. This was achieved through a set of 908 genes involved in metabolism, represented within the yeast GSMM and put in relation to the biochemical reactions they control. In the following, we will refer to the full transcriptomic profiles as "gene expression" (GE) data, and to the reduced transcript information from these 908 genes as "metabolic gene expression" (MGE), as depicted in Figure 1.

To create the strain-specific metabolic models, we altered the reaction bounds within the yeast GSMM based on expression fold change levels in the MGE dataset. To reproduce

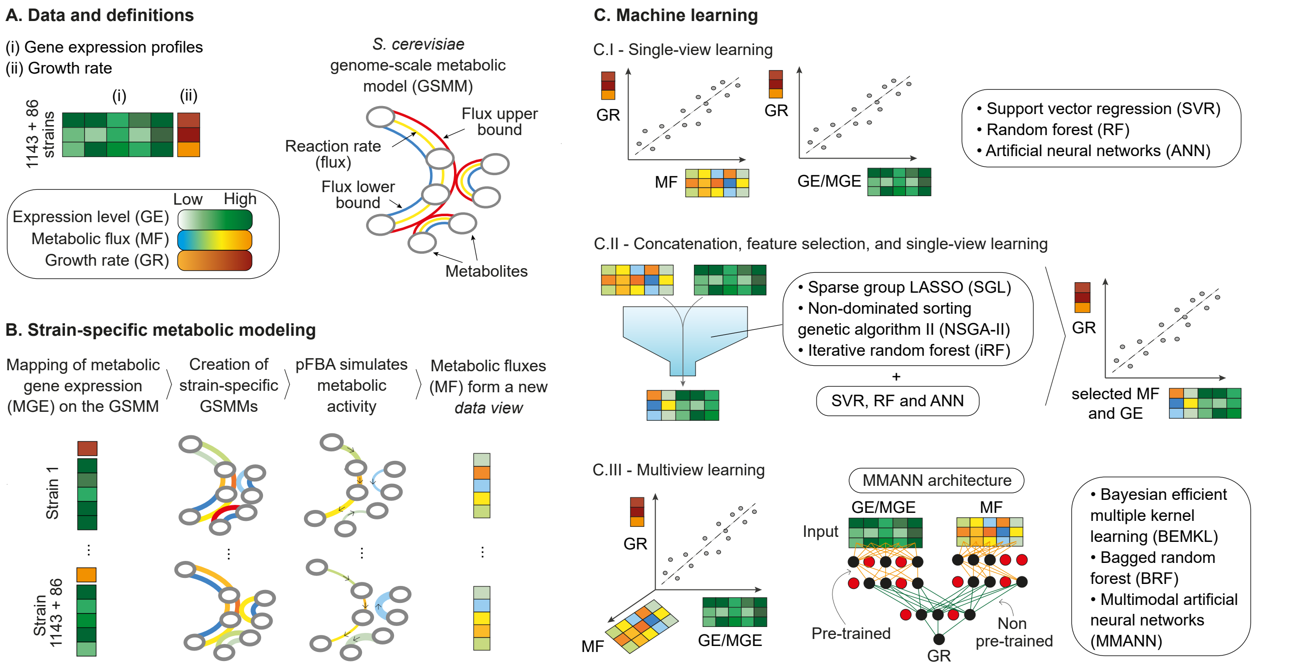
A. Data and definitions
(i) Gene expression profiles
(ii) Growth rate
(i)
(ii)
1143 + 86 strains
Low
High
Expression level (GE)
Metabolic flux (MF)
Growth rate (GR)
S. cerevisiae genome-scale metabolic model (GSMM)
Flux upper bound
Reaction rate (flux)
Flux lower bound
Metabolites
B. Strain-specific metabolic modeling
Mapping of metabolic gene expression (MGE) on the GSMM
Creation of strain-specific GSMMs
pFBA simulates metabolic activity
Metabolic fluxes (MF) form a new data view
Strain 1
Strain 1143 + 86
C. Machine learning
C.I - Single-view learning
GR
MF
GR
GE/MGE
• Support vector regression (SVR)
• Random forest (RF)
• Artificial neural networks (ANN)
C.II - Concatenation, feature selection, and single-view learning
• Sparse group LASSO (SGL)
• Non-dominated sorting genetic algorithm II (NSGA-II)
• Iterative random forest (iRF)
+
SVR, RF and ANN
GR
selected MF and GE
C.III - Multiview learning
GR
MF
GE/MGE
MMANN architecture
GE/MGE
MF
Input
Pre-trained
Non pre-trained
GR
• Bayesian efficient multiple kernel learning (BEMKL)
• Bagged random forest (BRF)
• Multimodal artificial neural networks (MMANN)

Fig. 1. Our multiomic integration and prediction framework, including all the datasets and machine learning methods used in this study. The input is a gene expression screen of 1143 single-knockout yeast strains (plus 86 single- and double-knockout strains used for independent validation), coupled with their relative growth rate and a genome-scale metabolic model (GSMM) of *S. cerevisiae* (A). Our methodology divides into two main stages. In the metabolic modeling stage (B), we extracted the gene expression (GE) data for the genes involved in metabolism (MGE), and used it to tailor the flux constraints of the GSMM in a strain-specific manner. Next, we applied parsimonious flux balance analysis (pFBA) to such strain-specific GSMMs to obtain the associated metabolic fluxes (MF). In the machine learning stage (C), we used the GE, MGE, and MF data to construct machine learning models of yeast growth. This was achieved through: (i) single-view learning - using only GE, MGE, or MF; (ii) concatenation, feature selection, and single-view learning - reducing the number of GE and MF predictors; and (iii) multiview learning - integrating the multiomic data with algorithms designed for multiple data sources (also referred to as *data modes* or *data views*). In total, 27 dataset-model combinations were tested in this stage, including a custom multimodal neural network (MMANN).

nutritional conditions, we set the uptake rates according to the feed composition used in the original study (see *Materials and Methods*). We then used pFBA to determine the reaction fluxes for the entire network by maximizing the biomass accumulation rate subject to model constraints. In this setting, we ensure that metabolic activity is coupled with gene expression and independent of environmental conditions, which are homogeneous across all strains (Figure 2a). Figure 2b-c shows the relationship between the pFBA-predicted biomass accumulation rate and the experimentally measured relative cell doubling time in the two sets of mutants. As expected, we obtained a clear negative correlation between the two quantities, with a Pearson's correlation coefficient PCC $= -0.66$, $p < 10^{-15}$ in the first set, and PCC $= -0.76$, $p < 10^{-15}$ in the second set.

Metabolic modeling of the yeast mutant populations also allowed us to identify pathways of biological interest that are highly correlated with growth, therefore providing means to assess the mechanistic knowledge supporting the machine learning models we developed in the next stage. Figure 2d shows the mean absolute correlation of fluxes inside each pathway with the relative doubling time. Among those pathways that correlate most strongly with growth ($|\text{PCC}| \geq 0.6$) we found amino acid and aminoacyl-tRNA metabolism, as well as pathways involved in producing the fuel for growth such as starch, sucrose, riboflavin and fructose metabolism, in keeping with previous experimental results (30). Other highly-correlated pathways act as intermediaries between processes that are important for cell growth, such as C5-Branched dibasic acid and galactose metabolism. Furthermore, we identified: purine metabolism, which has been found to regulate cell growth (31); RNA degradation, which has been shown to be strongly correlated with yeast growth rates (32); sulfur metabolism, which can actively promote initial cell division (33). Finally, the fact that growth rate is also correlated with pyrimidine supports recent research suggesting that its limitation causes the depletion of UTP and CTP, which in turn limits RNA biosynthesis, a limiting factor for cell growth (34).

Prediction of cellular growth based on transcriptomic and fluxomic profiles. Starting from gene expression (GE) and metabolic flux (MF) profiles of yeast mutants as two *data views*, we used the associated relative growth rate as a target to train our predictive machine learning models. As the nutritional conditions are fixed for all the strains, we assumed that variation on the level of gene regulation and expression is the main contributor to metabolism and growth. In this stage, we adopted the workflow depicted on the right-hand side of Figure 1.

Firstly, we explored three traditional machine learning techniques, each one with previous encouraging results in biological predictive tasks: (i) support vector regression (SVR) – often the learning tool of choice in computational biology due to its non-linear decision boundary and ability to handle high dimensional datasets (35, 36); (ii) random forest (RF) – able to handle heterogeneous data types in high dimensions and to account for both correlation and interaction among features, which has led to success in predictive modeling in multiple

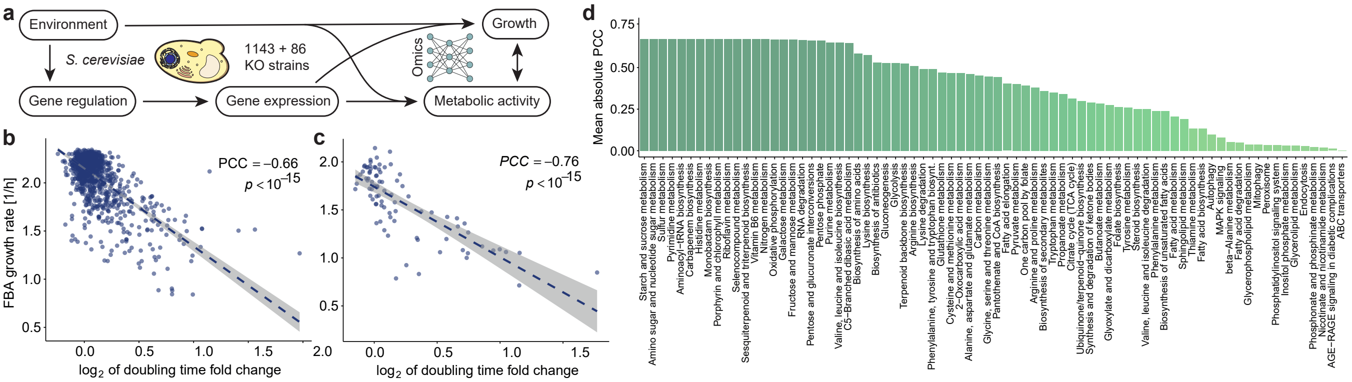

Fig. 2. Results of strain-specific metabolic modeling of yeast knockouts. **(a)** Relationship between cell growth and the main biological processes. While most models consider either gene expression or metabolism, here we seek to integrate both views within a unified computational framework. In our study, environmental conditions are fixed, hence cellular growth and metabolism are mainly driven by the influence of varying gene regulation and expression conditions. **(b-c)** Yeast mutant experimental relative doubling time plotted against their biomass accumulation rate, computationally estimated by strain-specific pFBA, both for the initial set (panel b) and for the experimentally-independent test set **(panel c)**. The negative correlation suggests that our strain-specific constraint-based modeling approach recapitulates the measured yeast growth. **(d)** Mean absolute correlation between experimental relative doubling time and strain-specific GSMM reaction fluxes within each metabolic pathway. High correlations were identified for meiosis, amino acids, and carbohydrates metabolism.

biological domains (37); and (iii) artificial neural networks (ANNs) – extremely effective in learning and modeling complex systems, with recent research reconstructing cell functionality (38) and predicting phenotypes from multiomic data (39). We applied these methods to GE, MGE, and MF data separately, in a single-view fashion, to obtain a baseline performance for the following steps.

In a second stage, we studied the integration of base omic datasets. Because our combined data represents two distinct *views* on the same biological systems, in order to thoroughly investigate the use of complementary information we explored three data strategies: (i) early integration – where GE and MF are concatenated and treated as a single dataset denoted as GE-MF; (ii) intermediate integration – where model building is carried out on a combined transformation of the input *views*; and (iii) late integration – where a model is separately built within each view and then the models are fused (3).

For intermediate and late integration, we used three multiview methods based on those employed in the single-view scenario. First, we considered Bayesian Efficient Multiple Kernel Learning (BEMKL) (40), applying separate radial basis kernels to the MF and GE datasets. Second, we used bagged random forest (BRF) with distinct forests learned on transcriptomic and fluxomic profiles. Finally, we designed and built a multimodal artificial neural network (MMANN) in order to independently extract latent information from the two omic views and then fuse it together via additional neural layers (see SI Appendix for details).

The multiomic datasets considered in our predictive framework have a large number of features, which in general can contribute to various extents towards the predicted growth value. Non-contributing features add noise to the data, therefore giving potentially weaker predictive models whilst increasing the training effort. To overcome this 'curse of dimensionality', feature selection and regularization techniques were incorporated with the aim of isolating the most predictive features. Also in this task, we explored three state-of-the-art approaches: (i) sparse group lasso (SGL) (41) – due to its ability to take into account the correlated and modular nature of biological functions; (ii) non-dominated sorting genetic algorithm II (NSGA-II) (42) – for its ability to optimize multiple objectives; and (iii) iterative random forests (iRF) (43) – for its ability in capturing non-linear interactions among features (see SI Appendix). Each of these techniques offers a different perspective on feature selection and is applied to GE-MF as an additional step of early integration. We thereby created three further datasets (SGL data, NSGA-II data, and iRF data, respectively) comprising the features identified by each of these approaches.

Comparison of 27 multiomic machine learning models of yeast growth. The methods outlined in the previous section globally constitute a wide and diversified collection of state-of-the-art data-driven prediction tools, applicable to different sets of omic data. In order to identify the most effective approach, we performed a systematic comparison of their predictive accuracy, covering 27 dataset-method combinations. We evaluated each combination by training and optimizing a model with 80% of the 1143 samples in our primary dataset, and testing it with the remaining 20%. The hyperparameters were selected by grid search as described in *Materials and Methods*. The entire procedure was repeated 100 times to capture the random variation in training and validation, while maintaining the same final test set.

Table 1 and Figure 3 give a breakdown of the predictive modeling results. Firstly, we found highly variable scores for single-omic predictions, depending on whether they referred to transcriptomic or fluxomic data. In fact, both GE and MGE consistently achieved higher accuracy than MF profiles. Analogously, the complete GE performs better than the MGE subset, therefore highlighting the importance of metabolic or non-metabolic genes that are not currently used by the yeast GSMM. Secondly, our results suggest that early- and late-integration approaches on average do not improve single-omic accuracy, although also this trend is associated with large variation depending on the specific data-method combination. Conversely, a small but tangible improvement was observed for intermediate integration approaches. Thirdly, SVR- and ANN-based approaches generally tend to be more accurate than tree-based approaches. It is interesting to observe that, overall, the most accurate dataset-method combination is the

MMANN model using both GE and MF, immediately followed by SVR trained on GE alone, with statistically significant MDAE differences between the two (see Figure 3d).

By examining the predictive scores achieved by single-view and multiview ANNs, we notice a clear improvement of multiomic models against the stand-alone GE- and MGE-based models, in contrast to other multiview methods. It thus emerges that ANNs constitute the most suitable framework for the integration of transcriptomic and fluxomic data in terms of predictive benefits, among those considered here. Our results also suggest that, despite the relatively weak performance of the fluxes alone, their useful information cannot be discerned from GE and is therefore complementary to it. This is supported by examining the prediction output correlations shown in Figure 3d, where the models produced using the fluxomic data have a prediction set that largely differs from the other models. MMANNs seem thus to use the metabolic modeling to gain information that can not be acquired from the gene expression alone. Additionally, using fluxes as additional features improves the ability to mechanistically explain the predictions from ANNs, making them biologically-interpretable.

Furthermore, data condensation through feature selection (SGL, NSGA-II and iRF data) increases the predictive capability of SVR and occasionally RF, but our results indicate that this is not the case with ANNs. Since our ANNs include at least two hidden layers, this suggests that ANNs can identify predictive non-linear relationships among genes and metabolic reactions that involve a larger set of features.

Table 1. Full set of accuracy scores across all 27 dataset-algorithm combinations, shown in Figure 3: root mean squared error (RMSE), mean absolute error (MAE), median absolute error (MDAE), Pearson's correlation coefficient (PCC) and fluxomic features representation (FFR, the percentage of metabolic flux features over the total number of features). Values in bold represent the best scores for each data integration scenario, while the best global performance for each measure is highlighted by an asterisk. The MMANN model consistently outperforms all other models and, with 36% of the features being fluxomic, demonstrates the utility of the additional metabolic modeling stage in our pipeline.

Dataset(s)	Method	RMSE	MAE	MDAE	PCC	FFR
Single omics						
GE	SVR	**0.102 ± 3e-04***	**0.067 ± 0.001***	**0.045 ± 0.004**	**.902 ± .001**	0%
GE	RF	0.127 ± 0.001	0.077 ± 4e-04	0.049 ± 0.001	.864 ± .002	0%
GE	ANN	0.122 ± 0.007	0.079 ± 0.008	0.053 ± 0.010	.876 ± .004	0%
MGE	SVR	0.115 ± 0.003	0.070 ± 4e-04	0.046 ± 2e-04	.872 ± .006	0%
MGE	RF	0.130 ± 0.001	0.079 ± 4e-04	0.050 ± 0.001	.855 ± .002	0%
MGE	ANN	0.139 ± 0.008	0.091 ± 0.008	0.065 ± 0.011	.838 ± .005	0%
MF	SVR	0.203 ± 0.006	0.117 ± 0.003	0.065 ± 3e-04	.504 ± .033	100%
MF	RF	0.185 ± 0.002	0.109 ± 0.001	0.065 ± 0.002	.611 ± .009	100%
MF	ANN	0.196 ± 0.009	0.125 ± 0.016	0.083 ± 0.021	.588 ± .003	100%
Early integration						
GE-MF	SVR	0.132 ± 0.009	0.079 ± 0.004	**0.048 ± 0.004**	.828 ± .029	36%
GE-MF	RF	0.126 ± 0.001	0.077 ± 0.001	**0.048 ± 0.001**	.866 ± .003	36%
GE-MF	ANN	0.132 ± 0.007	0.085 ± 0.009	0.057 ± 0.011	.847 ± .006	36%
SGL data	SVR	0.117 ± 0.001	0.082 ± 3e-04	0.058 ± 0.001	.867 ± .002	34%
SGL data	RF	0.130 ± 0.001	0.082 ± 5e-04	0.053 ± 0.001	.844 ± .003	34%
SGL data	ANN	0.163 ± 0.011	0.105 ± 0.013	0.072 ± 0.019	.805 ± .005	34%
NSGA-II data	SVR	0.178 ± 0.014	0.103 ± 0.005	0.063 ± 0.002	.653 ± .069	24%
NSGA-II data	RF	0.179 ± 0.020	0.110 ± 0.010	0.067 ± 0.004	.653 ± .077	24%
NSGA-II data	ANN	0.154 ± 0.011	0.100 ± 0.014	0.067 ± 0.017	.804 ± .013	24%
iRF data	SVR	**0.108 ± 0.002**	**0.072 ± 0.001**	0.050 ± 0.001	**.891 ± .002**	0%
iRF data	RF	0.120 ± 0.001	0.074 ± 3e-04	0.049 ± 0.001	.870 ± .002	0%
iRF data	ANN	0.136 ± 0.008	0.090 ± 0.010	0.065 ± 0.014	.854 ± .003	0%
Intermediate and late integration						
GE and MF	BEMKL	0.182 ± 1e-04	0.110 ± 2e-04	0.066 ± 1e-04	.626 ± .001	36%
GE and MF	BRF	0.145 ± 0.001	0.086 ± 3e-04	0.053 ± 0.001	.810 ± .003	36%
GE and MF	MMANN	**0.102 ± 0.001***	**0.067 ± 0.001***	**0.043 ± 0.002***	**.906 ± .002***	36%
MGE and MF	BEMKL	0.182 ± 7e-05	0.110 ± 1e-04	0.067 ± 2e-04	.625 ± 3e-04	79%
MGE and MF	BRF	0.147 ± 0.001	0.087 ± 4e-04	0.054 ± 0.001	.803 ± .003	79%
MGE and MF	MMANN	0.112 ± 0.001	0.073 ± 0.001	0.047 ± 0.002	.882 ± .003	79%

Generalization to an experimentally-independent dataset. For a machine learning model to be considered generalizable and of high utility, performance stability is paramount. Especially in those settings where new data is collected in environments that differ from those of the training data, it is imperative that the prediction accuracy does not degrade under this new and "unseen" setting. However, this can be challenging to achieve when all the training, validation and test data originates from a single experiment (44). To verify the ability of our MMANN model to generalize to experimentally-independent data, we applied it to a different set of yeast mutants cultivated in the same nutritional conditions. Importantly, the new mutants not only comprise single knockout strains, but also double knockouts, exposing our model to epistatic effects on which it was not trained (28). This analysis therefore allowed us to investigate the additional question of whether our multiomic MMANN model, trained only on single mutants, could also generate reasonable predictions for double mutants (further details can be found in *Materials and Methods*).

Figure 3c shows the results on the experimentally-independent test set. In the single knockout case, MAE and MDAE increase, but RMSE and PCC improve compared to the first test case. This might be caused by potential batch effects across experiments that represent a source of systematic error, often particularly visible on the level of MDAE (45). However, the key patterns are captured as RMSE and PCC are consistent with previous tests. Double knockouts were not present in the training dataset and therefore, expectedly, the model performs less well in this scenario. We note also that, even in this out-of-distribution double-gene knockout setting, the correlation with target growth rates is particularly strong. This suggests that, if a relative rather than absolute strain identification is required, then training on single knockouts and testing on double knockouts using the MMANN approach would give a setting from which strains could be compared with confidence. Taken together, assuming an appropriate training environment and batch effect corrections, these results support the use of MMANN as a strong predictive method for this task, and demonstrate robust generalization across experiments.

Functional classification of relevant multiomic predictors. As described above, the application of feature selection methods allowed us to reduce the number of biological variables to facilitate model learning. At the same time, it provided us with concise sets of predictors that hold a strong association with the cellular growth from a data-driven point of view. We found that SGL yields 71 GE and 36 MF features as most relevant, while iRF identifies 68 unique GE features. Thirdly, with the NSGA-II feature selection, nine variable sets are selected as members of the Pareto front of possible optimal solutions (see SI Appendix), which include 218 GE and 51 MF unique features. Figure 4a shows the metabolic pathways associated with the GSMM reactions selected by each of these algorithms, while Figure 4b illustrates the main functional categories for the selected genes, obtained by querying the PANTHER classification system (46).

Among all biological processes, metabolic processes are the most prominent class for all three feature selection algorithms. By examining the organic metabolic processes, we found that a large proportion of reactions and pathways correspond to the biosynthesis and metabolism of macromolecules and organic

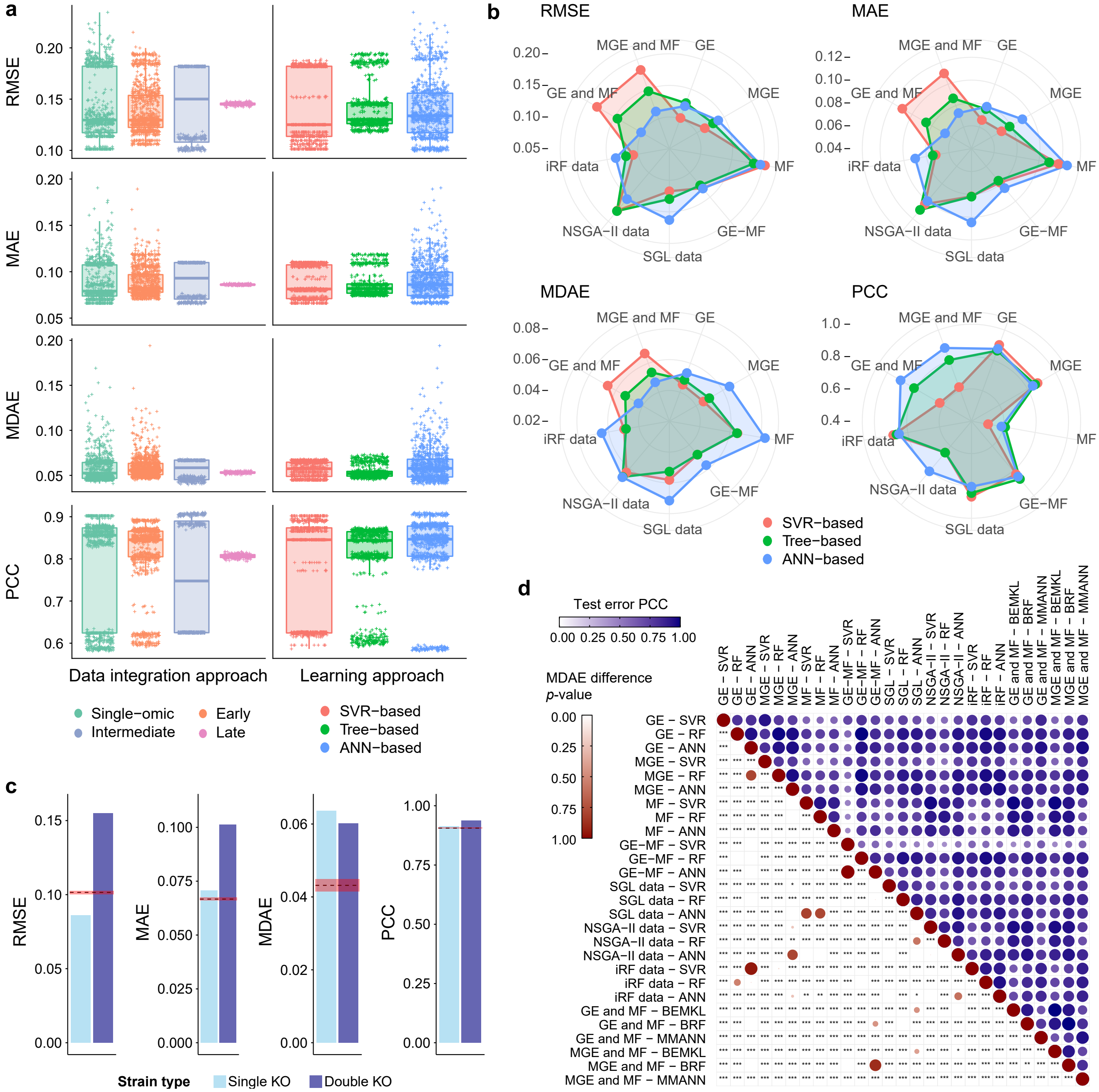

Fig. 3. Machine learning yeast growth prediction results. **(a)** Comparison of model predictive performance across data integration strategy and machine learning model type. Intermediate integration is overall the most effective approach, and notably better than single-omic models. Concomitantly, ANN- and SVR-based techniques appear generally more effective than tree-based techniques. **(b)** Comparison of model accuracy for all dataset-learning algorithm combinations, corresponding to numeric results shown in Table 1. The MMANN using both GE and MF profiles is overall the most accurate model, followed by GE-based SVR. **(c)** Error scores on the experimentally-independent test set. Dashed red lines represent the corresponding error score on the main test set, while shading areas represent their associated standard deviation. **(d)** In blue, Pearson's correlation between error score vectors on the test set, for each pair of data-method combination. In red, *p*-values of Wilcoxon rank-sum tests assessing the significance of MDAE differences, for each pair of data-method combination. One, two, and three stars represent significance at thresholds of 0.05, 0.01, and 0.001 respectively, rescaled by Bonferroni correction.

compounds, such as factors for transcription, translational initiation, and elongation (Figure 4c). This is consistent with the role in protein synthesis played by the translational machinery, which is critical for cell growth (47). No functional class was found statistically enriched, indicating that the joint contribution of multiple processes determines the actual growth rate. As regards MF features, SGL selected reactions largely involved in the metabolism of glycerolipids, glycerophospholipids and secondary metabolites, whereas reactions selected by NSGA-II encapsulate a more diverse variety of functions (see Figure 4a), ranging from the biosynthesis of amino acids and secondary metabolites to the metabolism of fatty acids, glycerophospholipids, and nucleotides.

The gene YDR472W (also known as TRS31) was selected by all three feature selection methods and encodes a core component of a subunit present in TRAPP complexes, which are responsible for Rab-mediated vesicle trafficking (48). All other selected genes and metabolic reactions are exclusive to one or two methods. Among the nine features selected by both iRF and NSGA-II, there are genes encoding binding proteins and transporters (see Supplementary Data 1). Similarly, the genes selected by SGL and NSGA-II also coded for mitochondrial transport and mRNA binding. The selection of genes linked to tRNA and cellular amino acid-related metabolic processes is consistent with the process of translational elongation during the assembly of amino acids into proteins, which consequently affects cellular growth and maximization of biomass. Despite the limited overlap among the features selected by the three methods (Figure 4d), their high-level functional classification is statistically coherent (χ^2 tests of independence, null hypothesis retained, $p = 0.72$ for biological processes and $p = 0.18$ for metabolic processes). This is consistent with the nature of cell systems, based on functional modularity and redundancy, and characterized by widespread cross-correlated omic cues.

For metabolic genes or reactions, their contribution to cell growth could be inferred also through CBM-only approaches, e.g. by simulating the effect of their artificial alterations. To compare a CBM-only approach with our multimodal machine learning approach, we performed a sensitivity analysis through *in silico* single-gene knockdown directly within the metabolic model, examining the impact on the biomass accumulation rate (SI Appendix). The genes and pathways that have the greatest effect on the biomass are listed in Supplementary Data 1, among which we found some overlap with the features selection algorithms. The downregulation of genes related to tRNA metabolic processes and the biosynthesis of amino acids such as arginine and phenylalanine resulted in zero biomass flux, consistently with the features identified by SGL and NSGA-II. From the perspective of individual algorithms, overlapping iRF-selected genes are related to pyrimidine and phospholipid biosynthesis, and to the pentose phosphate pathway. The NSGA-II genes whose deletion resulted in zero biomass are related to the metabolism of vitamin D and sphingolipid biosynthesis.

Analogously, we carried out a flux-coupling analysis to identify reaction fluxes on which growth rate is mutually dependent (fully coupled) or unilaterally dependent (directionally coupled) (49) (see SI Appendix for details). A total of 234 reactions were classified in either one of the two categories (Supplementary Data 1). Also in this case, we observed an overlap between some features that were selected by SGL or NSGA-II. Out of the 36 reactions selected by SGL, only three reactions are coupled with the biomass pseudo-reaction (with one fully coupled and two directionally coupled reactions), whereas 19 out of the 51 reactions selected by NSGA-II were found to be coupled (with one fully coupled and 18 directionally coupled reactions). However, it should be noted that CBM approaches are limited to the enzymes included in the genome-scale metabolic model and overlook the role of external biological factors. Thus, we argue that our integrative framework can be complementary to more traditional CBM approaches, and capture cross-omic relationships missed by them.

