Computers in Biology and Medicine 151 (2022) 106244

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Clinical stratification improves the diagnostic accuracy of small omics datasets within machine learning and genome-scale metabolic modelling methods☆

Giuseppe Magazzù [a], Guido Zampieri [a,b], Claudio Angione [a,c,d,*]

[a] School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, England, United Kingdom
[b] Department of Biology, University of Padova, Padova, Italy
[c] Centre for Digital Innovation, Teesside University, Middlesbrough, England, United Kingdom
[d] National Horizons Centre, Teesside University, Darlington, England, United Kingdom

ARTICLE INFO

Dataset link: https://github.com/Angione-Lab/Hepatoblastoma-Children-Classification

Keywords:
Machine learning
Metabolic modelling
Hepatoblastoma
Clinical stratification
Multi-omics
Cancer
Systems biology

ABSTRACT

Background: Recently, multi-omic machine learning architectures have been proposed for the early detection of cancer. However, for rare cancers and their associated small datasets, it is still unclear how to use the available multi-omics data to achieve a mechanistic prediction of cancer onset and progression, due to the limited data available. Hepatoblastoma is the most frequent liver cancer in infancy and childhood, and whose incidence has been lately increasing in several developed countries. Even though some studies have been conducted to understand the causes of its onset and discover potential biomarkers, the role of metabolic rewiring has not been investigated in depth so far.
Methods: Here, we propose and implement an interpretable multi-omics pipeline that combines mechanistic knowledge from genome-scale metabolic models with machine learning algorithms, and we use it to characterise the underlying mechanisms controlling hepatoblastoma.
Results and Conclusions: While the obtained machine learning models generally present a high diagnostic classification accuracy, our results show that the type of omics combinations used as input to the machine learning models strongly affects the detection of important genes, reactions and metabolic pathways linked to hepatoblastoma. Our method also suggests that, in the context of computer-aided diagnosis of cancer, optimal diagnostic accuracy can be achieved by adopting a combination of omics that depends on the patient's clinical characteristics.

1. Introduction

Systems biology is a branch of biology aiming at explaining biological entities and processes through the integration of experimental data and mathematical computational frameworks, thus joining aspects and approaches of theoretical and experimental biology [1,2]. The sub-field of constraint-based modelling, in particular, has allowed the research community to investigate changes in biological systems according to the medium, internal deficiencies and even single individuals, leading the way to precision-medicine approaches [3–5].

The application of systems biology offers mechanistic explanations of biological processes not fully understood so far [6]. Several successful results have been achieved by the adoption of this framework in cancer research [7–9]. However, the exploration of machine learning in conjunction with techniques from systems biology has only recently started to be explored, with clear opportunities for further development [10,11]. The flexibility of this framework is such that several different types of tasks can be solved, from machine learning-guided metabolic engineering [12] to the prediction of drug side-effects [13] and gene regulatory network reconstruction via transfer learning [14].

Hepatoblastoma is the most frequent epithelial liver tumour in infancy and childhood, with over 90% of cases diagnosed earlier than 4 years of age. This tumour is characterised by a high recurrence rate and metastatic aggressiveness, especially below this threshold age [15], which makes it paramount to be able to obtain an accurate prediction early on the onset of the disease. Additionally, its incidence is increasing in several developed countries. The recent development of

☆ We would like to acknowledge a Research Award from the Children's Liver Disease Foundation, grant number SG/2019/06/03, and a Network Development Award from The Alan Turing Institute, grant number TNDC2-100022.

* Corresponding author at: School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, England, United Kingdom.
E-mail address: c.angione@tees.ac.uk (C. Angione).

https://doi.org/10.1016/j.compbiomed.2022.106244
Received 15 June 2022; Received in revised form 7 October 2022; Accepted 22 October 2022
Available online 26 October 2022

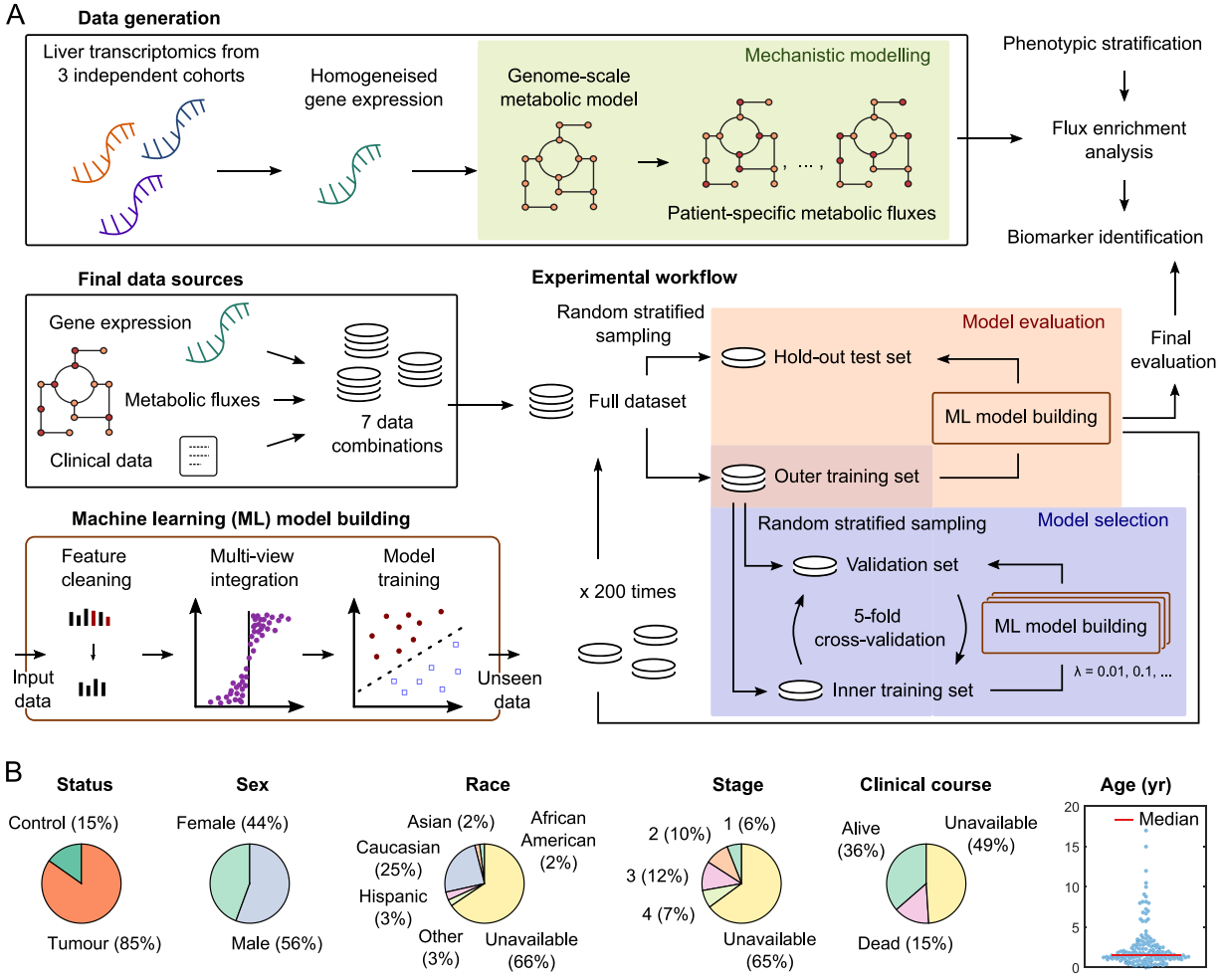


Fig. 1. A. Multi-omics and machine learning pipeline adopted in this study. Starting from liver gene expression profiles for hepatoblastoma patients and control subjects, we computed the associated genome-scale metabolic fluxes via FVA. For each of the combinations of transcriptomic, metabolic, and clinical data, we then performed a random stratified sampling to obtain a hold-out test set and an outer training set for machine learning model evaluation. Starting from this training set, we conducted a 5-fold cross-validation across hyperparameter values, and then evaluated the best model on the hold-out test set. Within each round of cross-validation, we also performed feature standardisation and cleaning and omics integration when necessary, in order to avoid any data leakage (brown box). We repeated the entire procedure 200 times to ensure the robustness of the results and re-ran the entire pipeline with a randomised dataset, whose phenotypes had been randomly permuted, so as to verify that the learned models correctly identified biologically meaningful patterns. B. Clinical data for the combined dataset used in the study. Race, tumour stage and clinical course had widespread missing entries, due to the original datasets having different information available.

molecular methods allowed extending the general subtype classification of primary childhood liver cancers, including hepatoblastoma [11,16], whose heterogeneity complicates the diagnosis of the disease. Moreover, clinical studies suggest that biomolecular mechanisms are associated with diverse prognostic outcomes and chemotherapy responses. Very recently, a few studies have started to explore the biological variability underlying hepatoblastoma, focusing on genomic biomarkers [17]. Likewise, machine learning has been adopted in the study of hepatoblastoma with encouraging results [18–20]. However, the role of metabolic rewiring – which is one of the main hallmarks of tumour cells [21] – has not been studied so far in hepatoblastoma. As a result, there is a general lack of robust biomarkers for this disease [22].

In this study, we investigate how different omics (and their combinations), interplaying with the patient's characteristics, affect the accuracy of a machine learning-based diagnosis by using a systems biology framework (flux balance analysis) in conjunction with machine learning. This concept has started to be explored only recently, and follows from recent work conducted on yeast metabolism [23,24]. Transcriptomics cannot be easily outperformed by other omic data, as observed by [25]. However, based on our experience [26], systems biology-derived data can inform mathematical models more comprehensively. We also examine metabolic markers for hepatoblastoma in the hope that this will guide future research in the field. In particular, we study how experimentally measured gene expression plays a role in diagnosing hepatoblastoma when paired with both synthetic *in silico* (i.e. simulated) metabolic data and clinical data such as gender and age of the patient.

Starting from a set of transcriptomic profiles, we use genome-scale metabolic modelling (GSMM) to estimate the associated metabolic activity across pathways in a sample-specific fashion. We then use support vector machines [27] as a predictor to identify hidden patterns that discriminate between phenotypic groups, and compare the performance of the different omics and their combinations, achieved by integrating the omics via Partial Least Squares Discriminant Analysis (PLSDA), in four alternative scenarios. For each scenario, we examine and present potential biomarkers, validating them against the existing literature. We report how specific omics combinations can be beneficial to the diagnosis of hepatoblastoma in different patients, and that the predictive power of each combination varies with their age, gender and clinical status.

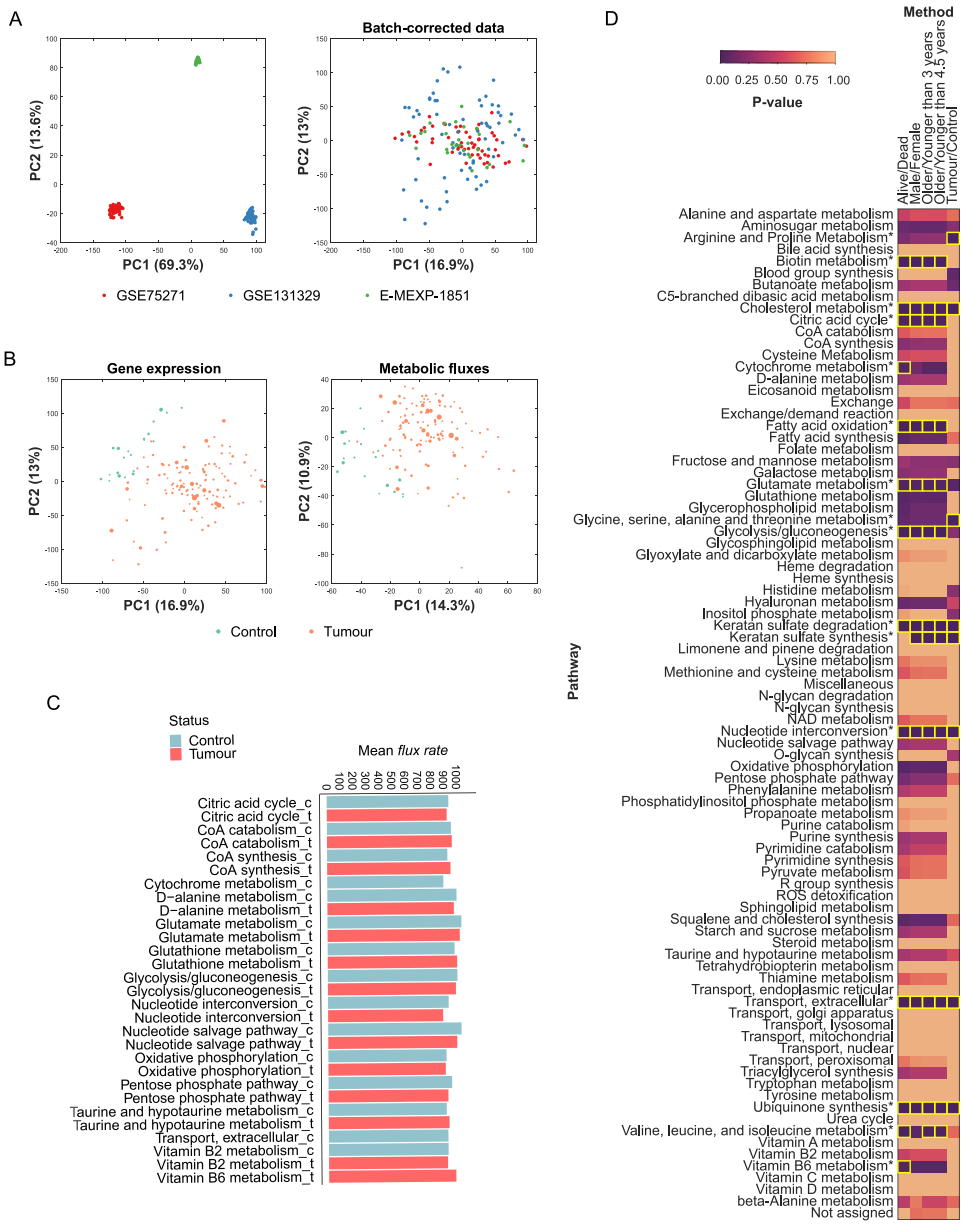


Fig. 2. A. Principal component visualisation of the three transcriptomic datasets considered. Upon batch correction through ComBat, the datasets correctly overlap, indicating that confounding experiment-specific variation has been reduced. B. Principal component visualisation of the aggregated cohort in terms of transcriptomic and fluxomic state, displaying the main phenotypic groups. The two groups appear circumscribed to well-defined areas of the principal component space for both omics across subjects, indicating that they describe distinct characteristics in the two groups. In contrast, no clear trend can be observed in terms of subject age, here represented by the circle size. An alternative representation of these graphs is Fig. E.9 in Appendix. C. Average flux in each pathway across patients and controls, obtained through FVA. Pathways associated with glutathione and CoA metabolism were found up-regulated, while the ones linked to nucleotide salvage D-alanine metabolism, and central metabolism were down-regulated. D. Flux enrichment analysis over the pathways in the genome-scale metabolic reconstruction for the flux rates from the FVA (maximal fluxes). Pathways with * and yellow contour are statistically significantly enriched in at least one stratification. In particular, extracellular transport, nucleotide interconversion, ubiquinone synthesis and keratan and cholesterol metabolism are the pathways enriched in all the stratifications. No significant difference was detected between the two age-based stratification enrichments.

2. Materials and methods

Data gathering and homogenisation

We gathered relevant transcriptomic data from liver samples of children diagnosed with hepatoblastoma and for control subjects within the same age range sets [17,28]. We selected three datasets whose gene expression profiles and clinical information have been retrieved from the Gene Expression Omnibus portal (www.ncbi.nlm.nih.gov/gds) under the accession numbers GSE75271, GSE131329 and from the BioStudies ArrayExpress portal (https://www.ebi.ac.uk/biostudies/arrayexpress) under the accession code E-MEXP-1851. The selection of these datasets considered the experimental platform utilised and, given the need for numerous samples to train a machine learning model, we prioritised the platform with the most abundant publicly-available data, which was in our case the Affymetrix microarray (Affymetrix Human Genome U133 Plus 2.0, Affymetrix Human Gene 1.0 ST and Affymetrix HG-U133 A 2.0 GeneChipTM respectively). The gathered data comprise a total of 151 subjects including 128 hepatoblastoma patients and 23 controls (see Fig. 1B). The average age is 2.6 years, while the sex distribution is 84 and 67 male and female subjects, respectively. The transcriptomic profiles cover 12,712 genes.

Appropriate high-throughput biological data pre-processing is fundamental to a meaningful analysis clear of technical biases [29–31].

Having gathered independently generated data, their joint analysis required ensuring uniformity and batch effect removal, which we performed through ComBat [32]. Fig. 2A shows the effect of homogenisation across the three datasets. Where possible, clinical data were also integrated and homogenised. Due to heterogeneous clinical formats, however, some information remained sparse, such as race, tumour stage and clinical course (Fig. 1B).

Patient-specific metabolic modelling of hepatoblastoma

To obtain metabolic information tailored to patient-specific metabolism, we adopted a GSMM approach. The base requirement is a mathematical representation of all the known biochemical reactions and transmembrane transporters present in an organism. Previous work has been done with GSMM to mechanistically characterise various human disorders, including liver diseases [33] and a range of cancer types [34–37]. GSMMs can be integrated with omics data to obtain context-specific models, representing the metabolic status across various conditions or tissues [38,39]. Notably, tissue- and cell-specific metabolic models have been successfully used to identify, and successively validate, specific drug targets that inhibit cancer proliferation but do not affect normal cell proliferation [40,41]. Through the mathematical manipulation of metabolic networks, GSMM can provide mechanistic insights regarding how hepatoblastoma works, with both the biochemical detail and completeness to interpret large molecular datasets.

Transcriptomics data integration

In our experiments, the human metabolic reconstruction Recon2.2 [42] was used in order to estimate the metabolic activity associated with transcriptional cues in tumour and control liver. Following a precision medicine approach, we derived a different metabolic model for each patient [43,44]. In doing so, we mapped the gene expression levels of the patients onto the metabolic network, thus determining the metabolic conditions from which to infer the reaction activity for each individual. Specifically, this process uses gene–protein–reaction relationships encoded within Recon2.2 and generates sample-specific constraints that describe the maximal and minimal activity that can be sustained by a given transcriptional state:

$$\begin{aligned} \mathbf{v}_{ub} &\leftarrow \mathbf{v}_{ub}\,[1+\gamma|\log\Theta|]^{\mathrm{sign}(\Theta-1)} \\ \mathbf{v}_{lb} &\leftarrow \mathbf{v}_{lb}\,[1+\gamma|\log\Theta|]^{\mathrm{sign}(\Theta-1)} \end{aligned} \tag{1}$$

where $\mathbf{v}_{ub}$ and $\mathbf{v}_{lb}$ represent the upper and lower bounds of the metabolic fluxes respectively, while Θ represents the gene expression level of the gene sets present in the genome-scale metabolic model, and $\gamma = 2$ (see Appendix A in the Appendix).

Furthermore, we imposed additional experimental constraints (see Table E.1 from Appendix E in the Appendix) directly onto the genome-scale metabolic model, which are orthogonal to those given by gene expression. To this end, we performed a literature search on liver metabolism, collecting experimentally supported bounds to metabolic exchanges in the liver. In particular, we followed previous work on hepatocyte modelling [33], correcting for the modelling convention according to which exchange reactions that assume uptakes are represented by negative lower bounds. These secretion and uptake rates were taken from previous measurements [45], which investigated the changes in intracellular pathway fluxes of primary rat hepatocytes in response to low-insulin preconditioning and amino acid supplementation. Among the involved reactions, we set uptake bounds for glucose, glutamate and glutamine.

We performed these steps for all the 151 samples in our dataset, in parallel, thus obtaining 151 context-specific metabolic models, each associated with a specific individual.

Flux variability analysis

To quantify the genome-scale metabolic state associated with collected transcriptomic profiles, we adopted flux variability analysis (FVA), which provides complete maximal (and minimal) cell metabolic capabilities across the biochemical network [46]. FVA operates by sequential maximisation and minimisation of each reaction activity to explore the boundaries of the feasible flux space. This algorithm yields a profile of maximal and minimal reaction rates (fluxes) for every biochemical reaction in Recon2.2, which collectively constitute a fluxomic profile. However, unlike the transcript levels, these metabolic fluxes do not belong to a single metabolic state, rather they represent the metabolic capabilities and limits of the individual's metabolic network, because the reactions are maximised and minimised in turn. While a transcriptomic profile represents the full set of transcription levels for a patient's genes in the liver, a fluxomic profile $\mathbf{v}$ is a set of reaction rates within a patient's liver metabolic network. The rationale behind this is that it should be possible to distinguish between healthy and cancer cells by looking at their metabolism. The optimisation problem was the following:

$$\begin{aligned} &\max\ (\min)\ v_i \\ &\text{subject to} \qquad \mathbf{c}^{\top}\mathbf{v} = f_{max}, \\ &\qquad\qquad\quad \mathbf{S}\,\mathbf{v} = 0, \\ &\qquad\qquad\quad \mathbf{v}_{lb} \le \mathbf{v} \le \mathbf{v}_{ub}, \quad \text{for } i = 1, 2, \ldots, n, \end{aligned} \tag{2}$$

where $\mathbf{S}$ is the stoichiometric matrix that defines the chemical reactions present in the metabolic model, $\mathbf{c}$ is a vector for characterising the objective function f starting from $\mathbf{v}$, f_{max} is the maximum value of f, and $\mathbf{v}_{lb}, \mathbf{v}_{ub}$ are the lower and upper bounds, respectively, of the metabolic reactions, as per Eq. (1).

As an alternative to FVA, we also used parsimonious flux balance analysis (pFBA) [47], following recent advances in mammalian metabolic modelling [48]. This approach, however, involves the adoption of specific cellular objectives that in this case did not provide sufficiently diversified metabolic profiles across all samples, which prompted us to employ FVA as it provides more unbiased estimates of metabolic variation across individuals. This could be explained by the fact that cancer cells present complex behaviour which may not be easily modelled with a single optimisation objective [49]. The interested reader can find more results regarding pFBA in Appendix B in Appendix. The COBRA Toolbox [50] was used with the Gurobi solver to compute the metabolic fluxes in MATLAB R2021b.

Biomarker identification framework

The study was divided into two parts: we first analysed the metabolism of the patients with respect to the possible stratifications in the population, and then applied machine learning techniques to determine possible biomarkers and how different omics could affect the precision of diagnosis of hepatoblastoma. We decided to follow this two-fold approach (flux-based metabolic analysis first and machine learning-led knowledge discovery after) as this is the most promising for the delivery of robust biomarker insights. Conversely, enrichment by itself does not guarantee predictive power nor does it help prioritise candidate biomarkers for future studies [51].

Flux enrichment analysis

To determine whether the over-represented pathways associated with the resulting metabolic reactions in the pool were overly present "by chance" or because of the existence of real biological mechanisms linked to the reactions, we decided to run a Flux Enrichment Analysis (FEA), which is a statistical testing technique that tests for the statistical relevance of biological pathways associated with a pool of reactions.

Before applying FEA, we removed all the reactions which had an absolute flux lower than 1e-7, considering them non-active, to account for the tolerance of the FVA solver. All the other reactions were instead

included in the analysis. FEA was conducted on all the samples, in a stratified and non-stratified way, by using hypergeometric tests, and the Benjamini–Hochberg correction was used to take into account the multiple hypothesis testing scenario. We set 0.05 as a threshold for the p-value to determine whether the presence of an over-represented pathway was statistically significant or not. Whenever specific covariate information was not available for a sample, we discarded the sample for that stratification and conducted the analysis on the remaining data.

We followed this approach because we were interested in assessing whether different groups of individuals (healthy/ill, male/female, etc…) showed changes in metabolic activity highly concentrated in specific pathways. The different cohorts were based on the available covariates and were organised as follows: tumour–control; male–female; older–younger than 4.5 years; older–younger than 3 years; alive–dead. The choice regarding the thresholds for the age was driven by the need for a deeper granularity in the analysis within the range [3, 4.5], which is considered to be critical to the diagnosis of the disease [15].

Machine learning-led biomarker discovery

Support vector machines (SVMs) are machine learning models that can be trained to distinguish samples belonging to different groups, such as patients and control individuals [27]. Here, we trained and applied SVM models to identify predictive variables that best discriminate between phenotypic groups (tumour and control). Once identified, these variables could thus be regarded as biomarkers. The objective function for the training of our SVMs was the following:

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^\top\mathbf{w} + \lambda \sum_i \max\Big(0, 1 - y_i(\mathbf{w}^\top x_i + b)\Big), \tag{3}$$

where λ is a regularisation hyperparameter to optimise, $\mathbf{w}$ and b, respectively, weights and bias of the model, and (x_i, y_i) the pair (features, class) of the ith sample. In addition to SVMs, we also tested another machine learning algorithm, Random Forest (RF) [52], and a Neural Network (NN), a deep learning approach that usually achieves state-of-the-art performance in many modern artificial intelligence tasks. The performance of the three models was compared and we found out that the SVM model performed better or equally well in all the studied scenarios. Further information can be found in Appendix C and E of Appendix. The choice of reporting the results for the SVMs only in the rest of the paper was also driven by the fact that the SVM algorithm is computationally inexpensive if compared, for instance, with the NNs.

As integrative approaches were also investigated, partial least squares discriminant analysis (PLSDA) was adopted in order to mitigate the problems deriving from the high dimensionality of the data combined with the small number of samples. In particular, the omics (transcriptomic and fluxomic) were projected onto two-dimensional spaces (one dimension per phenotypic trait; each omic was projected onto an independent space) in the explored integrative settings, explained below.

Our general training-evaluation pipeline, as reported in Fig. 1A, was the following: starting from the complete sample set (151 samples), we performed a random stratified sampling of 10 samples (5 patients and 5 controls) to put aside as a test set. The remaining samples were used as training data for an SVM model with a linear kernel, which we then employed to predict the phenotypic group for the 10 hold-out samples. This train–test process was repeated on random data partitions 200 times in order to ensure the robustness of the results, given such a small test set. The exact number of iterations was a result of a trial-and-error procedure, through which we determined that a lower number of repetitions would increase the standard deviation of the performance distributions (thus making our results less robust), while a higher number would simply increase the duration of the experiments, with negligible gains in terms of results robustness. Given the over-representation of tumour samples (see Fig. 1B), at each iteration we employed random under-sampling of tumour samples and over-sampling of control samples in order to obtain 30 samples for both groups (60 samples in total). We did this after the generation of the hold-out test sets, to avoid any data leakage that could affect the robustness of our pipeline. In other words, we randomly sampled, in a stratified fashion, 30+30 samples out of the 141 samples which did not belong to the test set. During the model building stage, we also performed feature standardisation and hyperparameter optimisation of λ through grid search. This, together with the under- and over-sampling of the 60 samples described above, was conducted within a 5-fold cross-validation framework on the remaining 141 samples, thus controlling for overfitting. The optimisation procedure for λ was selected for its robustness, but alternative approaches are possible. For instance, several meta-heuristics have been developed recently based on animal group behaviour and particle dynamics [53,54]. Such algorithms have previously been applied in combination with metabolic modelling [55], and it has been shown that they can be beneficial when optimising hyperparameters of SVM [56]. In this work, however, we opted for a more standard procedure that was applicable to all the investigated machine learning models.

Moreover, we performed feature selection by removing all the constant variables and the ones which were not unique (in the case of fluxomic data, for instance, reactions in a pathway with a locally linear topology could share the same value at all times). The feature selection procedure was itself performed within the cross-validation framework, in order to avert any overly optimistic performance evaluation of the SVM models during the hyperparameter optimisation.

We conducted these experiments in 6 different scenarios, with the aim of investigating how different combinations of omic data would influence the predictive power of the SVM models and their sensitivity to different biological entities (genes, fluxes, pathways): (i) use of transcriptomic data only; (ii) use of fluxomic data only; (iii) use of transcriptomic and clinical data (age, gender); (iv) use of fluxomic and clinical data; (v) use of transcriptomic and fluxomic data; (vi) use of transcriptomic, fluxomic and clinical data. To the above scenarios, we added also a final setting in which we trained the SVM models only with the clinical data, in order to eradicate any possible bias caused by the collection of the data (*sampling bias*).

To verify that the learned models correctly identified biologically meaningful patterns, we tested (through the same evaluation process) SVM models built starting from a permuted version of the dataset [57]. Specifically, we performed an additional 200 test iterations while randomly reassigning phenotypic labels to each sample prior to conducting the cross-validation, as previously suggested [58]. We did this for each of the 6+1 scenarios described above for completeness of the analysis.

Since we wanted to investigate how the discriminative power and sensitivity to biological mechanisms would change with different omics integrations, we decided to analyse the weights assigned by the SVMs to each variable during training, with the rationale that a higher weight in absolute value corresponds to a higher relevance. For the integrative experiments, the weights were computed by projecting the weights attributed to the latent dimensions back onto the original feature space. Moreover, in order to have a broader picture of the main metabolic pathways detected in the four integrative scenarios, we conducted FEA in each of them. For each scenario, we selected only the fluxes whose weight was in the 99.5th percentile. 0.05 was set as the threshold value for significance.

All the analyses were conducted in python, and the SVM and PLSDA algorithms were implemented with the library scikit-learn [59].

3. Results and discussion

The scope of this study was to investigate how different omics and their combinations may contribute to a computer-aided diagnosis of hepatoblastoma both in terms of accuracy and understanding of the biological mechanisms underlying the disease. In this framework, we focused on the use of individuals' transcriptomes and model-generated

fluxomic profiles in order to capture the metabolic alterations associated with the disease. These omics readouts were integrated and used to build predictive models through the machine learning pipeline displayed in Fig. 1A.

Genome-scale model characterisation of hepatoblastoma metabolism

Following a condition-specific modelling approach, we estimated the metabolic activity differences associated with varying transcriptional patterns across individuals. In brief, a genome-scale stoichiometric model of human metabolism was used as a platform for gene expression profiles obtained from three independent cohorts of individuals. As a result, we obtained maximal and minimal rates achievable through each biochemical reaction present in the model under the given transcriptional states. Fig. 2B shows a principal component analysis (PCA) of transcriptomic and fluxomic (maximal fluxes only) profiles. In both cases, hepatoblastoma patients and healthy controls display an almost linear separation. From a machine learning standpoint, this suggested that patient phenotypic classification could be achieved with high accuracy even with a limited number of samples. On the other hand, PCA revealed no obvious global relationship between subject age and multi-omics variation. An alternative graphical representation of Fig. 2B, in which age is replaced by gender, can be found in Fig. E.9 in Appendix.

The metabolic flux variation can be decomposed into metabolic capabilities across the pathways in the tumour and control groups described above, illustrated in Fig. 2C. From the figure, generated from the maximal fluxes, it is possible to observe a widespread reduced activity in several pathways associated with hepatoblastoma, such as in the central metabolism, nucleotide salvage and interconversion. However, up-regulation was found in glutathione and CoA metabolism.

We then used flux enrichment analysis (FEA) to obtain a picture of the most relevant metabolic pathways for groups of individuals defined based on their health status, sex, age, and clinical course. When doing so, FEA showed several statistically significant differences among the chosen cohorts (alive–dead, younger–older than 3 years, younger–older than 4.5 years, male–female, tumour–control). FEA computed over the maximal reaction fluxes generated by FVA returned several statistically significant differences. When considering the forward reaction direction, all the enrichments had in common many relevant pathways, since the reaction rates were generally higher, which meant many more active reactions in the metabolism (Fig. 2D). The generally enriched pathways were the ones associated with extracellular transport and nucleotide interconversion (as in the cases above), ubiquinone synthesis and keratan and cholesterol metabolism. The only exception to this was the alive–dead contrast, which did not present the reactions associated with the keratan sulphate synthesis. In all three cases, no significant differences were found across age groups (older–younger than 4.5 years and older–younger than 3 years), which probably indicates that within this age range there are no specific metabolic changes. When considering the backward direction of reversible reactions, all the enrichments had in common the reactions associated with extracellular transport and nucleotide interconversion as the most relevant, while the citric acid cycle and the nuclear transport reactions were not critical for the tumour–control comparison, unlike the other stratifications. Moreover, the two age stratifications and the alive–dead contrast showed as important reactions the ones related to the metabolism of valine, leucine and isoleucine.

Biochemical marker identification

The analyses above could identify changes in metabolic activity associated with a range of subject sub-cohorts. To understand which changes can be more strictly linked to carcinogenesis, we adopted machine learning techniques. Since the maximal fluxes presented more diversified metabolic profiles, we decided to focus only on them for the rest of the study. Fig. 3A shows the classification results obtained for all the 6+1 omics combinations studied, including gene expression, metabolic fluxes, and clinical information. On average, SVM models achieved a mean accuracy and MCC of around 0.9, with the exception of models trained only on clinical data. In contrast, SVM models obtained from permuted versions of the dataset on average proved no better than a random model for all the omics integrations, with a mean accuracy close to 0.6 and a mean Matthews correlation coefficient (MCC) around 0.1, and with a standard deviation much larger than the one of the models trained on the original dataset. This indicates that the original models captured meaningful patterns underlying the clinical state of the subjects, as expected.

In order to gain some insights regarding potential biomarkers for hepatoblastoma, we analysed the weights $\mathbf{w}$ from Eq. (3) given to the input features by the SVM models as a proxy for feature importance. Unlike previous work [26], we did not notice any regularisation-like phenomenon caused by the integration of fluxomic with transcriptomic data (see Figs. E.1–E.3 from Appendix E in the Appendix). However, the integration did redistribute the weights across the features in a varying way depending on the omics used.

Figs. 3D–E show the total weight each gene and reaction were given across the 200 iterations conducted. To facilitate comparison between the scenarios, the weights were quantile-normalised. The plot reveals that the most critical genes and reactions for subject classification vary depending on the data sources employed. In particular, the integrations highlighted as more relevant genes EPCAM, FRRS1L and ZBED8, whereas the base scenario with the SVMs trained only with gene expression had determined as more important the genes TMPRSS15, NPVF and HHLA2 (not relevant in the integrative experiments). It is interesting to note that none of these genes is associated with the reactions deemed relevant by the SVM models. EPCAM is a gene classified as tumour antigen on the database UniProt [60] while FRRS1L, more specifically, concerns the regulation of the glutamate receptor signalling pathway. The role of glutamate metabolism in hepatocytes is well-known and established [61]. Gene ZBED8 is instead a gene for which not much information has been collected yet, which suggests it might be involved in the metabolism of hepatoblastoma in an indirect way. Among the genes which were instead detected solely in the single-omic scenarios, TMPRSS15 is responsible for the activation of pancreatic proteolytic proenzymes, while NPVF is a neuropeptide and HHLA2 participates in the proliferation of T cells and regulation of cytokine production in lieu, which have a prominent role in inhibiting (but sometimes even stimulating) growth of cancer cells [62,63]. Among these, EPCAM, TMPRSS15 and HHLA2 can be found in blood samples [64], as also reported in The Human Protein Atlas database (https://www.proteinatlas.org) [65], whereas EPCAM and HHLA2 can be also found in urine samples [66,67].

When considering which reactions were deemed important by the SVM models, the results highlighted that reactions DASCBR (dehydroascorbate reductase, which participates in glutamate and ascorbate metabolism), RNMK (ribosylnicotinamide kinase, which is involved in the metabolism of nicotinamide adenine dinucleotide, a potential target for treating cancer [68]), AASAD3 m (L-aminoadipate-semialdehyde dehydrogenase, which participates in the production of lysine, whose acetylation is responsible for cancer development [69–71]) and EX_lys_L(e) (lysine exchange) were present in all the experimental scenarios.

When exploiting only the information contained in fluxomic data, the following reactions were identified as useful for the diagnosis of hepatoblastoma: LNS14DM (lanosterol 14-alpha-demethylase, which has shown to be able to decrease the proliferation of cancer cells [72]), G6PDA (glucosamine-6-phosphate deaminase), NAHCO3_HCLt (bicarbonate transport, that may be used in a diagnostic setting [73] but controversially in therapy [74,75]), THYMDtm (thymidine transport) and MI1PP (myo-inositol 1-phosphatase, regulating myo-inositol which can be used in cancer treatment [76]).

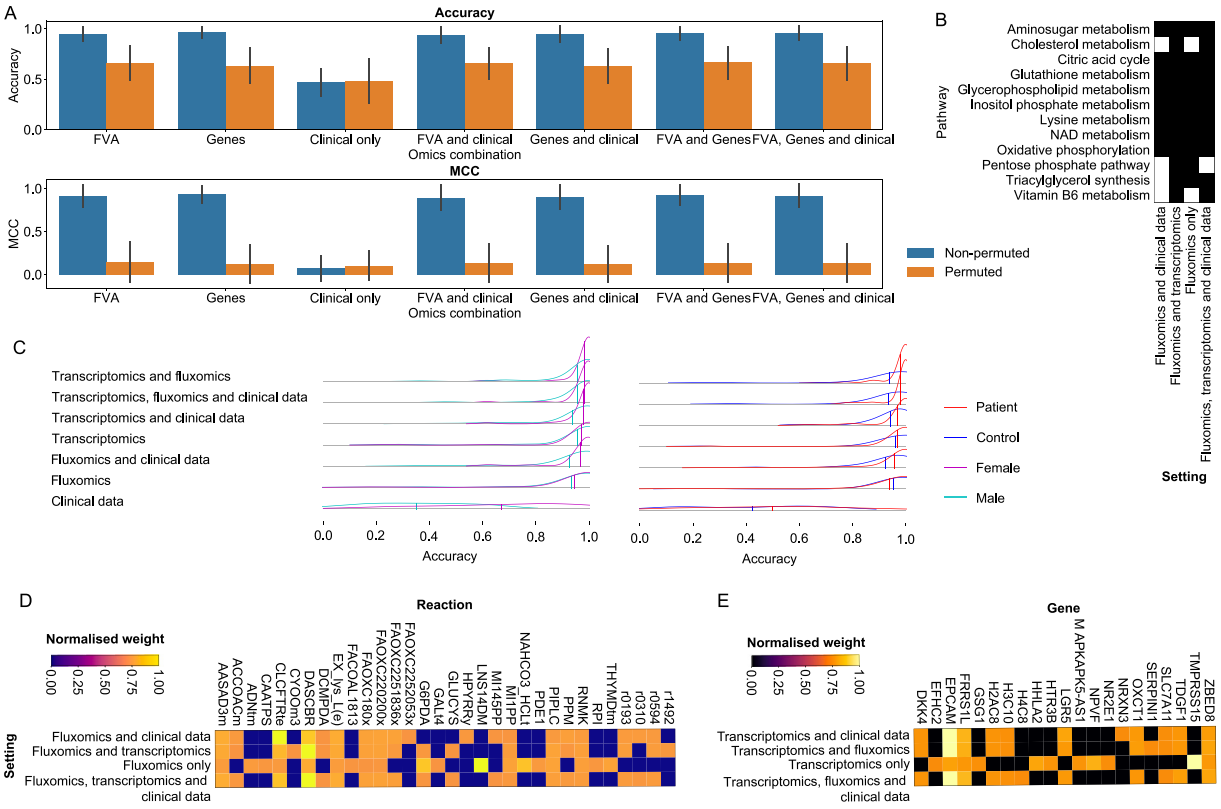


Fig. 3. A. Classification accuracy (top) and Matthews Correlation Coefficient (MCC, bottom) for SVM models that recognise tumour and control samples. Blue and orange bars respectively represent the performance of SVM models built using the original datasets and the same datasets with random sample labelling to phenotypic groups. The results with the original labels significantly outperform those with permuted labels, which approximate the performance of a random classifier. This proves further that our models are capable of learning from transcriptomic and fluxomic data the most relevant features which can be then used for biological interpretation. B. Statistically significantly enriched pathways from flux enrichment analysis across the four experimental settings. A black entry means that the pathway is significantly enriched in the cohort. When combining fluxomic and transcriptomic data enrichment returned more statistically significant pathways than in the other settings. C. Distribution of the accuracy of the SVM models trained, according to clinical data. Different groups of individuals can find beneficial a machine learning-aided diagnosis with different omics combinations. For female patients (in purple), all omics combinations tend to obtain a more accurate diagnosis. Moreover, if the patient has hepatoblastoma (in red), the best predictive performance can be achieved by integrating both transcriptomic and fluxomic data, while in the case of healthy control subjects, these two omics must be used separately to obtain a more accurate computer-aided diagnosis. Means are represented by vertical lines. D–E. Total weight attributed to reactions (D) and genes (E) in the four integrative scenarios. The weight distributions were first quantile-transformed so that they could be comparable, and the weights were then normalised in [0, 1]. An alternative representation of graphs D–E is available in Appendix E of Appendix.

Finally, the reactions that were useful for the prediction only in the integrative settings were CLCFTRte (CFTR chloride transport), FAOX∂ 00x and FAOXC180x (beta-oxidation of long-chain fatty acid). An alternative graphical representation of these plots, which allows for easier comparison between integrative scenarios, is provided in Figs. E.4–E.5 in Appendix.

When using FEA, we noticed that the enriched pathways were more stable across the four settings than the genes. In particular, we found that Vitamin B6 metabolism and Cholesterol metabolism were observed only when gene expression data was integrated with fluxomic data (regardless of the presence of clinical data in the integration), while the Pentose phosphate pathway was enriched only in the single-omic setting and when transcriptomics was integrated with fluxomics (but not in the presence of clinical data). Triacylglycerol synthesis was instead absent only when integrating the metabolic fluxes with clinical data, as opposed to Inositol phosphate metabolism, Aminosugar metabolism, Glutathione metabolism, Citric acid cycle, Glycerophospholipid metabolism, Lysine metabolism, NAD metabolism and Oxidative phosphorylation, which were found to be enriched in all scenarios. Overall, the integration of transcriptomic data with fluxomic data contributes to a greater number of enriched pathways. These results are summarised in Fig. 3B, and others showing the models' weights distributions are presented in Figs. E.1–E.3 from Appendix E in Appendix. In order to test for the robustness of our biomarker identification pipeline, we used a different approach to determine the weight of the biochemical features, obtaining similar results. These are presented in Appendix D of Appendix.

Relation between clinical data and diagnosis accuracy

We then asked if the constructed SVM models could be used to gain insights that can more accurately diagnose hepatoblastoma. We therefore analysed more in depth the trained models, in order to find directly applicable heuristics for guiding their use. In particular, we looked at how age, gender and health status could affect the predictive performance of different omics combinations.

In Fig. 3C, the accuracy distribution of the SVM models for different omics combinations is reported. It can be noticed that, across all omics combinations, female patients (in purple) tend to obtain a more accurate diagnosis (with the worst performance, in the case of a healthy subject, being above 94%, achieved by using only transcriptomic and clinical data) on average. On the other hand, in presence of a male, ill patient, the only use of transcriptomics will provide the best diagnostic performance. Similarly, the figure shows the desired property of our training and evaluation pipeline, namely the ability to discriminate with higher accuracy patients suffering from the tumour (in red).

Notably, when a patient has hepatoblastoma, the best predictive performance is achieved by integrating both transcriptomic and fluxomic data. Conversely, in the case of healthy control subjects, transcriptomic and fluxomic data separately represent the best two options for a correct computer-aided diagnosis. In both stratifications, as expected, the use of merely clinical data corresponds instead to trying to guess the phenotype of the individual, since there is no relation between age, gender and health status. This simple analysis allowed us to double-check that there were no spurious associations in the data due to their collection. Even though the information regarding clinical status cannot be exploited in a diagnostic setting, it is always possible to make use of other clinical information such as age and gender when choosing which omics combination to adopt for the diagnosis.

In particular, we investigated the performance of the SVM models with a more fine-grained detail to find less visible patterns in the performance distribution. Interestingly, we found that patients of age = 0.7 are much less likely to receive a correct diagnosis than patients of different ages (p-value < 0.001), while for patients of age = 7 the integrations of omics perform better than the use of single omics in general (p-value = 0.057). Finally, as a further addition to Fig. 3C, we found that omics integrations achieve overall better accuracy than single omics (p-value = 0.057) when the patient is female and has hepatoblastoma.

In general, both the characteristics of the dataset and biological factors might contribute to the above patterns. The highest accuracy observed for hepatoblastoma patients could be due to their over-representation in the dataset, which grants a more complete distribution for this class of subjects. Similarly, the omics integration might result more effective for the patients for this reason. In contrast, the higher discriminatory power found for the female patients cannot be explained in this way, given that no clear connection was seen between gender and health status (Fig. E.9). Besides, differences between genders in terms of omics accuracy could be underlain by specificities in developmental programming of growth and metabolism, which present sex differences not only in normal development but also in disease [77], and are linked to specific risk factors in childhood cancers [78]. Thus, critical aspects of metabolic rewiring in female subjects could be better captured by the GSMMs here developed, leading to better accuracy.

4. Conclusions and future work

In this study, we adopted an interpretable multi-omic framework to investigate molecular biomarkers and metabolic mechanisms in order to shed light on the onset of hepatoblastoma, potentially helping diagnose more accurately this disease in young patients. An aspect of this involved also the investigation of the sensitivity that such a method could have with respect to the characteristics of such patients, namely age, gender and the observed phenotypic trait. Moreover, we examined how different combinations of data can interplay with these and highlight different aspects of the metabolism of the patient. In particular, important genes as revealed by the integrations, are linked to cancer metabolism and hallmarks. Starting from gene expression profiles, we generated metabolic fluxes representing the metabolic state of the patients and, with the addition of clinical data, integrated all this information within a machine learning pipeline to determine whether an integrative approach could lead to an improvement in the diagnostic performance of our models.

We investigated genes, reactions and metabolic pathways by resorting to a feature importance approach in quest of potential biomarkers to guide future research in hepatoblastoma. We demonstrated that different omics combinations can achieve optimal predictive performance for different patients according to the patients' clinical data, even though the individual omics used can have a significantly skewed performance distribution [25], and that machine learning models can be endowed with different sensitivity to distinct biological entities based on the omics combinations they are trained on. Finally, we extracted novel mechanistic biomarkers whose study could be relevant in the research regarding the mechanisms underlying hepatoblastoma.

Overall, our study suggests that using systems biology approaches in conjunction with machine learning methods can provide valuable insights into the biological mechanisms of rare cancer conditions, for which omics data and biological knowledge are not as widely available as for the most commonly studied diseases (such as breast or lung cancer).

One of the limitations of our study is the practical impossibility, with current technology, to directly measure metabolic fluxes in human patients [79]. Here, we mitigated this by adopting genome-scale metabolic models, but these require some experimentally measured information such as transcriptomics data. Another potential limitation of our approach is the challenging direct applicability in some cases. Specifically, we have determined when to use which combination of omics, but this information is exploitable only in a limited number of cases, such as when the conditions determining the omics combination to use are based on clinical information such as gender or age (and, within this, only when we are within certain ranges). Yet, this work can serve as a guide for further research in hepatoblastoma, and the biomarkers found could potentially lead to the development of new diagnostic or therapeutic tools. Our approach has the advantage of elucidating how molecular entities can be related, and their importance in hepatoblastoma, with a granularity that is based on the patient's clinical information.

In the future, our work will focus on improving the omics integrative approaches, as well as investigating other omics data such as proteomics. Moreover, we will explore other phenotypic variations and try to determine whether some of the conclusions reached in this work are of general validity. For larger datasets, alternative optimisation methods based on heuristics will be investigated and adopted to improve the speed and quality of the training phase for the studied machine learning models [53,54].

CRediT authorship contribution statement

Giuseppe Magazzù: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Guido Zampieri:** Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Claudio Angione:** Conceptualization, Methodology, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

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

Acknowledgements

We thank Dr Jane Hartley, Consultant Paediatric Hepatologist at the Birmingham Women's and Children's NHS Foundation Trust, for inspiring discussions and advice regarding this work.

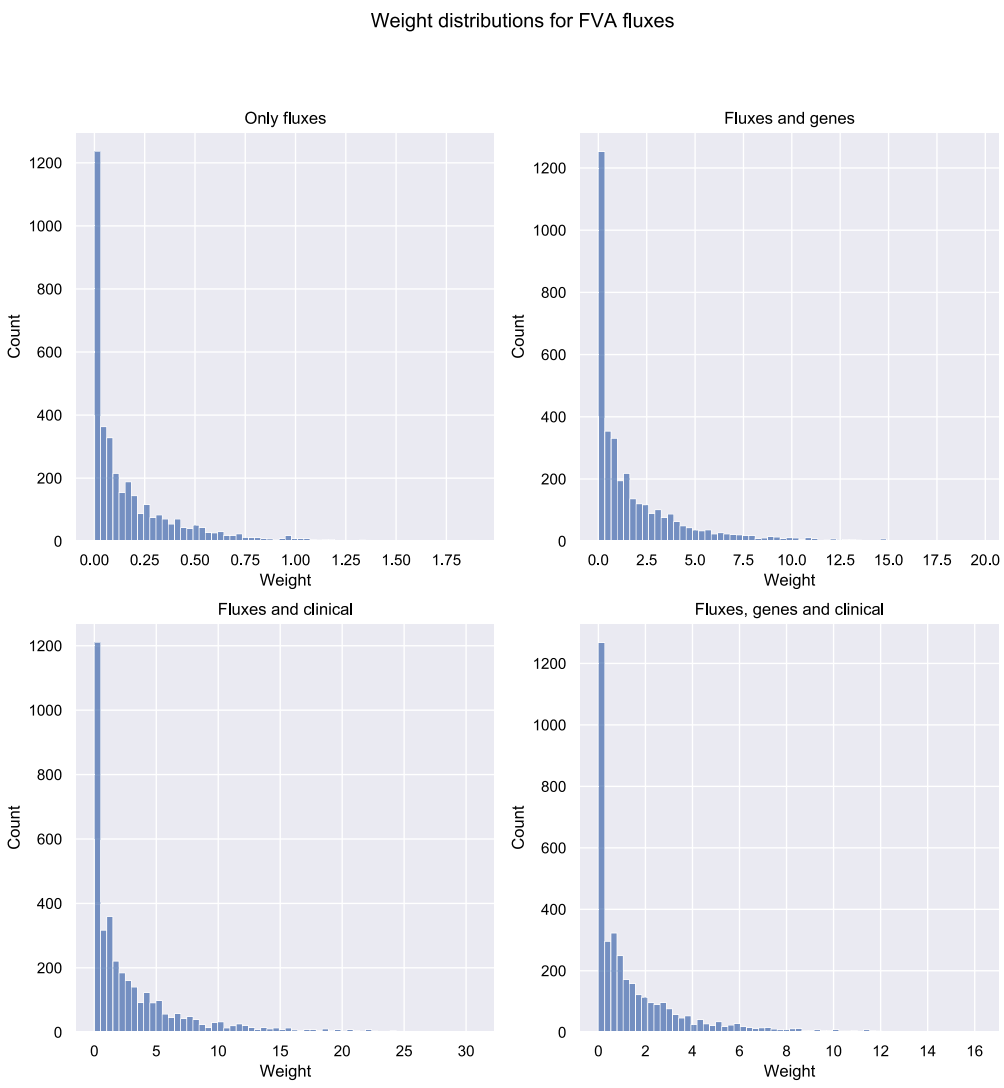


Fig. E.1. Weight distributions for metabolic reaction fluxes computed via FVA (Flux Variability Analysis) in the four different experimental settings. The shape of the distribution does not change in the various scenarios.

Appendix A. Context-specific metabolic modelling formulation

In a genome-scale metabolic model, the relation between a reaction and the genes that codify it is expressed in the form of gene-reaction rules. Such rules are formulae using logical operators `AND`, `OR` to combine different genes. The defined combination serves the purpose of determining the activity level of the reaction based on the gene expression level of the genes associated with it, which compose its *gene set*. Starting from a gene set composed of two genes, g_1 and g_2, whose respective gene expression levels are $\theta(g_1)$ and $\theta(g_2)$, we followed [44] and determined the gene expression level of the entire gene set by adopting the following mapping:

$$\begin{aligned} \Theta(g_1 \wedge g_2) &= \min\{\theta(g_1), \theta(g_2)\} \\ \Theta(g_1 \vee g_2) &= \max\{\theta(g_1), \theta(g_2)\}, \end{aligned} \tag{4}$$

Appendix B. Parsimonious flux balance analysis

We used an approach based on parsimonious Flux Balance Analysis (pFBA) [47] to study the metabolic network at a steady state, i.e. when the concentration of metabolites in liver cells can be approximated as constant over time [80]. We were therefore able to determine which metabolic fluxes (the rates of each metabolic reaction) were active and thus which metabolic pathways were predominant on a whole-cell scale.

Mathematical framework

The optimisation problem that was solved was formulated as:

$$\begin{aligned} &\min_{\mathbf{v}} \ \|\mathbf{v}\|_1 \\ &\text{subject to} \quad \mathbf{c}^\top \mathbf{v} = f_{max}\,, \\ &\qquad\qquad \mathbf{S}\,\mathbf{v} = 0\,, \\ &\qquad\qquad \mathbf{v}_{lb} \le \mathbf{v} \le \mathbf{v}_{ub} \end{aligned} \tag{5}$$

where f_{max} is the maximum value of the objective function f, and $\mathbf{v}_{lb}, \mathbf{v}_{ub}$ are the lower and upper bounds, respectively, of the metabolic reactions.

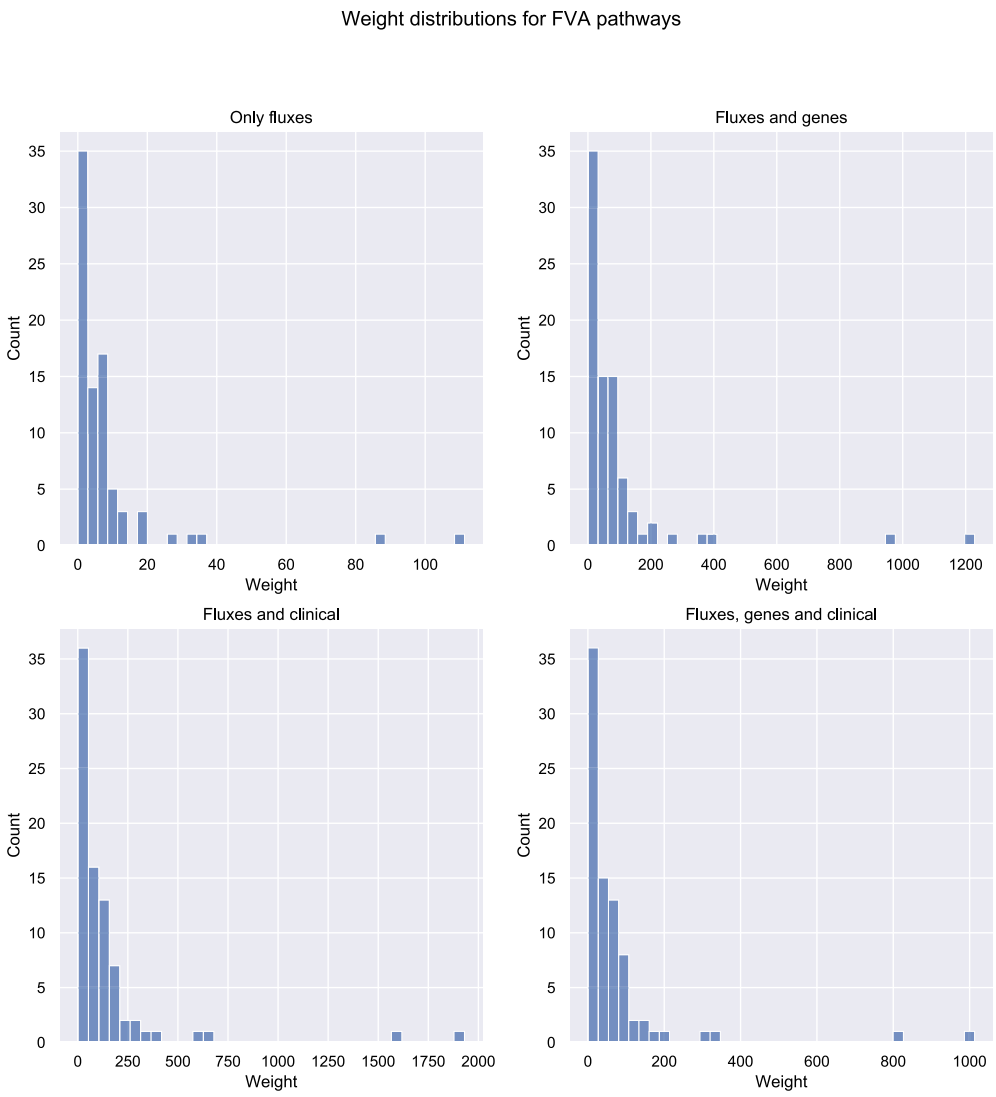


Fig. E.2. Weight distributions for metabolic pathways. The weights were computed in the experiments reported, i.e. when training the SVMs with reaction fluxes generated via FVA (Flux Variance Analysis). It is easy to notice that the shape of the distribution has not changed in the various settings.

Results

We conducted a Flux Enrichment Analysis (FEA) also on the fluxes generated via pFBA (using as objective function the maximisation of the biomass, thus simulating the uncontrolled growth of cancer cells). The enrichment for all the stratifications displayed as statistically relevant the reactions associated with nucleotide interconversion and glutamate metabolism. Furthermore, extracellular transport reactions were the most relevant for all the stratifications. All the enrichments but the one associated with the tumour–control stratification showed the importance of the reactions composing the citric acid cycle as well.

Appendix C. Machine learning models

In addition to the SVM models, in this study we adopted other two machine and deep learning models, namely Random Forests (RFs) and Neural Networks (NNs).

We compared their performances for each omics combination by using Wilcoxon signed-rank test. This statistical test was used to determine whether the accuracies and MCCs of the models belonged to an identical distribution or not. In particular, only the RF models presented performance distributions that were determined as different from the SVMs' ones, and exclusively when the models were trained with transcriptomic data only. In this case, the RF models performed slightly worse than the SVM models.

A graphical comparison of the performances of the three model types can be found in Appendix E, while for implementation details the reader is referred to Section 6.

Appendix D. Robustness of biomarker feature importance

In this study, we determined relevant biomarkers for hepatoblastoma by looking at the weights these were given by the trained SVM models.

In particular, we adopted two approaches for determining the usefulness of the available biomarkers: we aggregated the weights by summation over the 200 iterations of our training pipeline conducted,

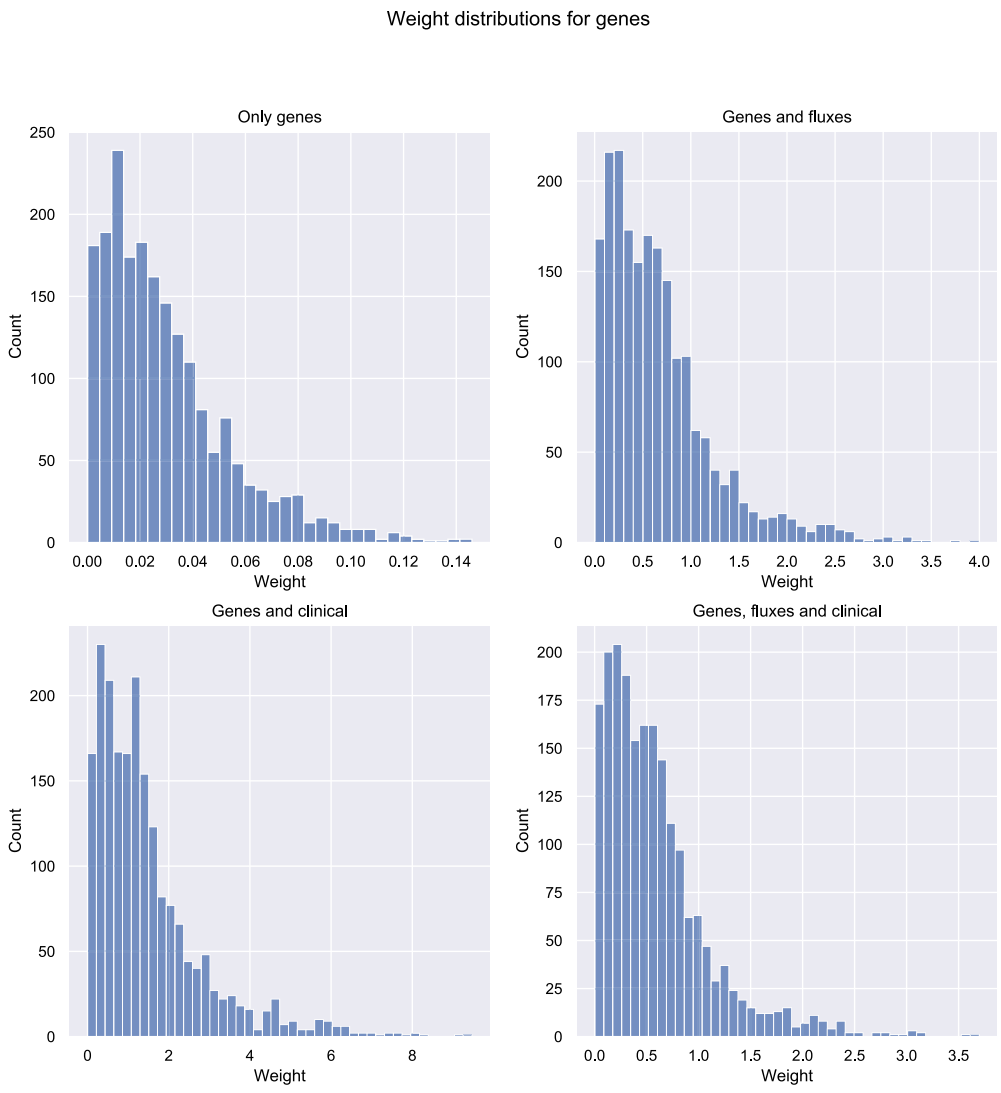


Fig. E.3. Weight distributions for the gene expression data in the four integration settings. The shape of the distribution does not significantly change in the various scenarios.

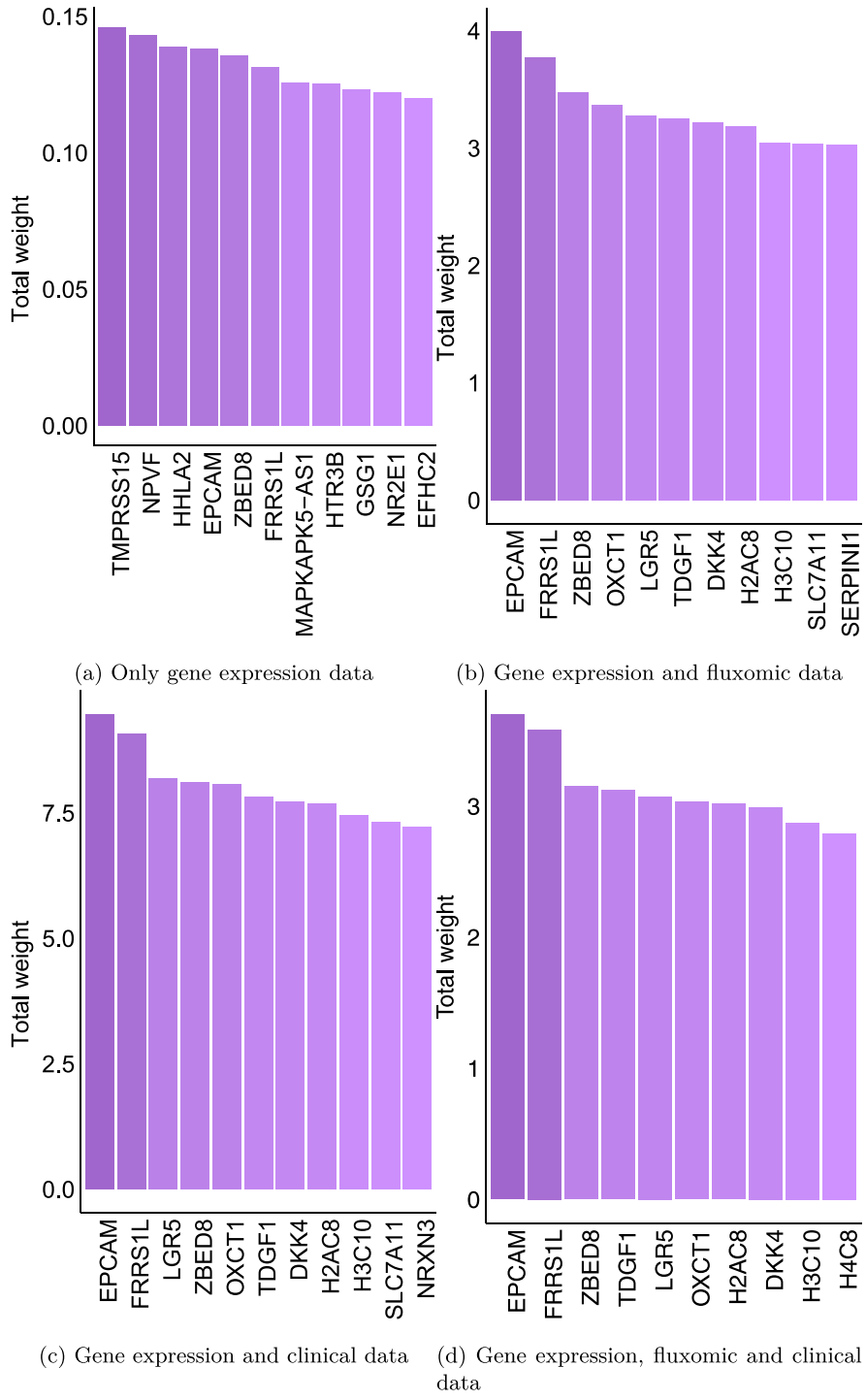
and we considered the median of the weight distribution for each feature. Both approaches led to very similar results, with the former being presented in the main text of the article. When taking the median instead of the sum (which in this scenario is qualitatively equivalent to taking the mean, since all the weights would need to be divided by 200 iterations, thus leaving their relative "importance ranking" unchanged), the following differences were found.

Among the reactions that the models considered important for the prediction, AASAD3 m was not present in the single-omic scenario, while RNMK was not relevant when the metabolic fluxes were integrated with the genes. FAOXC2251836x replaced FAOXC180x in the integrative settings, while ADNtm was found to be significant only in the metabolic fluxes without any other omic.

The results for the genes and pathways were identical to the ones reported in the main text, whereas for the weight-informed FEA the following results were found: NAD metabolism, Aminosugar metabolism, Glutathione metabolism, Citric acid cycle, Lysine metabolism, Oxidative phosphorylation were enriched in all four scenarios; Glycerophospholipid metabolism was observed only in the integrative settings, unlike Triacylglycerol synthesis which was present only in the single-omic scenario. Finally, Inositol phosphate metabolism was enriched in all four settings except when integrating metabolic fluxes and clinical data.

Appendix E. Additional tables and figures

See Figs. E.1–E.9 and Table E.1.



(a) Only gene expression data

(b) Gene expression and fluxomic data

(c) Gene expression and clinical data

(d) Gene expression, fluxomic and clinical data

Fig. E.4. Total weight bar plots for the weights attributed to the genes in the four integrative scenarios. The bar plots display only the genes with weight in the 99.5th percentile.

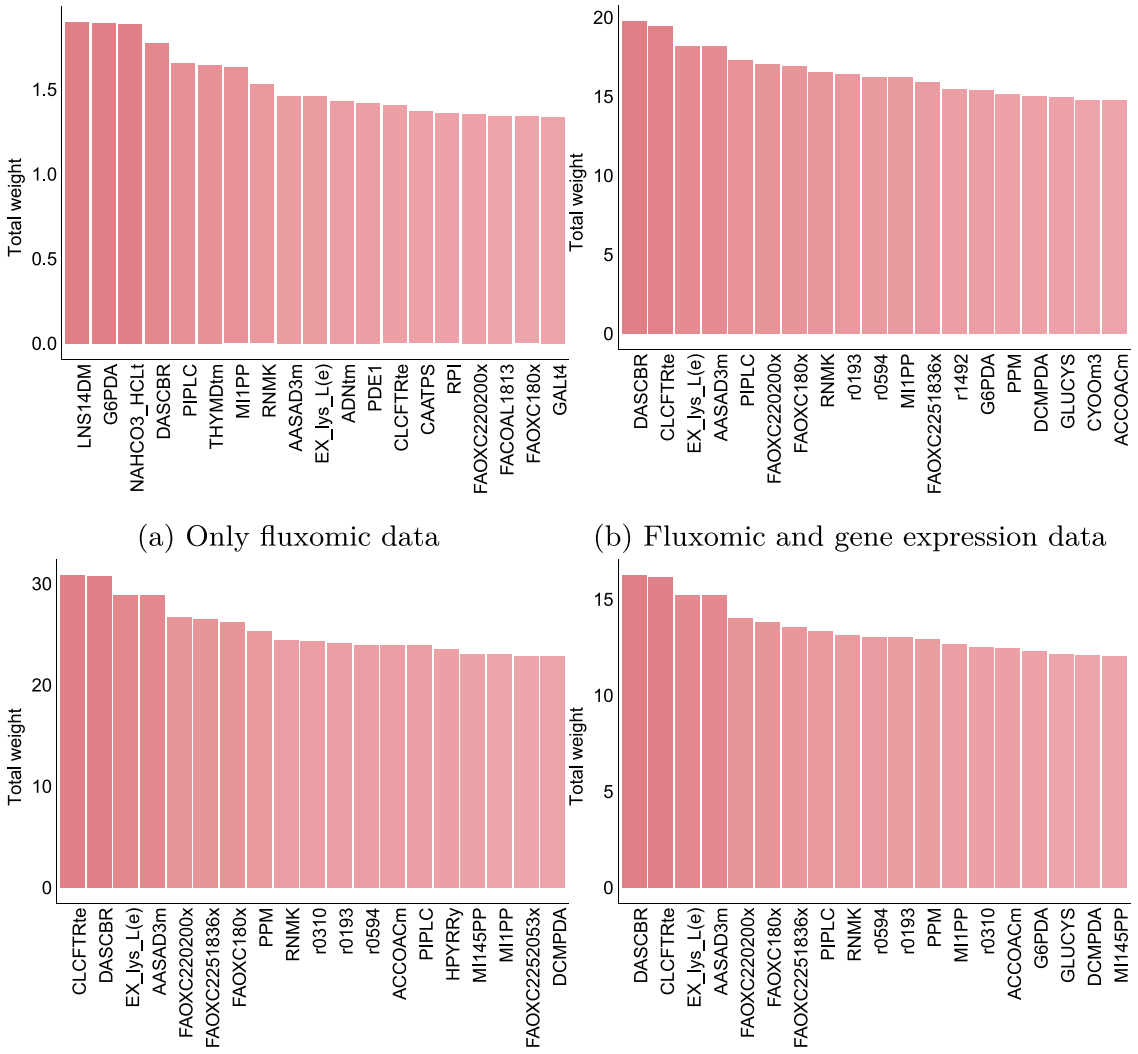
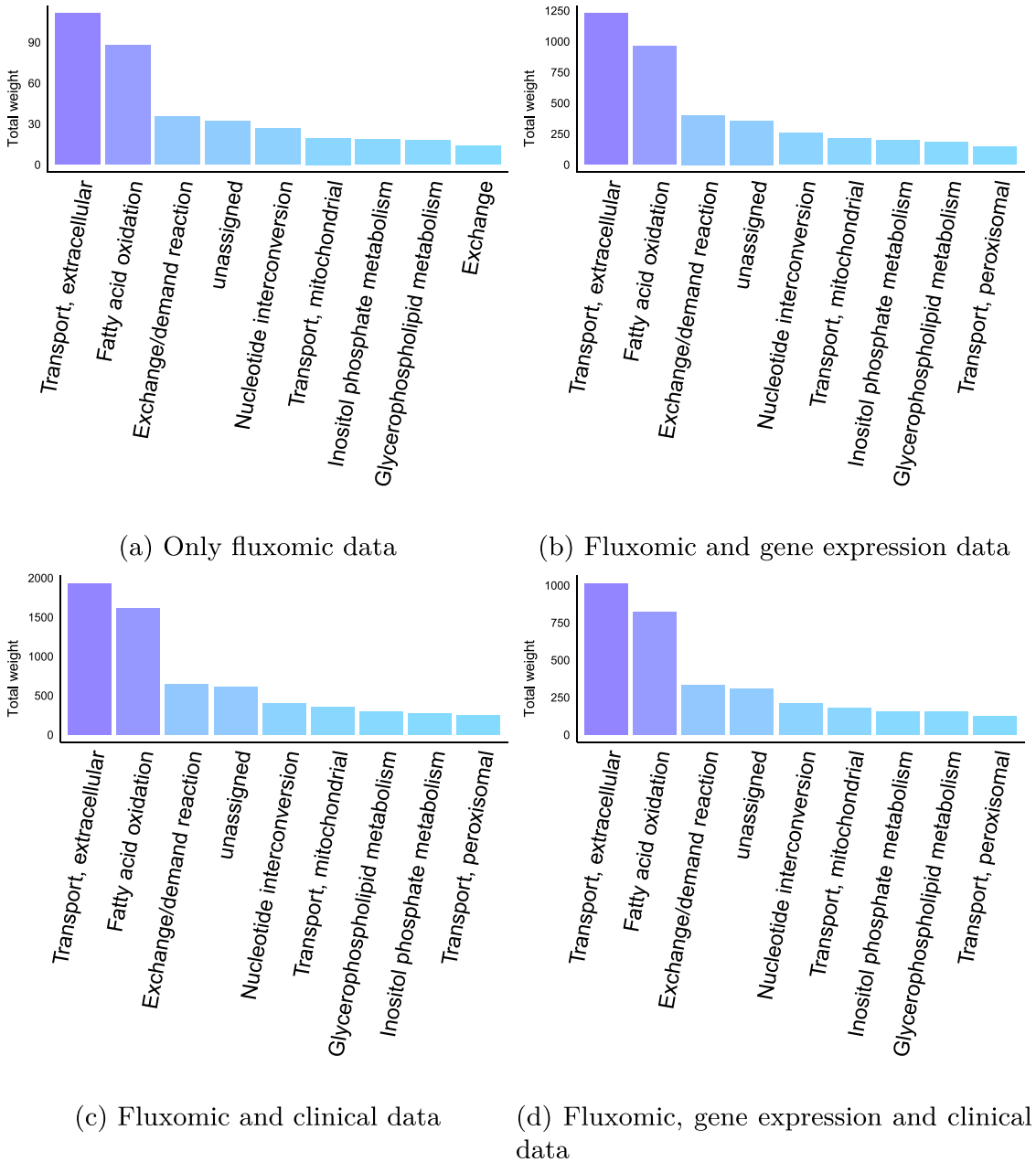


Fig. E.5. Total weight bar plots for the weights attributed to the metabolic reactions in the four integrative scenarios. The bar plots display only the reactions with weight in the 99.5th percentile.



(a) Only fluxomic data

(b) Fluxomic and gene expression data

(c) Fluxomic and clinical data

(d) Fluxomic, gene expression and clinical data

Fig. E.6. Total weight bar plots for the weights attributed to the metabolic pathways in the four integrative scenarios. The bar plots display only the pathways with weight in the 90th percentile. The most relevant pathways are consistent across the four settings, however, the integrations reveal the importance of peroxisomal transport, not shown in the single-omic scenario.

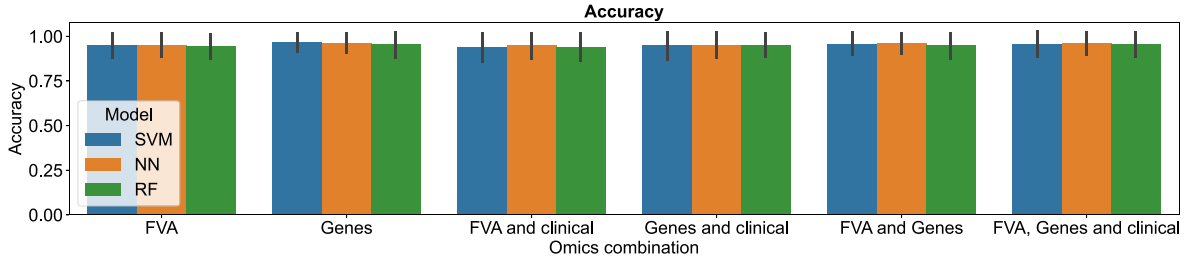


Fig. E.7. Classification accuracy for Support Vector Machine (SVM), Random Forest (RF) and Neural Network (NN) models. The three models perform comparably well (no statistically significant difference could be detected) in all omics combinations.

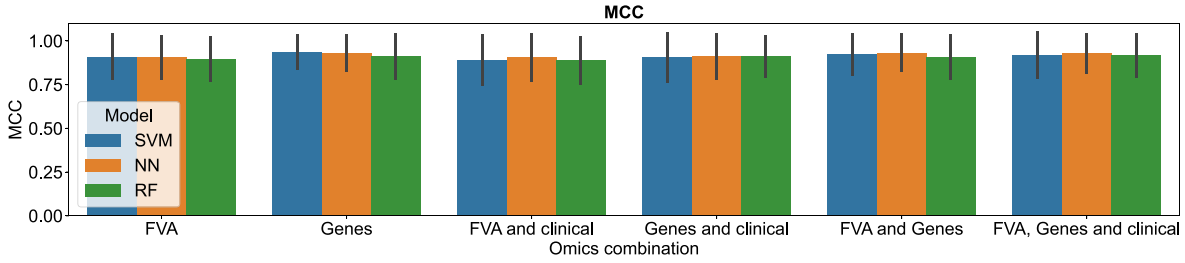


Fig. E.8. Matthew correlation coefficient for Support Vector Machine (SVM), Random Forest (RF) and Neural Network (NN) models. The three models perform comparably well (no statistically significant difference could be detected) in all omics combinations.

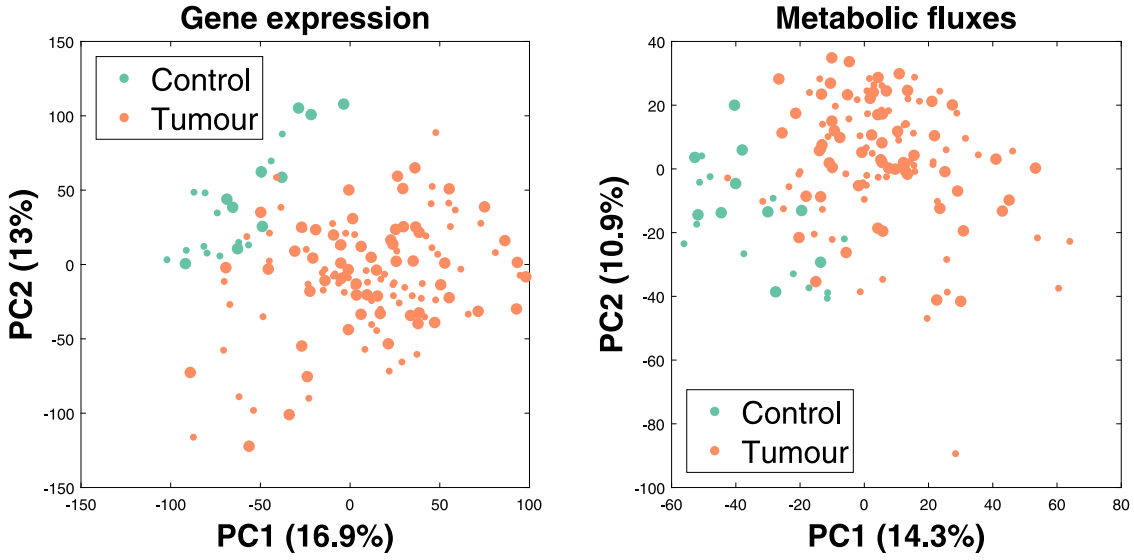


Fig. E.9. Principal component visualisation of the aggregated cohort in terms of transcriptomic and fluxomic state, displaying the main phenotypic groups. The two groups appear circumscribed to well-defined areas of the principal component space for both omics across subjects, indicating that they describe distinct characteristics in the two groups. In contrast, no clear trend can be observed in terms of subject gender, here represented by the circle size (small circle = male, big circles = female).

Table E.1
Experimental values used to constrain the model. The values were corrected by changing their sign, according to the convention by which lower bounds for exchange reactions are negative when the reaction admits uptakes.

Reaction	Reaction lower bound
D-Glucose exchange	2.025
L-histidine exchange	−0.04425
L-Isoleucine exchange	−0.0585
L-Leucine exchange	−0.0825
L-Lysine exchange	−0.2325
L-Methionine exchange	−0.12
L-phenylalanine exchange	−0.202774
L-Threonine exchange	−0.12
L-Tryptophan exchange	−0.0075
L-Valine exchange	−0.04125
H2O exchange	25.3228
O2 exchange	−28.05
CO2 exchange	21.7219
L-alanine exchange	−0.02325
L-asparagine exchange	−0.00135
L-glutamine exchange	−2.325
L-Tyrosine exchange	−0.05775
L-cysteine exchange	−0.0555
L-Arginine exchange	−0.2175
Glycine exchange	−0.2625
L-Proline exchange	0.02925
L-serine exchange	−0.1425
L-Aspartate exchange	0.00825
L-Glutamate exchange	0.15
Ammonia exchange	−0.165
Sulphate exchange	0.16121
Proton exchange	−0.42825
Glycerol exchange	−6.675
Ornithine exchange	0.125
Acetoacetate exchange	0.1275
(R)-3-Hydroxybutanoate transport via H+ symport	0.05775
L-Lactate exchange	−0.063
Urea exchange	3.375

References

[1] C. Kreutz, J. Timmer, Systems biology: Experimental design, FEBS J. 276 (4) (2009) 923–942.
[2] A.R. Kherlopian, T. Song, Q. Duan, M.A. Neimark, M.J. Po, J.K. Gohagan, A.F. Laine, A review of imaging techniques for systems biology, BMC Syst. Biol. 2 (1) (2008) 1–18.
[3] D.B. Kell, M. Brown, H.M. Davey, W.B. Dunn, I. Spasic, S.G. Oliver, Metabolic footprinting and systems biology: The medium is the message, Nat. Rev. Microbiol. 3 (7) (2005) 557–565.
[4] B. Palsson, Systems Biology, Cambridge University Press, 2015.
[5] J. Nielsen, Systems biology of metabolism: A driver for developing personalized and precision medicine, Cell Metabolism 25 (3) (2017) 572–579.
[6] I. Brigandt, Systems biology and the integration of mechanistic explanation and mathematical explanation, Stud. Hist. Phil. Sci. Part C: Stud. Hist. Phil. Biol. Biomed. Sci. 44 (4) (2013) 477–492.
[7] L. Xie, T. Evangelidis, L. Xie, P.E. Bourne, Drug discovery using chemical systems biology: Weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir, PLoS Comput. Biol. 7 (4) (2011) e1002037.
[8] M. Rahman, T. Islam, E. Gov, B. Turanli, G. Gulfidan, M. Shahjaman, N. Akhter Banu, M. Mollah, N. Haque, K.Y. Arga, et al., Identification of prognostic biomarker signatures and candidate drugs in colorectal cancer: Insights from systems biology analysis, Medicina 55 (1) (2019) 20.
[9] X. Lai, M. Eberhardt, U. Schmitz, J. Vera, Systems biology-based investigation of cooperating microRNAs as monotherapy or adjuvant therapy in cancer, Nucleic Acids Res. 47 (15) (2019) 7753–7766.
[10] T.M. Weiskittel, C. Correia, G.T. Yu, C.Y. Ung, S.H. Kaufmann, D.D. Billadeau, H. Li, The trifecta of single-cell, systems-biology, and machine-learning approaches, Genes 12 (7) (2021) 1098.
[11] N. Kubota, N. Fujiwara, Y. Hoshida, Clinical and molecular prediction of hepatocellular carcinoma risk, J. Clin. Med. 9 (12) (2020) 3843.
[12] J. Zhang, S.D. Petersen, T. Radivojevic, A. Ramirez, A. Pérez-Manríquez, E. Abeliuk, B.J. Sánchez, Z. Costello, Y. Chen, M.J. Fero, et al., Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism, Nature Commun. 11 (1) (2020) 1–13.
[13] M. Ben Guebila, I. Thiele, Predicting gastrointestinal drug effects using contextualized metabolic models, PLoS Comput. Biol. 15 (6) (2019) e1007100.
[14] G. Pio, P. Mignone, G. Magazzù, G. Zampieri, M. Ceci, C. Angione, Integrating genome-scale metabolic modelling and transfer learning for human gene regulatory network reconstruction, Bioinformatics 38 (2) (2022) 487–493.
[15] J. Feng, G. Polychronidis, U. Heger, G. Frongia, A. Mehrabi, K. Hoffmann, Incidence trends and survival prediction of hepatoblastoma in children: A population-based study, Cancer Commun. 39 (1) (2019) 1–9.
[16] S. Bharti, J.N. Bharti, A. Sinha, T. Yadav, Common and rare histological variants of hepatoblastoma in children:A pathological diagnosis and review of the literature, Gastrointest. Tumors 8 (2) (2021) 41–46.
[17] P. Sumazin, Y. Chen, L.R. Treviño, S.F. Sarabia, O.A. Hampton, K. Patel, T.A. Mistretta, B. Zorman, P. Thompson, A. Heczey, et al., Genomic analysis of hepatoblastoma identifies distinct molecular and prognostic subgroups, Hepatology 65 (1) (2017) 104–121.
[18] S. Naeem, A. Ali, S. Qadri, W. Khan Mashwani, N. Tairan, H. Shah, M. Fayaz, F. Jamal, C. Chesneau, S. Anam, Machine-learning based hybrid-feature analysis for liver cancer classification using fused (MR and CT) images, Appl. Sci. 10 (9) (2020) 3134.
[19] B.D. de Senneville, F.Z. Khoubai, M. Bevilacqua, A. Labedade, K. Flosseau, C. Chardot, S. Branchereau, J. Ripoche, S. Cairo, E. Gontier, et al., Deciphering tumour tissue organization by 3D electron microscopy and machine learning, 2021, bioRxiv.
[20] R.A. Khan, Y. Luo, F.X. Wu, Machine learning based liver disease diagnosis: A systematic review, Neurocomputing 468 (2022) 492–509.
[21] D. Hanahan, R.A. Weinberg, Hallmarks of cancer: The next generation, Cell 144 (5) (2011) 646–674.
[22] N. Kremer, A.E. Walther, G.M. Tiao, Management of hepatoblastoma: An update, Curr. Opin. Pediatr. 26 (3) (2014) 362–369.
[23] G. Zampieri, S. Vijayakumar, E. Yaneske, C. Angione, Machine and deep learning meet genome-scale metabolic modeling, PLoS Comput. Biol. 15 (7) (2019) e1007084.
[24] C. Culley, S. Vijayakumar, G. Zampieri, C. Angione, A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth, Proc. Natl. Acad. Sci. 117 (31) (2020) 18869–18879.
[25] B. Ray, M. Henaff, S. Ma, E. Efstathiadis, E.R. Peskin, M. Picone, T. Poli, C.F. Aliferis, A. Statnikov, Information content and analysis methods for multi-modal high-throughput biomedical data, Sci. Rep. 4 (1) (2014) 1–10.
[26] G. Magazzù, G. Zampieri, C. Angione, Multimodal regularised linear models with flux balance analysis for mechanistic integration of omics data, Bioinformatics (2021).
[27] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[28] S. Cairo, C. Armengol, A. De Reyniès, Y. Wei, E. Thomas, C.A. Renard, A. Goga, A. Balakrishnan, M. Semeraro, L. Gresh, et al., Hepatic stem-like phenotype and interplay of Wnt/β-catenin and Myc signaling in aggressive childhood liver cancer, Cancer Cell 14 (6) (2008) 471–484.
[29] J. Fu, Y. Zhang, Y. Wang, H. Zhang, J. Liu, J. Tang, Q. Yang, H. Sun, W. Qiu, Y. Ma, et al., Optimization of metabolomic data processing using NOREVA, Nat. Protoc. 17 (1) (2022) 129–151.
[30] J. Tang, J. Fu, Y. Wang, B. Li, Y. Li, Q. Yang, X. Cui, J. Hong, X. Li, Y. Chen, et al., ANPELA: Analysis and performance assessment of the label-free quantification workflow for metaproteomic studies, Brief. Bioinform. 21 (2) (2020) 621–636.
[31] W.W.B. Goh, C.H. Yong, L. Wong, Are batch effects still relevant in the age of big data? Trends Biotechnol. (2022).
[32] W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics 8 (1) (2007) 118–127.
[33] A. Mardinoglu, R. Agren, C. Kampf, A. Asplund, M. Uhlen, J. Nielsen, Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease, Nature Commun. 5 (1) (2014) 1–11.
[34] A. Occhipinti, Y. Hamadi, H. Kugler, C.M. Wintersteiger, B. Yordanov, C. Angione, Discovering essential multiple gene effects through large scale optimization: An application to human cancer metabolism, IEEE/ACM Trans. Comput. Biol. Bioinform. 18 (6) (2020) 2339–2352.
[35] K. Yizhak, E. Gaude, S. Le Dévédec, Y.Y. Waldman, G.Y. Stein, B. van de Water, C. Frezza, E. Ruppin, Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer, Elife 3 (2014) e03641.
[36] M. Uhlen, C. Zhang, S. Lee, E. Sjöstedt, L. Fagerberg, G. Bidkhori, R. Benfeitas, M. Arif, Z. Liu, F. Edfors, et al., A pathology atlas of the human cancer transcriptome, Science 357 (6352) (2017).
[37] A. Nilsson, J. Nielsen, Genome scale metabolic modeling of cancer, Metab. Eng. 43 (2017) 103–112.
[38] S. Vijayakumar, M. Conway, P. Lió, C. Angione, Optimization of multi-omic genome-scale models: Methodologies, hands-on tutorial, and perspectives, Metab. Network Reconstruct. Model. (2018) 389–408.
[39] J.S. Cho, C. Gu, T.H. Han, J.Y. Ryu, S.Y. Lee, Reconstruction of context-specific genome-scale metabolic models using multiomics data to study metabolic rewiring, Curr. Opin. Syst. Biol. 15 (2019) 1–11.
[40] O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin, T. Shlomi, Predicting selective drug targets in cancer through metabolic networks, Mol. Syst. Biol. 7 (1) (2011) 501.
[41] R. Agren, A. Mardinoglu, A. Asplund, C. Kampf, M. Uhlen, J. Nielsen, Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling, Mol. Syst. Biol. 10 (3) (2014) 721.
[42] N. Swainston, K. Smallbone, H. Hefzi, P.D. Dobson, J. Brewer, M. Hanscho, D.C. Zielinski, K.S. Ang, N.J. Gardiner, J.M. Gutierrez, et al., Recon 2.2: From reconstruction to model of human metabolism, Metabolomics 12 (7) (2016) 1–7.
[43] C. Angione, P. Lió, Predictive analytics of environmental adaptability in multi-omic network models, Sci. Rep. 5 (1) (2015) 1–21.
[44] C. Angione, Integrating splice-isoform expression into genome-scale models characterizes breast cancer metabolism, Bioinformatics 34 (3) (2018) 494–501.
[45] C. Chan, F. Berthiaume, K. Lee, M.L. Yarmush, Metabolic flux analysis of cultured hepatocytes exposed to plasma, Biotechnol. Bioeng. 81 (1) (2003) 33–49.
[46] R. Mahadevan, C.H. Schilling, The effects of alternate optimal solutions in constraint-based genome-scale metabolic models, Metab. Eng. 5 (4) (2003) 264–276.
[47] N.E. Lewis, K.K. Hixson, T.M. Conrad, J.A. Lerman, P. Charusanti, A.D. Polpitiya, J.N. Adkins, G. Schramm, S.O. Purvine, D. Lopez-Ferrer, et al., Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models, Mol. Syst. Biol. 6 (1) (2010) 390.
[48] S.M. Schinn, C. Morrison, W. Wei, L. Zhang, N.E. Lewis, Systematic evaluation of parameters for genome-scale metabolic models of cultured mammalian cells, Metab. Eng. 66 (2021) 21–30.
[49] Z. Dai, S. Yang, L. Xu, H. Hu, K. Liao, J. Wang, Q. Wang, S. Gao, B. Li, L. Lai, Identification of cancer–associated metabolic vulnerabilities by modeling multi-objective optimality in metabolism, Cell Commun. Signal. 17 (1) (2019) 1–15.
[50] L. Heirendt, S. Arreckx, T. Pfau, S.N. Mendoza, A. Richelle, A. Heinken, H.S. Haraldsdóttir, J. Wachowiak, S.M. Keating, V. Vlasov, et al., Creation and analysis of biochemical constraint-based models using the COBRA toolbox v. 3.0, Nat. Protoc. 14 (3) (2019) 639–702.
[51] F. