Systems biology

Multimodal regularised linear models with flux balance analysis for mechanistic integration of omics data

Giuseppe Magazzù [1], Guido Zampieri [1] and Claudio Angione [1,2,3,*]

[1]School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK
[2]Healthcare Innovation Centre, Teesside University, Middlesbrough, UK
[3]Centre for Digital Innovation, Teesside University, Middlesbrough, UK

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: High-throughput biological data, thanks to technological advances, have become cheaper to collect, leading to the availability of vast amounts of omic data of different types. In parallel, the *in silico* reconstruction and modelling of metabolic systems is now acknowledged as a key tool to complement experimental data on a large scale. The integration of these model- and data-driven information is therefore emerging as a new challenge in systems biology, with no clear guidance on how to better take advantage of the inherent multi-source and multi-omic nature of these data types while preserving mechanistic interpretation.

Results: Here we investigate different regularisation techniques for high-dimensional data derived from the integration of gene expression profiles with metabolic flux data extracted from strain-specific metabolic models to improve cellular growth rate predictions. To this end, we extend previous regularisation frameworks including group, view-specific and principal component regularisation, and experimentally compare them using data for 1,143 *Saccharomyces cerevisiae* strains. We observe a divergence between methods in terms of predictive accuracy and integration effectiveness based on the type of regularisation employed. In general, no method results superior in both aspects, highlighting a widespread limitation in this type of model.

Availability: All data, models, and code produced in this work are available on GitHub at `https://github.com/Angione-Lab/HybridGroupIPFLasso_pc2Lasso`.

Contact: C.Angione@tees.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Keywords: Machine learning, flux balance analysis, multi-omics, regularisation

1 Introduction

In the past two decades, technological advances have led to the production of enormous amounts of biological data of different types, named *omics*. Each of these data types represents a different facet of an organism and its functioning. Recently, fuelled by the increasing availability of biological information, machine learning and deep learning have proven to be capable of revealing hidden relationships and patterns (Libbrecht and Noble, 2015), otherwise impossible to highlight by human operators, due to the vast dimensionality of the data and complexity of its inner relationships. Owing to the great heterogeneity of these data, the development of multi-view learning methods in a biological setting has been promoted (Li *et al.*, 2016).

Unfortunately, even though the most recent machine learning methods can disclose some relationships among the omics data, important limitations remain. These technologies are mainly used as black boxes and, depending on their architecture, may also not be able to produce new knowledge on the underlying biological mechanisms. Moreover, limited

computing resources or data availability may hamper the application of advanced machine learning methods. In many situations, the use of appropriate linear models for high-dimensional data can hence be a preferable option.

To compensate for these limitations, the reconstruction of genome-scale models of metabolic systems is opening new avenues for injecting biological knowledge into data-driven models (Zampieri *et al.*, 2019). Constraint-based modelling (CBM) can be used to simulate metabolism on a cellular scale, providing further insights into the biological mechanisms underlying cell operation. Metabolic flux data generated *in silico* have been previously used with machine learning models (Shaked *et al.*, 2016; Kim *et al.*, 2016; Yaneske and Angione, 2018; Yang *et al.*, 2019).

In this work, we compare multi-view learning frameworks that utilise both transcriptomics data and strain-specific metabolic models to predict cellular growth of *Saccharomyces cerevisiae*, which is one of the main eukaryotic platforms for bio-industrial production (Castillo *et al.*, 2019). Understanding and controlling cellular growth is important in biotechnology for the development of efficient cell factories (Dikicioglu *et al.*, 2013; Lian *et al.*, 2018). Due to the common problem of interpretability encountered when using machine learning algorithms, here we focus primarily on those algorithms that are directly interpretable, thus being able to provide immediate biological clues to be further investigated.

Mathematical modelling techniques such as CBM have been developed to simulate the possible outcomes from organisms in different environmental conditions and genetic characteristics. All the models devised so far offer quantitative mechanistic representations of biomolecular processes, but to achieve accurate estimates they often require detailed knowledge on uptake rates from the environment. On the other hand, it is also possible to find correlations between gene expression and cell growth only through data-driven machine learning methods. Previous research focused on building linear predictive models for yeast growth (Airoldi *et al.*, 2009), and more recently machine learning both for *E. coli* and *S. cerevisiae* (Wytock and Motter, 2019). Metabolic activity in combination with machine learning techniques was taken into consideration and evaluated only lately (Culley *et al.*, 2020).

Here, we combine regularised statistical learning methods with flux balance analysis (FBA) for omic data integration, in a setting designed to exploit the partial information present in the two different views. The goal is to reveal what characteristics a model should have to take advantage of this heterogeneous information. Despite superior prediction accuracy recently observed for multimodal neural networks (Culley *et al.*, 2020), as noted above there may be several factors hindering their utilisation in some case studies. Moreover, the interest in combining regularised linear methods with fluxomics data could also be motivated by the enhancement of model interpretation in biological terms.

We investigate a range of regularisation techniques, proposing expansions of previous frameworks and empirically evaluating them on a common benchmark. To this end, we use a compendium of 1,143 single gene knockout yeast strain expression profiles to predict cell doubling rates. Fluxomics data are obtained through a parsimonious implementation of flux balance analysis (pFBA) using the transcriptomics data to generate strain-specific genome-scale metabolic models. The metabolic model-generated reaction flux rates are then added to the gene expression profiles as additional features.

We show that, in our setting, group and view-specific regularisation achieve higher performance than principal component regularisation, outperforming multimodal neural networks. On the other hand, the latter obtains a larger performance improvement when integrating transcriptomic and fluxomic data. Overall, our results demonstrate the competitiveness of multimodal regularised linear models compared to data-hungry methods in a multi-omic task using experimental and model-generated omic data. At the same time, it highlights the lack of a clearly superior method for effective and transparent omic data integration, further underlying the importance of a bespoke selection of both features and machine learning models for each case study.

2 Methods

2.1 Dataset

We used a transcriptomics dataset generated in a previous study (Kemmeren *et al.*, 2014), which contains two-channel microarray profiles for 1,484 single-gene deletion strains of *S. cerevisiae* during mid-log phase. The data were downloaded from the supplementary materials of a second study providing relative growth rates compared to the wild type for 1,312 of the same strains, expressed as $\log_2$ of the doubling times ratio (O'Duibhir *et al.*, 2014). The final gene expression dataset was composed of 1,143 samples, and is here denoted as TRSC. Pre-processing was applied separately on the fluxomics data (denoted as FLUX) and the gene expression profiles. For the fluxomics data, all the reaction fluxes for which the value was $< 10^{-7}$ for all the samples were discarded (negligible fluxes in all samples). All data were standardised, following a preliminary exploration of normalisation techniques including also 0-1 normalisation and log-normalisation. Finally, in addition to these two datasets, a third one was built by joining the previous two (TRSC + FLUX).