Interestingly, when examining rules within the GSMM that dictate the gene-protein-reaction associations, some of the reactions selected uncover formerly overlooked connections. For instance, the reactions involved in glycerophospholipid metabolism are selected by SGL but the corresponding genes are not. In fact, a closer inspection of these results revealed that the functionality of the selected gene and reaction features hardly overlap. Five reactions that constitute part of the glycerophospholipid metabolic pathway are controlled either exclusively or partially by the gene YPR140W, which is essential for maintaining the phospholipid content of the mitochondrial membrane. Indeed, *S. cerevisiae* is a popular choice of organism for studying glycerophospholipid homeostasis in eukaryotes, owing to tolerance with respect to its membrane lipid composition (50). These results support the case for the inclusion of both flux and gene features in order to augment the machine learning model with more data, while improving our mechanistic understanding of the role that each omic plays in the wider biological context.

Finally, given the high prediction accuracy of MMANN models, we sought to determine their most contributing features. To this end, we exploited recent advances in ANN interpretation via the SHapley Additive exPlanations (SHAP) method (51), a general approach for determining the contribution (called SHAP value) of individual features to model outputs. We applied SHAP to a randomly selected model from the set of MMANN models, selecting features with absolute mean SHAP values in the top percentile as highly-relevant, and obtaining 71 belonging to the transcriptomic domain and 10 to GSMM reaction fluxes (Supplementary Data 1). MMANN-associated GE features yield statistically-significant differences from those selected by the feature selection methods in terms of functional classification (χ^2 tests of independence, null hypothesis rejected, $p = 6.3 \cdot 10^{-4}$ for biological processes and $p = 2.2 \cdot 10^{-3}$ for metabolic processes). The information extracted by these models thus seems notably distinct, which may explain the higher performance of MMANNs. Among the top-contributing genes in MMANNs, many produce proteins binding to RNA, with several genes acting as mRNA splicing factors involved in pre-processing via the spliceosome. Some genes encode proteins that bind to DNA to repair mismatched nucleotides, as well as proteins responsible for dephosphorylation and protein/tRNA modification. This, along with the presence of an amino acid transporter gene, reaffirms the role of protein synthesis in relation to growth. Among the top contributing reactions, the main pathways (glycerophospholipid and inositol metabolism) are very closely linked, since inositol signaling is responsible for homeostasis and regulation of lipid metabolism (52).

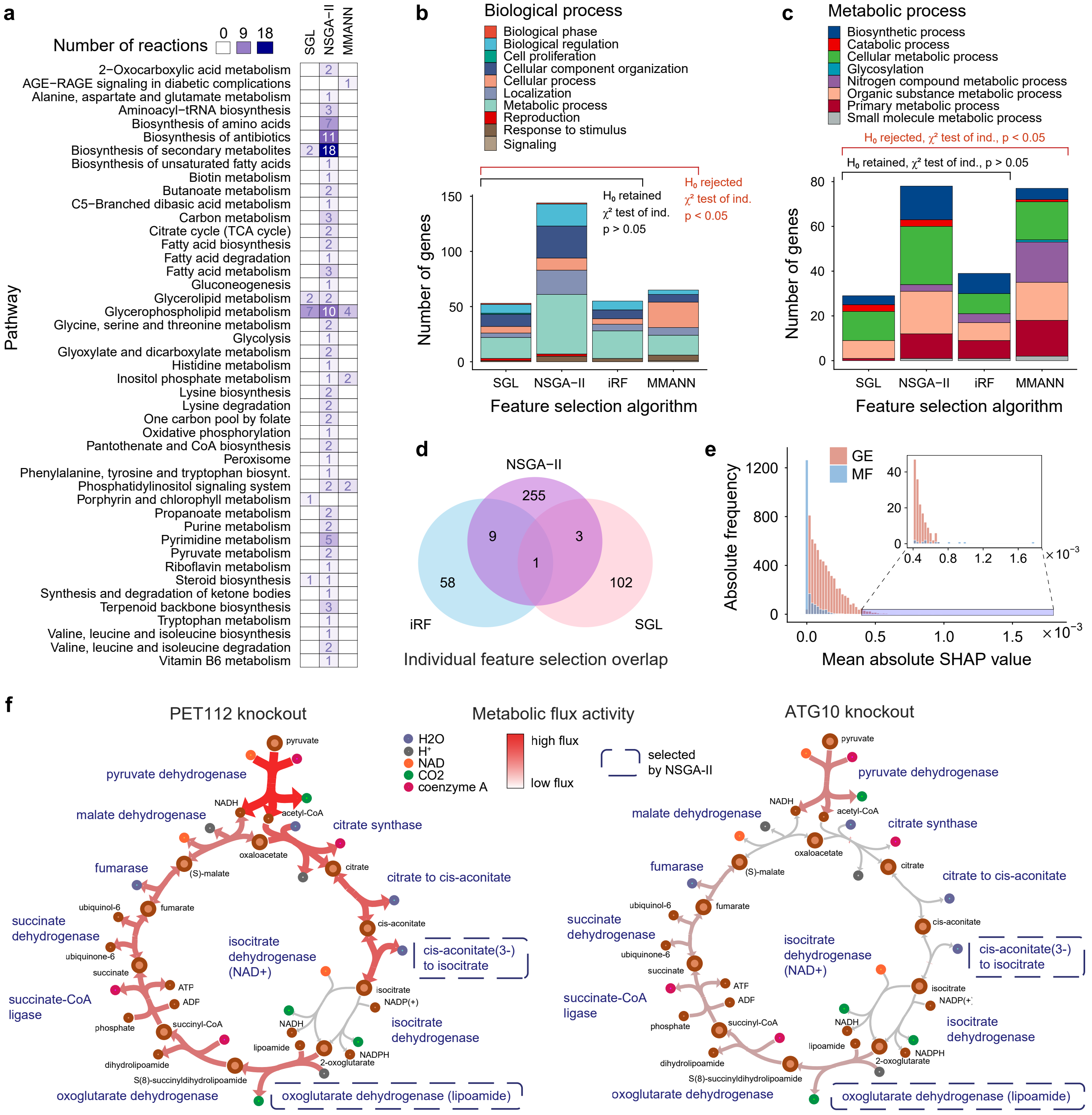

Fig. 4. Contribution of the omic features to the learning process. **(a)** Pathway classification of the metabolic features selected by SGL, NSGA-II, and MMANN. **(b-c)** Functional classification of the genes selected either by SGL, NSGA-II, iRF, or MMANN, based on GO biological processes and metabolic molecular functions, respectively. The number of features per functional class is independent of the selection method for SGL, NSGA-II, and iRF (χ^2 test of independence, null hypothesis H_0 retained, $p = 0.72 > 0.05$ for biological processes and $p = 0.18 > 0.05$ for metabolic processes), but dependent for MMANN (null hypothesis H_0 rejected, $p = 6.3 \cdot 10^{-4} < 0.05$ for biological processes and $p = 2.2 \cdot 10^{-3} < 0.05$ for metabolic processes). **(d)** Overlap in the individual features selected by SGL, NSGA-II, and iRF. A single feature is shared among iRF, NSGA-II, and SGL, represented by the expression of gene YDR472W. This suggests that individual features are used interchangeably by the feature selection methods (e.g., highly correlated gene expression values, or reactions with similar flux in a linear pathway) while, at a higher functional level, the pathway-level selected signal is consistent across all methods (as shown in panel b). **(e)** Distribution of feature importance in the MMANNs. These distributions are extracted from the MF and GE components of the MMANN models. Although the GE SHAP values have an overall higher contribution, the MF has a small number of features determined as highly contributing, demonstrating their predictive utility. **(f)** Metabolic flux through the citric acid cycle in two mutants: PET112 (left) and ATG10 (right), illustrating how condition-specific CBM can capture metabolic perturbations generated by the knockout of two genes not present in the GSMM, whose fluxes are exploited downstream by the machine-learning approaches. The color scale from grey (low) to red (high) indicates the amount of flux carried by each reaction in the pathway.

Contribution of fluxomic information in multiomic machine learning models. Although from the single-omic results it is clear that a large contribution in the most accurate multimodal learning model (MMANN) comes from the transcriptomic data, we showed that a significant and complementary amount of relevant signal is present in the metabolic view. Thus, we further investigated the extent to which this method exploits the information in MF rather than in GE. The variable importance distribution for each data source, estimated through SHAP, is plotted in Figure 4e. Although transcriptomic features have a higher mean absolute SHAP value and constitute the majority of the information used, fluxomic features also contribute a subset with high SHAP values. This shows that the predictive improvement obtained by the addition of MF profiles is directly attributable to active information sourcing from this *data view*.

Finally, in order to ascertain how the addition of MF affected the predictive accuracy on individual knockout strains, we compared the absolute error differences between ANNs (using only GE) and MMANNs (using both GE and MF). The knockout strains that recorded the highest differences between the mean errors were regarded as providing a more accurate prediction of growth rate due to the addition of MF to the model. The full list of strains for this analysis can be found in Supplementary Data 2. Among the 20 highest differences were many gene knockouts that played a role in DNA transcription or RNA processing, as well as enzymes involved in the sorting and modification of proteins. Interestingly, only two of these 20 genes are present within the GSMM. This shows that MF and machine learning can jointly contribute towards extracting more accurate and biologically-interpretable predictions by indirectly propagating perturbations on biological components into a GSMM, even when such components are not explicitly included in the GSMM. As an example, Figure 4f displays the difference in metabolic flux in the citric acid cycle between two different mutants, illustrating how our condition-specific CBM approach can capture metabolic perturbations generated by the knockout of genes not present in the GSMM (PET112 and ATG10), which in turn can be exploited by a data-driven model used downstream. This advocates the use of metabolic reactions as features for machine learning methods, using ad-hoc feature selections techniques for any given application.

Discussion

This work investigates the application of multiview and multistage learning to integrating experimental and *in silico*-generated omic data for the prediction of yeast cellular growth. To the best of our knowledge, this is the first time that such a framework is proposed and systematically evaluated across several machine learning approaches. The wide spectrum of models and data integration techniques considered here provides a useful starting point for future benchmarking. We verified that combining experimental transcriptomic and artificial fluxomic data can increase the prediction strength over individual omics, although the improvement is subject to the predictive model choice. In our study, the largest improvement was obtained through artificial neural networks, with multimodal neural networks being the strongest predictive model overall. Additionally, we demonstrated that the advantages in terms of prediction accuracy and biological insights can reach beyond what is directly captured mechanistically by the metabolic reconstruction used to generate the fluxomic profiles.

Although transcriptomic-constrained flux balance analysis is widely used in genome-scale metabolic modeling, there are additional methods that can inject further constraints in flux simulations (53–55). Similarly, additional information may lie in the solution space of strain-specific models. For instance, additional features could be extracted from a metabolic model, e.g. from the results of flux variability analysis or sampling. While in this work we focused on cross-comparing machine learning methods, an analogous survey could be performed on the level of constraint-based modeling techniques to generate reaction-level fluxes, as well as on the level of different base metabolic reconstructions. Furthermore, in this work, we adopted transcriptomic data as a benchmark, given their widespread use across biology and biotechnology studies. In the cases where further omic data are available, they could be implemented to perform predictions across different biological layers (5). Similarly, our framework could be extended to investigating varying environmental conditions.

It is interesting to note that multimodal artificial neural networks achieve higher accuracy compared to single-view neural networks, and to other methods overall, but also transcriptomics-based support vector regression achieves good performance scores. Indeed, multiomic data integration does not always guarantee improved predictions, especially when benchmarking over gene expression (56). While any difference in accuracy generally depends on the task, our findings demonstrate that the knowledge embedded in genome-scale metabolic models is complementary to gene expression and may support its exploitation by data-driven models in a variety of scenarios. Therefore, support vector regression also appears as a promising framework for further improving the predictions guided by transcriptomic and fluxomic data, once such complementarity is fully exploited.

Finally, it is important to note that metabolic flux information has a straightforward mechanistic interpretation, as it is directly linked to the underlying biochemistry. Data augmentation based on metabolic networks, combined with multiview learning, can therefore increase predictivity while providing direct mechanistic insights into the condition-specific interaction of metabolites that give rise to the phenotypic outcome. This can translate into advantages in terms of human ability to trust and employ more biologically-interpretable machine learning models, especially in scenarios where it is important to understand the effect of cell or metabolic engineering operations (10). Our results thus support the extension of such data- and knowledge-based multiomic machine learning to biological engineering and to other relevant phenotypic targets, such as the secretion of metabolites for drug development.

Materials and Methods

Transcriptomic and growth data. The main transcriptomic dataset used in this work was collected in a previous study (27), which provides two-channel microarray profiles for 1484 single-gene deletion strains of *S. cerevisiae* during the mid-log phase. We downloaded this data from the supplementary material of a second study (63), which provides also relative growth rates compared to the wild type for 1312 strains, expressed as the $\log_2$ of the doubling time ratio between each strain and the wild-type. After merging transcriptomic profiles and growth rates, we obtained 1143 samples with their associated growth rates, which we used in the following stages.

An independent dataset for testing the proposed MMANN was obtained from a third study (28), providing gene expression profiles for single and double gene deletion strains of *S. cerevisiae* on the same microarray platform. Among these strains, we selected the single mutants that do not overlap with those in our primary dataset (14 strains) and all the double mutants (72 strains). In this second dataset, 58 of the genes present in the main training dataset were missing. To ensure consistency of features, i.e the same gene sets, and feed this new data into our pre-trained models, we imputed the gene expression values for the missing genes by linear regression based on the other variables. Upon imputation of missing values, the obtained 86 mutants represented an experimentally-independent set of conditions and served as a real-case scenario for using our proposed MMANN method.

Genome-scale metabolic modeling. A genome-scale metabolic model (GSMM) is a collection of all known biochemical reactions and transmembrane transporters that occur within an organism. The reaction network is mathematically represented as a stoichiometric matrix $\mathbf{S}$, capturing the exact proportions of reactants and products involved in each biochemical transformation (57). Reaction rates (fluxes) are mass- and energy-balanced assuming a metabolic steady-state, and can be described by a vector $\mathbf{v}$ of reaction fluxes through the network, limited by their lower and upper bounds $\mathbf{v}_{lb}$ and $\mathbf{v}_{ub}$. The constraints given by $\mathbf{v}_{lb}$ and $\mathbf{v}_{ub}$ can be modified to model varying genetic or environmental factors, yielding a context-specific metabolic model consistent with experimental data.

We estimated the metabolic fluxes associated to each transcriptional condition by solving the following parsimonious FBA problem:

$$\begin{aligned} \min_{\mathbf{v}} \ & \|\mathbf{v}\|_1 \\ \text{subject to } \ & \mathbf{w}^\top \mathbf{v} = f\,, \\ & \mathbf{S}\,\mathbf{v} = 0\,, \\ & \mathbf{v}_{lb}\,\Theta \leq \mathbf{v} \leq \mathbf{v}_{ub}\,\Theta\,. \end{aligned} \qquad [1]$$

Here $\mathbf{w}$ is a binary vector expressing the biomass pseudo-reaction as unique objective, while f is the maximal growth rate achievable by the network under the given constraints. The impact of each transcriptional condition is represented by Θ, which is the *gene set expression* vector obtained by mapping the expression of the individual genes onto the associated reactions. This involves converting logical gene-protein-reaction association rules into max/min operations, as follows:

$$\begin{aligned} \Theta(g_1 \wedge g_2) &= \min\{\theta(g_1), \theta(g_2)\} \\ \Theta(g_1 \vee g_2) &= \max\{\theta(g_1), \theta(g_2)\}, \end{aligned} \qquad [2]$$

where $\theta(g)$ represents the expression level of a gene g, and Θ represents the effective expression level of the gene set $\{g_1, g_2\}$ (58). We refer the reader to the SI Appendix for more details regarding the nutritional conditions.

In this work, we used the *i*Sce926 yeast GSMM, which includes 926 genes, 3494 reactions, and 2223 metabolites (29). Among these 926 genes, a total of 908 (98%) are present in our main transcriptomic dataset. To solve Eq. 1, we used the COBRA toolbox 3.0 (59) with the PDCO solver. The solutions provide steady-state fluxes for every reaction in the *i*Sce926 GSMM across the 1143 yeast strains from the main dataset and the 86 strains from the experimentally-independent dataset.

Machine learning models. To predict the relative doubling time, expressed as the $\log_2$ of the doubling time ratio with respect to the wild type, we started from the transcriptomic and fluxomic profiles as features, and we used the following supervised learning methods: support vector regression (SVR) (35), random forest (RF) (37), artificial neural networks (ANNs) (60). To integrate omic profiles and obtain multiomic machine learning models, we employed the following multiview methods: Bayesian efficient multiple kernel learning (BEMKL) (40), bagged random forest (BRF), and multimodal artificial neural networks (MMANNs). Further, to reduce the number of omic predictors, we employed sparse group LASSO (SGL) (41), non-dominated sorting genetic algorithm II (NSGA-II) (42), and iterative random forest (iRF) (43) (see SI Appendix for details on each of these methods).

Machine learning model selection, training, and testing. To assess model generalization, we randomly split our samples into train and test subsets comprising 80% and 20% of the main dataset, respectively. Training data was used for fitting the models and learn latent patterns present in the data, which can predict the relative doubling time of yeast mutants. Since many of the adopted methods have hyperparameters that can impact the learning process, we performed a grid-search to identify the optimal hyperparameter settings with the use of validation data subsets. Using the 80% data portion, we applied 5-fold cross-validation repeated three times for all methods, except the ANN-based models, for which we used a fixed 10% of the training set for validation. After selecting the hyperparameters, we trained each model again, this time using the full training data - validation samples included. In order to measure model performance, we used the obtained models to make predictions on all the samples in the test set, which are disjoint from those in the training and hyperparameter selection phases.

To account for stochastic variability - whether in cross-validation or during the optimization process in the case of ANN - we repeated the training-test procedure 100 times for each combination of dataset and ANN-based model, and repeated the selection-training-test procedure 100 times for each other dataset-method combination. Feature selection methods were optimized and applied one time only. Lastly, we applied a randomly selected MMANN model to the experimentally-independent test set to simulate a real-use scenario. To ensure full reproducibility, we provide the train-test split indexes and the random seed used, along with details on methods, software packages and hyperparameter search spaces in the SI Appendix.

Data normalization and performance metrics. When feeding the different *data views* to the machine learning techniques, we used z-score normalization, where the mean and standard deviation of the training data was also used to normalize the test data in order to prevent information leakage. We used the normalized data in all the learning approaches due to the different data distributions of the two views (fluxes and gene expression), also noting in general that normalization is a requirement for SVR and enables faster convergence in ANNs.

The hyperparameter selection focused on minimizing the root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}, \qquad [3]$$

where model predictions y_i are compared with observed growth rates $\hat{y}_i$ across all n strains. The RMSE emphasizes incorrect predictions. When evaluating and comparing models we used three additional metrics, namely the mean absolute error (MAE):

$$\text{MAE} = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n}, \qquad [4]$$

the median absolute error (MDAE):

$$\text{MDAE} = \mathit{median}(|\hat{y}_1 - y_1|, ..., |\hat{y}_n - y_n|), \qquad [5]$$

and the Pearson's correlation coefficient (PCC). MDAE statistical differences across data-method pairs were estimated by Wilcoxon rank-sum tests through the *wilcox.test* R function, whose $p-$values were adjusted via Bonferroni correction.

Artificial neural network interpretation. To quantify the variable contributions in the MMANN models, we used the SHapley Additive exPlanations (SHAP) method (51). SHAP uses a game-theoretic approach to determine the importance of a particular feature to individual data inputs. SHAP values are thus feature importance scores defined to satisfy local accuracy, missingness, and consistency properties. We used a variant of the SHAP method specifically designed for ANN models, called Deep SHAP (51), whose working principle is the back-propagation of unit activation differences to input features. The top contributing features inspected in terms of biological classification were chosen as those in the largest mean SHAP value percentile, where the mean was computed over the training samples.

Biological feature classification. The biological classification for the genes identified by the feature selection methods and SHAP was obtained with the PANTHER classification system (46). The KEGG pathway annotation (61) for GSMM reactions was obtained from a curated *S. cerevisiae* GSMM (62). The statistical enrichment tests on PANTHER were run with default parameters. To assess associations between the feature selection methods and the selected gene features, χ^2 independence tests were run on biological and metabolic process classification classes via the *chisq.test* R function. These tests were performed first across SGL, NSGA-II, and iRF, and finally with the inclusion of the MMANN features obtained through SHAP.

Data availability. The microarray and growth data obtained for this study is available on GEO (accession numbers GSE42526, GSE42527, and GSE42536), Array Express (E-MTAB-1383, E-MTAB-1384 and E-MTAB-1385), and as flat files from the authors of the original studies (27, 28, 63). The yeast metabolic model can be found in the supplementary material of the corresponding paper (29). All data, models, and code used in this work are also available on GitHub at `https://github.com/multiOmicMechanismAwareML/CodeBase`, along with the information for replicating the results presented.

ACKNOWLEDGMENTS. CC was supported by the UKRI CDT in Machine Intelligence for Nano-electronic Devices and Systems [EP/S024298/1]. CA received funding from BBSRC, grants CBMNet-PoC-D0156 and NPRONET-BIV-015 (BB/L013754/1). GZ and CA were also supported by Teesside University, and by UKRI Research England's THYME project.

1. Niedenführ S, Wiechert W, Nöh K (2015) How to measure metabolic fluxes: a taxonomic guide for 13c fluxomics. *Current opinion in biotechnology* 34:82–90.
2. Libbrecht MW, Noble WS (2015) Machine learning applications in genetics and genomics. *Nature Reviews Genetics* 16(6):321.
3. Li Y, Wu FX, Ngom A (2016) A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics* 19(2):325–340.
4. Shaked I, Oberhardt MA, Atias N, Sharan R, Ruppin E (2016) Metabolic network prediction of drug side effects. *Cell systems* 2(3):209–213.
5. Kim M, Rai N, Zorraquino V, Tagkopoulos I (2016) Multi-omics integration accurately predicts cellular state in unexplored conditions for escherichia coli. *Nature communications* 7:13090.
6. Yaneske E, Angione C (2018) The poly-omics of ageing through individual-based metabolic modelling. *BMC bioinformatics* 19(14):415.
7. Yang JH, et al. (2019) A white-box machine learning approach for revealing antibiotic mechanisms of action. *Cell* 177(6):1649–1661.
8. Culley C, Vijayakumar S, Zampieri G, Angione C (2019) Combining metabolic modelling with machine learning accurately predicts yeast growth rate in *11th International Workshop on Bio-Design Automation*.
9. Guebila MB, Thiele I (2019) Predicting gastrointestinal drug effects using contextualized metabolic models. *PLoS computational biology* 15(6):e1007100.
10. Zampieri G, Vijayakumar S, Yaneske E, Angione C (2019) Machine and deep learning meet genome-scale metabolic modeling. *PLoS computational biology* 15(7):e1007084.
11. Yu R, Nielsen J (2020) Yeast systems biology in understanding principles of physiology underlying complex human diseases. *Current Opinion in Biotechnology* 63:63–69.
12. Levy S, Barkai N (2009) Coordination of gene expression with growth rate: a feedback or a feed-forward strategy? *FEBS letters* 583(24):3974–3978.
13. Pacheco MP, Bintener T, Sauter T (2019) Towards the network-based prediction of repurposed drugs using patient-specific metabolic models. *EBioMedicine* 43:26–27.
14. Bao Z, et al. (2018) Genome-scale engineering of saccharomyces cerevisiae with single-nucleotide precision. *Nature biotechnology* 36(6):505.
15. Gardner TS (2013) Synthetic biology: from hype to impact. *Trends in biotechnology* 31(3):123–125.
16. David F, Siewers V (2015) Advances in yeast genome engineering. *FEMS Yeast Res* 15(1):1–14.
17. Shahrezaei V, Marguerat S (2015) Connecting growth with gene expression: of noise and numbers. *Current opinion in microbiology* 25:127–135.
18. De Jong H, et al. (2017) Mathematical modelling of microbes: metabolism, gene expression and growth. *Journal of The Royal Society Interface* 14(136):20170502.
19. Herrgård MJ, et al. (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature biotechnology* 26(10):1155–1160.
20. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T (2010) Interdependence of cell growth and gene expression: origins and consequences. *Science* 330(6007):1099–1102.
21. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? *Nature biotechnology* 28(3):245.
22. Chen Y, Li G, Nielsen J (2019) Genome-scale metabolic modeling from yeast to human cell models of complex diseases: latest advances and challenges in *Yeast Systems Biology*. (Springer), pp. 329–345.
23. Pelechano V, Pérez-Ortín JE (2010) There is a steady-state transcriptome in exponentially growing yeast cells. *Yeast* 27(7):413–422.
24. Airoldi EM, et al. (2009) Predicting cellular growth from gene expression signatures. *PLoS Computational Biology* 5(1):e1000257.
25. Wytock TP, Motter AE (2019) Predicting growth rate from gene expression. *Proceedings of the National Academy of Sciences* 116(2):367–372.
26. Slavov N, Botstein D (2011) Coupling among growth rate response, metabolic cycle, and cell division cycle in yeast. *Molecular biology of the cell* 22(12):1997–2009.
27. Kemmeren P, et al. (2014) Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* 157(3):740–752.
28. Sameith K, et al. (2015) A high-resolution gene expression atlas of epistasis between gene-specific transcription factors exposes potential mechanisms for genetic interactions. *BMC biology* 13(1):112.
29. Chowdhury R, Chowdhury A, Maranas CD (2015) Using gene essentiality and synthetic lethality information to correct yeast and cho cell genome-scale models. *Metabolites* 5(4):536–570.
30. Broach JR (2012) Nutritional control of growth and development in yeast. *Genetics* 192(1):73–105.
31. Kondo M, et al. (2000) The rate of cell growth is regulated by purine biosynthesis via atp production and g1 to s phase transition. *The Journal of Biochemistry* 128(1):57–64.
32. García-Martínez J, et al. (2015) The cellular growth rate controls overall mrna turnover, and modulates either transcription or degradation rates of particular gene regulons. *Nucleic acids research* 44(8):3643–3658.
33. Blank HM, Gajjar S, Belyanin A, Polymenis M (2009) Sulfur metabolism actively promotes initiation of cell division in yeast. *PloS one* 4(11):e8018.
34. Boer VM, Crutchfield CA, Bradley PH, Botstein D, Rabinowitz JD (2010) Growth-limiting intracellular metabolites in yeast growing under diverse nutrient limitations. *Molecular biology of the cell* 21(1):198–211.
35. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G (2008) Support vector machines and kernels for computational biology. *PLoS computational biology* 4(10):e1000173.
36. Huang S, et al. (2018) Applications of support vector machine (svm) learning in cancer genomics. *Cancer Genomics-Proteomics* 15(1):41–51.
37. Chen X, Ishwaran H (2012) Random forests for genomic data analysis. *Genomics* 99(6):323–329.
38. Ma J, et al. (2018) Using deep learning to model the hierarchical structure and function of a cell. *Nature methods* 15(4):290.
39. Guo W, Xu Y, Feng X (2017) Deepmetabolism: A deep learning system to predict phenotype from genome sequencing. *arXiv preprint arXiv:1705.03094*.
40. Gönen M (2012) Bayesian efficient multiple kernel learning in *Proceedings of the 29th International Coference on International Conference on Machine Learning*. (Omnipress), pp. 91–98.
41. Simon N, Friedman J, Hastie T, Tibshirani R (2013) A sparse-group lasso. *Journal of Computational and Graphical Statistics* 22(2):231–245.
42. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: Nsga-II. *IEEE transactions on evolutionary computation* 6(2):182–197.
43. Basu S, Kumbier K, Brown JB, Yu B (2018) Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences* p. 201711236.
44. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ (2018) Next-generation machine learning for biological networks. *Cell* 173(7):1581–1592.
45. Goh WWB, Wang W, Wong L (2017) Why batch effects matter in omics data, and how to avoid them. *Trends in biotechnology* 35(6):498–507.
46. Mi H, et al. (2019) Protocol update for large-scale genome and gene function analysis with the panther classification system (v. 14.0). *Nature protocols* 14(3):703–721.
47. Dever TE, Kinzy TG, Pavitt GD (2016) Mechanism and regulation of protein synthesis in saccharomyces cerevisiae. *Genetics* 203(1):65–107.
48. Zou S, Liu Y, Min G, Liang Y (2018) Trs20, trs23, trs31 and bet5 participate in autophagy through gtpase ypt1 in saccharomyces cerevisiae. *Archives of biological sciences* 70(1):109–118.
49. Larhlimi A, David L, Selbig J, Bockmayr A (2012) F2c2: a fast tool for the computation of flux coupling in genome-scale metabolic networks. *BMC bioinformatics* 13(1):57.
50. de Kroon AI (2017) Lipidomics in research on yeast membrane lipid homeostasis. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids* 1862(8):797–799.
51. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions in *Advances in neural information processing systems.* pp. 4765–4774.
52. Patton-Vogt J, de Kroon AI (2020) Phospholipid turnover and acyl chain remodeling in the yeast er. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids* 1865(1).
53. Machado D, Herrgård M (2014) Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS computational biology* 10(4).
54. Opdam S, et al. (2017) A systematic evaluation of methods for tailoring genome-scale metabolic models. *Cell systems* 4(3):318–329.
55. Vijayakumar S, Conway M, Lió P, Angione C (2017) Seeing the wood for the trees: a forest of methods for optimization and omic-network integration in metabolic modelling. *Briefings in bioinformatics* 19(6):1218–1235.
56. Ray B, et al. (2014) Information content and analysis methods for multi-modal high-throughput biomedical data. *Scientific reports* 4:4411.
57. Palsson BØ (2015) *Systems Biology: Constraint-Based Reconstruction and Analysis.* (Cambridge University Press).
58. Angione C (2018) Integrating splice-isoform expression into genome-scale models characterizes breast cancer metabolism. *Bioinformatics* 34(3):494–501.
59. Heirendt L, et al. (2019) Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0. *Nature protocols* p. 1.
60. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436.
61. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M (2019) New approach for understanding genome variations in kegg. *Nucleic acids research* 47(D1):D590–D595.
62. Sánchez B, Li F, Lu H, Kerkhoven E, Nielsen J (2018) Sysbiochalmers/yeast-gem: yeast8.2.0.
63. O'Duibhir E, et al. (2014) Cell cycle population effects in perturbation studies. *Molecular systems biology* 10(6):732.