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

Giuseppe Magazzù[1], Guido Zampieri[1,2] and Claudio Angione[1,3,4,*]
[1] School of Computing, Engineering and Digital Technologies
Teesside University, Middlesbrough, UK
[2] Department of Biology, University of Padova, Padova, Italy
[3] Healthcare Innovation Centre, Teesside University, Middlesbrough, UK
[4] Centre for Digital Innovation, Teesside University, Middlesbrough, UK
* Corresponding author

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 of a genome-scale model by assigning a gene expression value to each gene set, which then affects the lower and upper bound of the corresponding reactions. Such expression value 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 and hyperparameter choice

Let us consider a neural network with one-dimensional output and three hidden layers. Each node has a weight and a bias term, meaning that we can describe each layer in matrix notation with two matrices (W and B, the matrices of the weights and the biases respectively). If we indicate the input data as X and the output as o, then it is possible to describe it mathematically in the following way:

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

where f is the non-linear activation function. 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

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

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.

The following hyperparameters were selected as the best combinations for the 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: $\texttt{batch\_size} = 32, \texttt{epochs} = 400, \texttt{learning\_rate} = 10^{-5}, \texttt{neurons\_first\_layer} = 1200, \texttt{neurons\_second\_layer} = 1800, \texttt{optimiser} = SGD, \texttt{dropout} = 0.6, \texttt{loss} = Smooth\_L1$.

3 Supplementary Figures and Tables

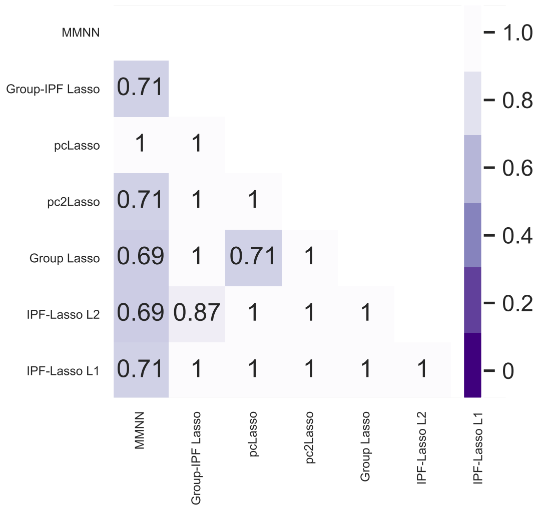


Figure S1: p-values from the Wilcoxon signed-rank test conducted for each couple of Regularised Linear Model and MMNN on the absolute error distribution, when using both the views.

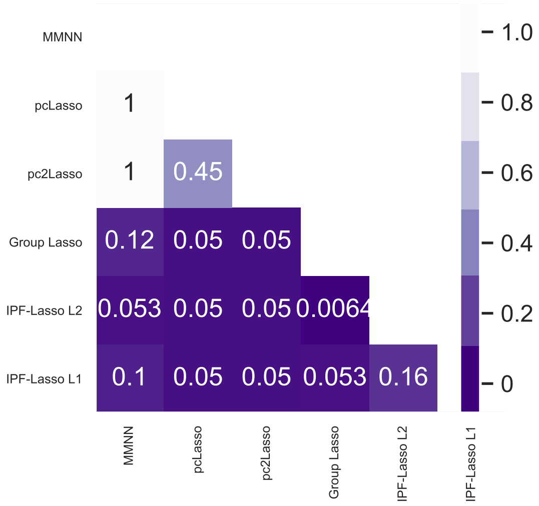


Figure S2: p-values from the Wilcoxon signed-rank test conducted for each couple of Regularised Linear Model and MMNN on the absolute error distribution, when using only fluxomic data.

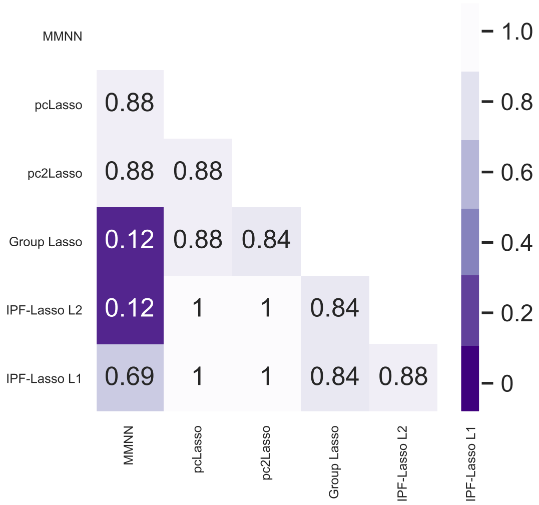
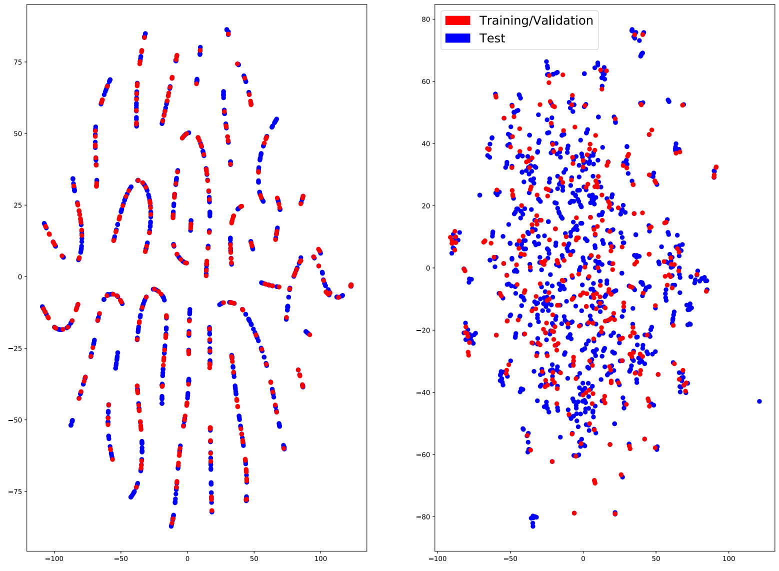


Figure S3: p-values from the Wilcoxon signed-rank test conducted for each couple of Regularised Linear Model and MMNN on the absolute error distribution, when using only gene expression.



Metabolic fluxes dataset

Gene expression dataset

Figure S4: t-SNE plots of the gene expression and fluxomic datasets (perplexity = 5). We conducted a post-mortem analysis to understand whether the two distributions of the training and test set were identical or not (here the split 70:30 is reported), which is the necessary condition so that our machine learning regression models could effectively generalise to unseen data. As shown, there is no significant distinction among the two distributions, thus reaffirming the legitimacy of our workflow.

Dataset	Average MSE ($\times 10^{-2}$)	Average MAE ($\times 10^{-2}$)	Average $\mathbf{R}^2$
70:30 split			
TRSC + FLUX	0.675	6.20	0.65
TRSC	0.848	6.98	0.72
FLUX	4.02	11.8	-0.33
80:20 split			
TRSC + FLUX	0.688	6.21	0.71
TRSC	0.854	6.83	0.76
FLUX	4.68	12.4	-0.34
90:10 split			
TRSC + FLUX	0.731	6.34	0.56
TRSC	0.785	6.90	0.76
FLUX	4.26	12.5	-0.31

Table S1: Robustness analyses for the ANN models with respect to the size of the dataset split. For each combination model/dataset, we ran 10 training-testing runs varying the split size (70:30, 80:20, 90:10), for a total of 30 runs for each model. The results are the averaged final scores.

Dataset	Average MSE ($\times 10^{-2}$)	Average MAE ($\times 10^{-2}$)	Average $\mathbf{R}^2$
70:30 split			
TRSC + FLUX	0.640	6.02	0.70
TRSC	0.679	6.18	0.64
FLUX	1.70	9.23	0.13
80:20 split			
TRSC + FLUX	0.658	6.03	0.75
TRSC	0.702	6.28	0.70
FLUX	2.00	9.97	0.24
90:10 split			
TRSC + FLUX	0.707	6.24	0.60
TRSC	0.786	6.52	0.54
FLUX	2.12	10.4	-0.11

Table S2: Robustness analyses for the MMNN models with respect to the size of the dataset split. For each combination model/dataset, we ran 10 training-testing runs varying the split size (70:30, 80:20, 90:10), for a total of 30 runs for each model. The results are the averaged final scores.

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 S3: 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].

Hyperparameter	**Hyperparameter search space**
`batch_size`	$\{32, 64, 128\}$
`epochs`	$\{400, 800, 1200, 1600, 2000, 2400\}$
`learning_rate`	$\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$
`neurons_first_layer`	range depending on the input data
`neurons_second_layer`	range depending on the input data
`optimiser`	$\{ADAM, SGD, RPROP, ADADELTA\}$
`dropout`	$\{0, 0.3, 0.6\}$
`loss`	$\{L1, MSE, Smooth\_L1\}$

Table S4: Hyperparameters spaces for the ANN explored during Random Search. For not mentioned parameters, default values were used.

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	$1.45 \cdot 10^{-2}$			

Table S5: 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 $2.06 \cdot 10^{-2}$.

Methods	**p-value**
Fluxomic data	
pcLasso & pc2Lasso	$1.09 \cdot 10^{-2}$
Transcriptomic data	
Hybrid Group-IPF Lasso & IPF-Lasso L2	$3.40 \cdot 10^{-2}$
Hybrid Group-IPF Lasso & IPF-Lasso L1	$1.79 \cdot 10^{-4}$
IPF-Lasso L1 & IPF-Lasso L2	$1.84 \cdot 10^{-6}$
IPF-Lasso L2 & pcLasso	$7.03 \cdot 10^{-5}$
IPF-Lasso L2 & Group Lasso	$1.86 \cdot 10^{-2}$
IPF-Lasso L1 & pc2Lasso	$7.26 \cdot 10^{-3}$
pc2Lasso & Group Lasso	$1.53 \cdot 10^{-2}$

Table S6: Spearman correlation among the methods reported in Figure 2 (b), computed for the average pathway weights. Only the statistically significant results were reported.

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.