2.2 Genome-scale metabolic modelling

Genome-scale metabolic models (GSMMs) are mathematical models, representing almost all the known biochemical reactions and transmembrane transporters present in an organism. The reaction network is described by a stoichiometric matrix $\mathbf{S}$, whose entries are the stoichiometric coefficients that characterise every biochemical transformation. The reaction rates (i.e. the fluxes) are mass- and energy-balanced assuming metabolic steady-state (no change in metabolite concentration). We utilised the *i*Sce926 yeast GSMM, which includes 926 genes, 3494 reactions and 2223 metabolites (Chowdhury *et al.*, 2015). Among all the genes in the TRSC data, a total of 908 (98%) are present in our transcriptomics dataset.

Parsimonious flux balance analysis (pFBA).

We used pFBA to control the global metabolic activity through a L1-regularisation for maximising our objective, at the same time making the solution as sparse as possible. The complete optimisation problem with constraints that we have to solve is the following:

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

where $\mathbf{v}$ is the vector of reaction fluxes in the network, whereas $\mathbf{v}_{lb}$ and $\mathbf{v}_{ub}$ represent their lower and upper bounds respectively. These two constraints are the mathematical representation of the several genetic or environmental factors under which the cell has to operate, and give a context-specific metabolic model that is consistent with experimental data. $\mathbf{w}$ is a one-hot encoding vector identifying the biomass pseudo-reaction as the unique objective, whereas Θ is a function mapping the gene expression levels to the reaction rates (Angione, 2018) (see details in **Supplementary Material**). Finally, g_{max} is the maximal growth rate achievable under these conditions. To perform the optimisation of Equation 1, the COBRA toolbox 3.0 (Heirendt *et al.*, 2019) was used with the PDCO solver. The solutions provide steady-state flux levels for each yeast strain and every reaction in the *i*Sce926 GSMM.

2.3 Regularised linear models for omic data

The models that were investigated belong to two different categories of machine learning techniques: statistical learning algorithms and neural networks. From the former group, we decided to consider only regularised linear models (RLMs) due to their inherent interpretability. The following multi-view approaches on the original omic profiles were employed:

IPF-Lasso L1. Integrative Lasso with Penalty Factors (Boulesteix *et al.*, 2017) is a variation of Lasso (Tibshirani, 1996) that accounts for different modalities being used. Specifically, it uses penalty factors λ_m to weight the L_1 penalty applied to the m-th modality. The objective to minimise is thus

$$\sum_{i=1}^{n}\left(y_i - \sum_{m=1}^{M}\sum_{j=1}^{p_m} x_{ij}^{(m)}\beta_j^{(m)}\right)^2 + \sum_{m=1}^{M}\lambda_m\|\beta^{(m)}\|_1, \tag{2}$$

where M is the number of modalities, p_m the number of covariates of the m-th modality, β the regression coefficients and n the total number of samples. The rationale behind this approach is that each modality has, in general, a different proportion of relevant variables, hence each contribution is weighted differently.

IPF-Lasso L2. We extended the originally proposed IPF-Lasso algorithm, replacing the L_1 norm with an L_2 norm, which was not tested in the original paper.

pcLasso. Principal component Lasso is a variation of elastic net that biases the solution coefficient vector towards the leading singular vectors of the feature matrix (in case of grouped features, towards the leading singular vector of each matrix associated with a group) (Tay *et al.*, 2018). In other words, the solution is pushed towards the most important/identified pattern to improve prediction accuracy. The objective to minimise is the following:

$$\frac{1}{2}\left\|Y - \sum_{p=1}^{P} X_p\beta_p\right\|^2 + \lambda\|\beta\|_1 + \frac{\theta}{2}\sum_k \beta_k^T(V_k D_{d_{k_1}^2 - d_{k_j}^2} V_k^T)\beta_k, \tag{3}$$

where k is a non-overlapping group (fluxomics or transcriptomics data in this study), β_k is the subvector of β corresponding to group k, V_k are the right singular vectors of the columns of X corresponding to group k, and D is a diagonal matrix with entries $d_{k_1}^2 - d_{k_j}^2$, which are the singular values of the columns of X related to group k (the former associated with the leading singular vector).

pc2Lasso. We also modified pcLasso and tested a new version, which shrinks the vector of coefficients towards the first and the second singular vectors associated with the two largest singular values. In our implementation, the entries $d_{k_1}^2 - d_{k_j}^2$ are substituted with $\alpha_1 d_{k_1}^2 + \alpha_2 d_{k_2}^2 - 2d_{k_j}^2$, where $d_{k_2}^2$ is the second-largest singular value, while α_1 and α_2 represent the quantity of variance explained by the first and the second largest singular values respectively.

Group Lasso. Group Lasso is a variation of Lasso regression in which the model is forced to include or disregard entire groups of variables defined by the user (Yuan and Lin, 2006). Notwithstanding the similarity with IPF-Lasso, there are two main differences: first, the groups are defined by the user without necessarily following a strict logic such as the one regarding the modalities; second, the algorithm makes a binary choice for each group, i.e. whether to include it or disregard it. In biological applications, this strategy can be justified based on the relationships among genes (e.g. whether they code the same protein, or regulate the same genes). In this study, the groups were defined looking at the correlation among the data in both views separately (TRSC and FLUX), while the number of groups was chosen to encourage a larger granularity. This was set to 50 groups for the fluxes, already fairly correlated, and 500 for the transcriptomics data. We varied these two parameters but a greater number of groups would lead to non-significant clusters, while a smaller number would lose information about the potential aggregations. The R function *hclust*, with default parameters, was used to find the clusters. When using both data sources, we used the same groups we had already defined when using the sources separately. The minimisation problem we solved is

$$\frac{1}{2}\left\|Y - \sum_{j=1}^{J} X_j\beta_j\right\|^2 + \lambda\sum_{j=1}^{J}\|\beta_j\|_{K_j}, \tag{4}$$

where $\|\beta\|_{K_j} = (\beta^{-1}K_j\beta)^{\frac{1}{2}}$, while X_j and K_j are respectively a design matrix identifying a group of covariates and an associated kernel.

Hybrid Group-IPF Lasso. We developed a hybrid method to take into account both the two modalities and the possible relationships within each of them. In order to do so, we combined the L1 penalty of IPF-Lasso and the L2 penalty of Group Lasso on the two different omic levels. This occurs when choosing, as we did following the suggestion of the original paper, K_j as the identity matrix multiplied by the square root of the size of the group. We chose the same groups chosen for the Group Lasso algorithm to make a fair comparison of the methods. The objective to minimise is therefore

$$\sum_{i=1}^{n}\left(y_i - \sum_{m=1}^{M}\sum_{j=1}^{p_m} x_{ij}^{(m)}\beta_j^{(m)}\right)^2 + \sum_{m=1}^{M}\lambda_m\|\beta^{(m)}\|_1 + \sum_{j=1}^{J}\lambda_j\|\beta_j\|_{K_j}, \tag{5}$$

where $\lambda_j = 1$ for $i = 1, 2, 3, ..., J$ to reduce the computational burden.

Artificial Neural Networks. Artificial Neural Networks (ANNs) are models capable of approximating any function, provided they are endowed with enough layers and/or neurons. An ANN is composed by an input layer, an output layer and one or more hidden layers in between. Each layer is made up of neurons, which are linked to the neurons assembling the other layers of the network. When a neural network presents more than one hidden layer it is defined as a Deep Neural Network (DNN). Numerous architectures were devised and studied, optimising several hyperparameters (e.g. number of layers, learning rate, optimisation strategy) in order to choose the best neural network architecture among the possible ones. Additional information can be found in **Supplementary Material**.

We used a naive neural network as a regressor, defining its architecture via hyperparameter optimisation (described in subsection 2.4). We also trained an identical network with the output of a variational autoencoder (VAE) (Higgins *et al.*, 2017), optimising it in the same way, the idea being to investigate whether it would be possible to find a shared projection onto a common latent space for both the gene expression and the metabolic fluxes.

Multi-Modal artificial Neural Networks. Multi-Modal artificial Neural Networks (MMNN) are a particular type of ANNs devised for learning from different sources of information, in general involving the use of an independent network for processing each modality and then a further network for integrating the gathered information and producing an output. In order to make a fairer comparison between the RLMs and the neural networks, in this study we trained the architecture devised in (Culley *et al.*, 2020), which inherently works in our scenario. This network is composed of two individual networks (one for the fluxomics data and one for the transcriptomics one) whose outputs are then concatenated and further processed by another separate network.

2.4 Training and testing pipeline

We split the dataset into a training set and a test set, with a 80:20 ratio. Then, we defined a subset of the training set as validation set, we trained only on the training set, and we optimised the hyperparameters based on the performance of the model on the validation set. All methods and models, when applicable, were optimised applying extensive grid-search over the hyperparameters (details can be found in **Supplementary Material**). In case a grid-search would be too computationally expensive we applied a consistent number of random-search iterations. In the case of the neural networks, the number of iterations exceeded 100. Finally, the best combination of hyperparameters was used to train the final model to make predictions on the unseen test set.

The explored machine learning models were evaluated over several metrics: the mean squared error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2, \tag{6}$$

where model predictions $\hat{y_i}$ are compared with observed growth rates y_i across all the n samples of the test set; the mean absolute error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|\,; \tag{7}$$

the coefficient of determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{8}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

We also computed for each method the standard deviation of the error distribution as a further metric:

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-1}}, \tag{9}$$

where e_i is the difference between the prediction and the ground truth and $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$.

2.5 Feature relevance analysis

We used enrichment analysis to analyse and interpret our results. For the fluxes, we performed hypergeometric tests through the hygecdf MATLAB function, and applied it to those to which the algorithms had attributed relevant weights (the threshold was chosen so as to reduce the number of fluxes to an easily interpretable amount). For the genes, we resorted to a different type of analysis since the lack of annotations for the transcriptomics data did not lead to meaningful results. The findings of these analyses are presented in subsection 3.3.

We also examined the final models by inspecting directly the weights attributed to the input features. While this is straightforward with the RLMs, with the neural network we developed a specific method in order to quantify the relevance that each feature has to the final prediction (see **Supplementary Material** for details).

3 Results

The aim of this study was to explore the predictive ability of multimodal regularised linear models integrating experimental and simulated omics data, expanding the current landscape of methods (Figure 1). As a case study, we focussed on predicting the growth rate of *S. cerevisiae*. We used constraint-based modelling (CBM) of metabolism to extract metabolic information of yeast mutants in the exponential growth phase, employing transcriptomics information. We then compared regularised linear models to evaluate the performance on our dataset, and how accurately they overlap the findings previously known from biological experiments. We also used ANNs to better understand the advantages and drawbacks of using a less interpretable method with a high predictive potential.

3.1 Multi-omics prediction of cellular growth

We started from three state-of-the-art RLMs that were previously proposed for biological data analysis: Integrative Lasso with Penalty Factors (IPF-Lasso) (Boulesteix *et al.*, 2017), Group Lasso (Yuan and Lin, 2006) and Principal Component Lasso (pcLasso) (Tay *et al.*, 2018). As described in Section 2, we then introduced Hybrid Group-IPF Lasso, which accounts both for different omic domains and intra-domain biological modules. Moreover, we considered the use of a modified regularisation term for IPF-Lasso and pcLasso (Section 2) Overall, we therefore tested the following RLMs: (i) IPF-Lasso, both L1 and L2, (ii) pcLasso, (iii) pc2Lasso, (iv) Group Lasso, (v) Hybrid Group-IPF Lasso. As a benchmark, we considered artificial neural networks (ANN) and multi-modal artificial neural networks (MMNN).

All the above methods (apart from the hybrid method) were tested over datasets containing three different types of information: (i) only

Data	Method	MSE ($\cdot 10^{-2}$)	MAE ($\cdot 10^{-2}$)	R^2	σ_e
	Regularised Linear Models				
TRSC + FLUX	Group Lasso	0.680	6.32	0.78	0.214
	IPF-Lasso L1	0.577	5.76	0.81	0.212
	IPF-Lasso L2	**0.551**	**5.61**	**0.82**	**0.215**
	Hybrid Group	**0.570**	**5.75**	**0.81**	**0.213**
	pcLasso*	0.812	6.70	0.73	0.206
	pc2Lasso*	**0.702**	**6.29**	**0.77**	**0.209**
TRSC	Group Lasso	0.558	5.65	0.82	0.219
	IPF-Lasso L1	0.577	5.76	0.81	0.212
	IPF-Lasso L2	**0.544**	**5.61**	**0.82**	**0.216**
	pcLasso	1.00	7.25	0.67	0.205
	pc2Lasso	**0.837**	**6.68**	**0.72**	**0.207**
FLUX	Group Lasso	1.74	9.70	0.43	0.206
	IPF-Lasso L1	1.76	9.74	0.42	0.207
	IPF-Lasso L2	**1.76**	**9.75**	**0.42**	**0.207**
	pcLasso	1.73	9.74	0.43	0.191
	pc2Lasso	**1.72**	**9.72**	**0.43**	**0.191**
	Artificial Neural Networks				
TRSC + FLUX	ANN	0.664	6.10	0.78	0.209
	MMNN	0.631	5.96	0.79	0.214
TRSC	ANN	3.33	11.57	-0.02	0.182
	MMNN	0.633	5.98	0.79	0.212
FLUX	ANN	3.09	11.9	0.05	0.176
	MMNN	1.74	9.46	0.42	0.205

Table 1. Multi-view results across all dataset-algorithm combinations. Values in bold represent the scores of the methods proposed in this work. Asterisks indicate a statistically significant improvement for methods using TRSC and FLUX data over TRSC only, showing that some methods benefit more than others when fluxomics data is added to transcriptomics data as predictive features. The best performance is held by our modified version of IPF-Lasso with L2 penalty, which outperforms the other algorithms over almost all the comparison metrics.

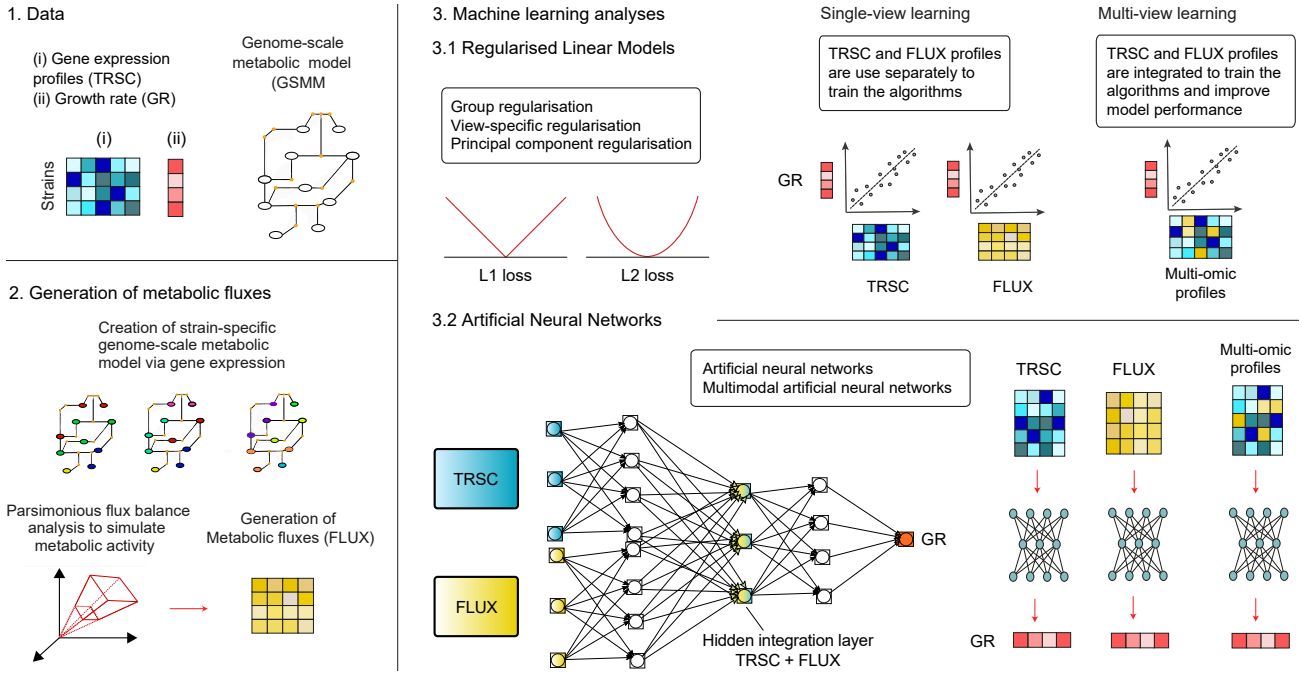

Fig. 1. General pipeline adopted. From 1,143 S. cerevisiae strains, the gene expression was used as a starting point (Kemmeren et al., 2014). A genome-scale metabolic model (GSMM) was then used (panel 1) to generate strain-specific GSMM models. From these GSMMs, metabolic fluxes were generated via parsimonious flux balance analysis (panel 2, see subsection 2.2). The machine learning methods were applied in two different settings: a single-view fashion and a multi-view one. In the former case, transcriptomics and fluxomics data were used separately as input for regularised linear models and artificial neural networks, while in the latter the two omics were integrated to let the two classes of methods leverage the different information of both sources (panel 3).

fluxomics data; (ii) only transcriptomics data; (iii) fluxomics and transcriptomics data. In the latter case, the integration was accomplished through either concatenation or previous elaboration by a beta-variational autoencoder architecture (β-VAE) (Higgins *et al.*, 2017), which learns a low-dimensional projection of the data in an unsupervised fashion. We also explored feature selection techniques prior to applying ANN models, but we did not proceed further as we obtained a performance decrease in all cases as also observed before (Culley *et al.*, 2020).

3.2 Comparison of multi-omics models of growth

For each of the described methods, when applicable, we fine-tuned the hyperparameters dividing the datasets into training, validation and test set. The algorithms were trained on the training set, while the selection of the hyperparameters was made using the validation set. The test set was used only to assess the ability of the algorithms with the best combination of hyperparameters for comparison. Figure 2 and Table 1 provide a detailed overview of the results. It can be noted that the performance based only on reaction fluxes is considerably lower than the performance based on gene expression, consistently to previous results (Culley *et al.*, 2020). This is likely to indicate that this source, when used in isolation, has a smaller amount of information compared to transcriptomics data, thus we considered the fluxes only for the comparison with the performance in a multi-view setting.

Amongst all the presented methods, only our proposed pc2Lasso managed to achieve an improvement in the performance when using more than one view, together with the original pcLasso and the MMNN. Instead, IPF-Lasso L1 fails to learn from the fluxes and the gene expression jointly. Indeed, its error scores remain unchanged when moving from one view to two and a Wilcoxon signed-rank test confirmed the overlap between their error distributions over the test set ($p = 0.19$). Moreover, we could have further confirmation of this by looking at the weights IPF-Lasso L1 attributes to the fluxes, which are all zeros.

Likewise, it is possible to gain some interesting insights by inspecting the weights IPF-Lasso L2 gives the fluxes and the transcriptomics data, albeit the method does not show actual improvement. While the weights of the fluxes are all zeros, the weights of the genes are significantly different from the weights the algorithm attributes when trained only on transcriptomics, and an even smaller amount of them is selected. This could be interpreted as a particular indirect form of regularisation that reaction fluxes apply over the gene expression with this algorithm, which suggests that this multi-modal approach utilises profitably metabolic modelling to gain information that cannot be acquired from the transcriptomics alone.

In general, the best-performing methods in the integration case (TRSC + FLUX), which adopt group and view-specific regularisation, do not display improved metrics over the TRSC case. On the other hand, methods employing principal component regularisation clearly display such improvement but remain with worse scores.

3.3 Interpretation of biological predictors

One of the purposes of adding a second view such as the metabolic fluxes was to improve the biological understanding and thus the interpretability of the input features, and consequently of the predicted output. From this perspective, we decided to look at the weights attributed to the metabolic fluxes by the algorithms and to conduct an enrichment analysis over the two different data types.

Regarding the weights, thanks to their transparent structure, RLMs can be interpreted immediately, as they directly assign a weight to each input feature. Since a typical characteristic of Lasso is the inner feature selection due to the fact that some input features are given zero as weight (which means that they are not used to make any prediction), all the RLMs share a similar property. Our analysis on the relevance of certain features

takes thus into account solely the features that are not disregarded by the methods (i.e. with a non-zero weight). Figure 2 illustrates the outcome for the most common pathways that were found enriched for the RLMs. In the case of the genes, the pathways most present in the pool of the selected genes were considered.

Looking exclusively at the metabolic fluxes that were given the highest weights by each method, it is possible to draw some parallels between the algorithms. All the algorithms, except IPF-Lasso L1, Group Lasso and our hybrid Group Lasso, select phenylalanine-involving reactions. Furthermore, all the algorithms except Group Lasso select tyrosine transaminase as a key reaction. It is widely known that in yeast these two compounds take part in the Ehrlich Pathway, which is directly related to fermentation. Moreover, it is noteworthy that while the two pcLasso versions find only one of the two PS decarboxylases, the two IPF-Lassos find the other one. Both these reactions have been found to support growth in *S. cerevisiae* (Griac, 1997). Finally, phosphatidyl-L-serine and phosphatidylethanolamine are once again, like tyrosine transaminase, common to all but IPF-Lasso L1, Group Lasso and the hybrid method. The former has been proven to be essential for cell growth (Kuge *et al.*, 1986) while the latter, under certain condition, takes on crucial importance for yeast growth (Kodaki and Yamashita, 1989).

Enrichment analyses were conducted to further validate the above findings. Due to the diverse nature of the two types of data, the analyses were performed in different ways based on the dataset considered. As regards the reaction fluxes, in addition to the importance of phenylalanine, tyrosine and tryptophan, the enrichment highlighted the relevance of cysteine and methionine as previously known (Sutter *et al.*, 2013; Yoshida *et al.*, 2011), 2-Oxocarboxylic acid and Lysine when considering IPF-Lasso and aminoacyl-tRNA synthesis, arginine, alanine aspartate and glutamate when looking at the results from pcLasso and pc2Lasso (the exact p-values can be found in **Supplementary Material**). We also made a comparison between the fluxes and the related genes, i.e. the genes associated with the enzymes that catalyse each reaction, selected by the algorithms in order to see whether there was a correspondence between them. Therefore, the genes associated with the reactions with the largest weights were considered and compared with the genes selected by the same method. The results showed that the genes associated with the selected reactions are not significantly present in the set of selected genes. This further strengthens our hypothesis for which fluxes and genes carry qualitatively different information, increasing the accuracy of a multimodal method compared to a single-view one.

4 Discussion

In this work, we proposed and tested approaches for multimodal learning, integrating information from metabolic models and experimentally-obtained gene expression data. We showed that the metabolic information represented by model-derived flux rates is relevant for interpreting the predictions from machine learning models, and for better understanding the interplay among genes, metabolism and growth. More specifically, we found that that multi-omics data integration through principal component regularisation leads to predictive improvements in our setting, while other forms of regularisation appear less effective in such task. While the metabolic fluxes were calculated through pFBA, it must be noted that there are several methodologies that can be used to compute flux rates, which can hence better exploit it and further improve the integration results.

Finally, we found that regularised linear models can outperform neural networks, without taking into account their intrinsic interpretability, even after extensive fine-tuning of the hyperparameters. This suggests that powerful methods such as neural networks cannot always be easily used as black boxes to improve the performance of a predictor, because their optimisation in this case would not be straightforward.

5 Conclusion

We investigated the predictive power of existing and novel multimodal regularised linear models in predicting *S. cerevisiae* growth using experimental and metabolic model-derived multi-omics data. Our experiments included state-of-the-art regularisation methods such as group-based, view-specific and principal component regularisations. These were applied to a combination of genome-wide gene expression data and model-generated metabolic information. In our task, we found that linear interpretable methods such as variations of Lasso can perform better than artificial neural networks even on a relatively large dataset.

We also observed that accurate state-of-the-art regularisation methods conceived for data integration fail in achieving accuracy improvements in our multi-omics setting, highlighting the need for new solutions that can exploit the cross-modal information, in addition to the information held in the individual modalities. Albeit still at the beginning of the exploration of this field, we believe that our results constitute a valuable benchmark for future investigation into multi-omics integration, again stressing the validity of linear regression methods in scenarios characterised by complex and high-dimensional data such as molecular biology.

Funding

This work was supported by Teesside University, by UKRI Research England's THYME project, and by a Children's Liver Disease Foundation Research Grant.

References

Airoldi, E. M., Huttenhower, C., Gresham, D., Lu, C., Caudy, A. A., Dunham, M. J., Broach, J. R., Botstein, D., and Troyanskaya, O. G. (2009). Predicting cellular growth from gene expression signatures. *PLoS Computational Biology*, **5**(1), e1000257.

Angione, C. (2018). Integrating splice-isoform expression into genome-scale models characterizes breast cancer metabolism. *Bioinformatics*, **34**(3), 494–501.

Boulesteix, A.-L., De Bin, R., Jiang, X., and Fuchs, M. (2017). Ipf-lasso: Integrative-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and mathematical methods in medicine*, **2017**.

Castillo, S., Patil, K. R., and Jouhten, P. (2019). Yeast genome-scale metabolic models for simulating genotype–phenotype relations. *Yeasts in Biotechnology and Human Health: Physiological Genomic Approaches*, page 111.

Chowdhury, R., Chowdhury, A., and Maranas, C. D. (2015). Using gene essentiality and synthetic lethality information to correct yeast and cho cell genome-scale models. *Metabolites*, **5**(4), 536–570.

Culley, C., Vijayakumar, S., Zampieri, G., and Angione, C. (2020). A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. *Proceedings of the National Academy of Sciences*.

Dikicioglu, D., Pir, P., and Oliver, S. G. (2013). Predicting complex phenotype–genotype interactions to enable yeast engineering: Saccharomyces cerevisiae as a model organism and a cell factory. *Biotechnology journal*, **8**(9), 1017–1034.

Griac, P. (1997). Regulation of yeast phospholipid biosynthetic genes in phosphatidylserine decarboxylase mutants. *Journal of bacteriology*, **179**(18), 5843–5848.

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., Haraldsdóttir, H. S., Wachowiak, J., Keating, S. M., Vlasov, V., *et al.* (2019). Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0. *Nature protocols*, page 1.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, **2**(5), 6.

Kemmeren, P., Sameith, K., van de Pasch, L. A., Benschop, J. J., Lenstra, T. L., Margaritis, T., O'Duibhir, E., Apweiler, E., van Wageningen, S., Ko, C. W., *et al.* (2014). Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, **157**(3), 740–752.

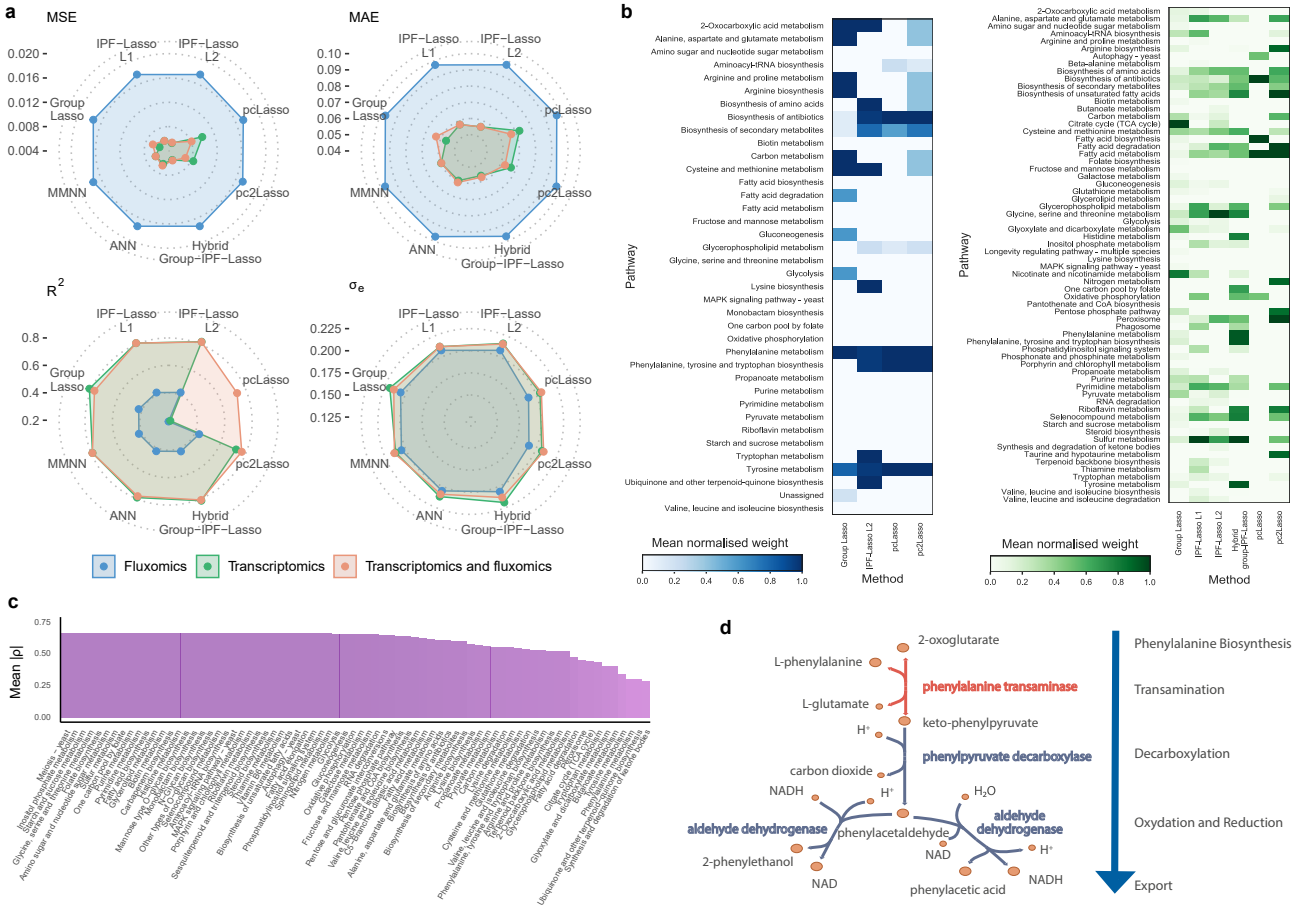

Fig. 2. (a) Comparison of RLMs and MMNN according to the evaluation metrics and the learning setting. The bigger the polygon drawn by the learning setting the worse the results for MSE, MAE and σ_e and the better for R^2. We show that the fluxomics data alone does not perform well for all the metrics (except for σ_e). On the other hand, for some methods, a combined learning with both transcriptomics and fluxomics data leads to better performance. (b) Average weight attributed to each of the related pathways according to the associated metabolic fluxes (left) and genes (right) for the regularised linear models. For better visualisation, we reported even the non-statistically significant pathways, and scaled the weights for each method separately. (c) Mean Pearson correlation along the pathways in the fluxomics dataset. The coefficients were computed calculating the absolute values of the Pearson correlation between each metabolic flux and the growth rate and then averaging them according to the pathways. (d) Ehrlich pathway for the catabolism of phenylalanine. The reaction in red is amongst the ones selected by IPF-Lasso with L2-norm as penalty. The main metabolites are represented by bigger circles.

Kim, M., Rai, N., Zorraquino, V., and Tagkopoulos, I. (2016). Multi-omics integration accurately predicts cellular state in unexplored conditions for escherichia coli. *Nature communications*, **7**, 13090.

Kodaki, T. and Yamashita, S. (1989). Characterization of the methyltransferases in the yeast phosphatidylethanolamine methylation pathway by selective gene disruption. *European journal of biochemistry*, **185**(2), 243–251.

Kuge, O., Nishijima, M., and Akamatsu, Y. (1986). Phosphatidylserine biosynthesis in cultured chinese hamster ovary cells. iii. genetic evidence for utilization of phosphatidylcholine and phosphatidylethanolamine as precursors. *Journal of Biological Chemistry*, **261**(13), 5795–5798.

Li, Y., Wu, F.-X., and Ngom, A. (2016). A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics*, **19**(2), 325–340.

Lian, J., Mishra, S., and Zhao, H. (2018). Recent advances in metabolic engineering of saccharomyces cerevisiae: New tools and their applications. *Metabolic Engineering*, **50**, 85–108.

Libbrecht, M. W. and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, **16**(6), 321.

O'Duibhir, E., Lijnzaad, P., Benschop, J. J., Lenstra, T. L., van Leenen, D., Koerkamp, M. J. G., Margaritis, T., Brok, M. O., Kemmeren, P., and Holstege, F. C. (2014). Cell cycle population effects in perturbation studies. *Molecular systems biology*, **10**(6), 732.

Shaked, I., Oberhardt, M. A., Atias, N., Sharan, R., and Ruppin, E. (2016). Metabolic network prediction of drug side effects. *Cell systems*, **2**(3), 209–213.

Sutter, B. M., Wu, X., Laxman, S., and Tu, B. P. (2013). Methionine inhibits autophagy and promotes growth by inducing the sam-responsive methylation of pp2a. *Cell*, **154**(2), 403–415.

Tay, J. K., Friedman, J., and Tibshirani, R. (2018). Principal component-guided sparse regression. *arXiv preprint arXiv:1810.04651*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267–288.

Wytock, T. P. and Motter, A. E. (2019). Predicting growth rate from gene expression. *Proceedings of the National Academy of Sciences*, **116**(2), 367–372.

Yaneske, E. and Angione, C. (2018). The poly-omics of ageing through individual-based metabolic modelling. *BMC bioinformatics*, **19**(14), 83–96.

Yang, J. H., Wright, S. N., Hamblin, M., McCloskey, D., Alcantar, M. A., Schrübbers, L., Lopatkin, A. J., Satish, S., Nili, A., Palsson, B. O., *et al.* (2019). A white-box machine learning approach for revealing antibiotic mechanisms of action. *Cell*, **177**(6), 1649–1661.

Yoshida, S., Imoto, J., Minato, T., Oouchi, R., Kamada, Y., Tomita, M., Soga, T., and Yoshimoto, H. (2011). A novel mechanism regulates h2s and so2 production in saccharomyces cerevisiae. *Yeast*, **28**(2), 109–121.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(1), 49–67.

Zampieri, G., Vijayakumar, S., Yaneske, E., and Angione, C. (2019). Machine and deep learning meet genome-scale metabolic modeling. *PLoS computational biology*, **15**(7), e1007084.

Multimodal regularised linear models with flux balance analysis for mechanistic integration of omics data

Supplementary Material

Giuseppe Magazzù[1], Guido Zampieri[1] and Claudio Angione[1,2,3,*]
[1] School of Computing, Engineering and Digital Technologies
Teesside University, Middlesbrough, UK
[2] Healthcare Innovation Centre, Teesside University, Middlesbrough, UK
[3] Centre for Digital Innovation, Teesside University, Middlesbrough, UK
* Corresponding author

August 16, 2020

1 Genome-scale metabolic modelling

Context-specific metabolic modelling. Each of the metabolic reactions is controlled by a specific combination of genes named gene sets. In a GSMM, the gene sets are represented using `AND`/`OR` operators. For example, if a reaction can be equally catalysed by two enzymes (namely, the two enzymes are *isozymes*), this relationship will be encoded through an `OR` operator between the two corresponding genes. Conversely, an `AND` relation identifies enzymatic complexes where both genes are necessary for the reaction to occur. GEMsplice [1] changes the reaction bounds by designating an effective gene expression value to each gene set. Such expression is obtained by converting the logical operations into maximum/minimum rules, according to the following map:

$$\begin{aligned} \Theta(g_1 \wedge g_2) &= \min\{\theta(g_1), \theta(g_2)\} \\ \Theta(g_1 \vee g_2) &= \max\{\theta(g_1), \theta(g_2)\}, \end{aligned} \tag{1}$$

where $\theta(g)$ represents the expression level of gene g and Θ represents the effective expression level of the gene set $\{g_1, g_2\}$. GEMsplice thus works as a further constraint inside the FBA optimisation. Following [2] and unlike its original version [3], we opted for the following map from gene set expressions Θ to reaction bounds $\mathbf{v}_{ub}$ and $\mathbf{v}_{lb}$:

$$\begin{aligned} \mathbf{v}_{ub} &\leftarrow \mathbf{v}_{ub}\, \Theta^{\gamma} \\ \mathbf{v}_{lb} &\leftarrow \mathbf{v}_{lb}\, \Theta^{\gamma}, \end{aligned} \tag{2}$$

where γ is a hyperparameter expressing the relevance of the gene expression in influencing the reaction bounds. We set $\gamma = 1$ according to [2], as this value minimises the linear correlation between predicted biomass accumulation rates and experimentally-available relative doubling times over all strains.

2 Interpretation of weights in neural networks

Let us consider a neural network with one-dimensional output and three hidden layers. Following the notation adopted in the paper, it is possible to describe it mathematically in the following way:

$$\begin{aligned} \mathrm{o} =& f(f(f(f(XW_1 + B_1)W_2 + B_2)W_3 + \\ & B_3)W_4 + B_o). \end{aligned} \tag{3}$$

Being almost all the activation functions currently used in research monotonic (included the ones used in the networks of interest in this study), and in view of the fact that only the relative importance of the features is of relevance for us, it is reasonable to ignore the functions and consider only the following expression

$$\begin{aligned} \mathrm{o} =& (((XW_1 + B_1)W_2 + B_2)W_3 + \\ & B_3)W_4 + B_o, \end{aligned} \tag{4}$$

from which, generalising, we can obtain that

$$\mathrm{o} = X \prod_{i=1}^{I} W_i + \sum_{j=1}^{I-1} B_j \prod_{k=j+1}^{I} W_k. \tag{5}$$

It is hence evident the fact that the weight influencing the input features is just the product of the weights that each linked neuron possesses.

3 Final neural network models

TRSC ANN. Selected hyperparameters: $\texttt{batch\_size} = 32, \texttt{epochs} = 2400, \texttt{learning\_rate} = 10^{-2}, \texttt{neurons\_first\_layer} = 3500, \texttt{neurons\_second\_layer} = 4000, \texttt{optimiser} = RPROP, \texttt{dropout} = 0.6, \texttt{loss} = Smooth\_L1$.

FLUX ANN. Selected hyperparameters: $\texttt{batch\_size} = 32, \texttt{epochs} = 400, \texttt{learning\_rate} = 10^{-5}, \texttt{neurons\_first\_layer} = 1200, \texttt{neurons\_second\_layer} = 1800, \texttt{optimiser} = SGD, \texttt{dropout} = 0.6, \texttt{loss} = Smooth\_L1$.

Table 1: List of nutrients allowed to be imported when performing flux balance analysis, together with their corresponding exchange reactions in the *i*Sce926 metabolic model [4]. These correspond to commonly used media [5, 6].

Medium component	**Exchange reaction name**	**Exchange reaction ID**
ammonium	ammonium exchange	r_1654
sulphate	sulphate exchange	r_2060
biotin	biotin exchange	r_1671
(R)-pantothenate	(R)-pantothenate exchange	r_1548
folic acid	folic acid exchange	r_1792
myo-inositol	myo-inositol exchange	r_1947
nicotinate	nicotinate exchange	r_1967
4-aminobenzoate	4-aminobenzoate exchange	r_1604
pyridoxine	pyridoxine exchange	r_2028
H+	H+ exchange	r_1832
riboflavin	riboflavin exchange	r_2038
thiamine(1+)	thiamine(1+) exchange	r_2067
sulphate	sulphate exchange	r_2060
potassium	potassium exchange	r_2020
phosphate	phosphate exchange	r_2005
sulphate	sulphate exchange	r_2060
sodium	sodium exchange	r_2049
L-alanine	L-alanine exchange	r_1873
L-arginine	L-arginine exchange	r_1879
L-asparagine	L-asparagine exchange	r_1880
L-aspartate	L-aspartate exchange	r_1881
L-cysteine	L-cysteine exchange	r_1883
L-glutamate	L-glutamate exchange	r_1889
L-glutamine	L-glutamine exchange	r_1891
glycine	glycine exchange	r_1810
L-histidine	L-histidine exchange	r_1893
L-isoleucine	L-isoleucine exchange	r_1897
L-leucine	L-leucine exchange	r_1899
L-lysine	L-lysine exchange	r_1900
L-methionine	L-methionine exchange	r_1902
L-phenylalanine	L-phenylalanine exchange	r_1903
L-proline	L-proline exchange	r_1904
L-serine	L-serine exchange	r_1906
L-threonine	L-threonine exchange	r_1911
L-tryptophan	L-tryptophan exchange	r_1912
L-tyrosine	L-tyrosine exchange	r_1913
L-valine	L-valine exchange	r_1914
oxygen	oxygen exchange	r_1992
adenine	adenine exchange	r_1639
uracil	uracil exchange	r_2090

Table 2: Hyperparameters spaces for the ANN and the β-VAE explored during Random Search. For not mentioned parameters, default values were used.

Method	**Hyper-parameters search space**
Artificial Neural Networks	`batch_size` $\in \{32, 64, 128\}$ `epochs` $\in \{400, 800, 1200, 1600, 2000, 2400\}$ `learning_rate` $\in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ `neurons_first_layer` = range depending on the input data `neurons_second_layer` = range depending on the input data `optimiser` $\in \{ADAM, SGD, RPROP, ADADELTA\}$ `dropout` $\in \{0, 0.3, 0.6\}$ `loss` $\in \{L1, MSE, Smooth\_L1\}$
β-Variational AutoEncoder	`batch_size` $\in \{32, 64, 128\}$ `epochs` $\in \{400, 800, 1200, 1600, 2000, 2400\}$ `learning rate` $\in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ `neurons_first_layer` = range depending on the input data `bottleneck_layer` = range depending on the input data `optimiser` $\in \{ADAM, SGD, RPROP, ADADELTA\}$ `dropout` $\in \{0, 0.3, 0.6\}$ `reconstruction_loss` $\in \{L1, MSE, Smooth\_L1\}$ `beta` $\in \{3, 4, 5\}$

Table 3: Flux Enrichment Analyses for all the regularised linear models. For each method we display the p-value associated to the pathway found (when present). As it can be noticed, phenylalanine- and tyrosine-related pathways are common to almost all the methods. All the p-values are below the defined threshold of 0.05. The results for pcLasso and the hybrid Group-IPF Lasso are not shown since the only enriched pathway for the former was the *Aminoacyl-t RNA biosynthesis*, with a p-value of $1.50 \cdot 10^{-2}$, while the latter was enriched in *Valine, leucine and isoleucine biosynthesis* with a p-value of $2.06 \cdot 10^{-2}$.

Pathway	IPF-Lasso L1	IPF-Lasso L2	pc2Lasso	Group Lasso
Phenylalanine, tyrosine and tryptophan biosynthesis	$1.33 \cdot 10^{-5}$	$1.52 \cdot 10^{-4}$	$9.30 \cdot 10^{-3}$	$1.79 \cdot 10^{-12}$
Phenylalanine metabolism	$1.79 \cdot 10^{-2}$	$8.21 \cdot 10^{-8}$	$9.30 \cdot 10^{-3}$	
Tyrosine metabolism	$4.71 \cdot 10^{-2}$	$1.52 \cdot 10^{-4}$	$9.30 \cdot 10^{-3}$	$2.74 \cdot 10^{-2}$
Biosynthesis of amino acids	$9.68 \cdot 10^{-4}$			$1.62 \cdot 10^{-7}$
Biosynthesis of antibiotics	$3.90 \cdot 10^{-3}$			$1.62 \cdot 10^{-7}$
Biosynthesis of secondary metabolites	$3.90 \cdot 10^{-3}$			$1.58 \cdot 10^{-4}$
Cysteine and methionine metabolism	$1.44 \cdot 10^{-2}$			
Aminoacyl-t RNA biosynthesis			$9.30 \cdot 10^{-3}$	
2-Oxocarboxylic acid metabolism	$1.45 \cdot 10^{-2}$			
Lysine biosynthesis	$1.45 \cdot 10^{-2}$			

References

[1] Claudio Angione. Integrating splice-isoform expression into genome-scale models characterizes breast cancer metabolism. *Bioinformatics*, 34(3):494–501, 2018.

[2] Christopher Culley, Supreeta Vijayakumar, Guido Zampieri, and Claudio Angione. A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. *Proceedings of the National Academy of Sciences*, 2020.

[3] Claudio Angione and Pietro Lió. Predictive analytics of environmental adaptability in multi-omic network models. *Scientific reports*, 5:15147, 2015.

[4] Ratul Chowdhury, Anupam Chowdhury, and Costas D Maranas. Using gene essentiality and synthetic lethality information to correct yeast and cho cell genome-scale models. *Metabolites*, 5(4):536–570, 2015.

[5] Yeast drop-out mix complete media. `https://www.usbio.net/media/D9515`, 2018. Accessed : 16/01/2018.

[6] Yeast nitrogen base (ynb) media. `https://www.usbio.net/media/Y2025`, 2018. Accessed : 16/01/2018.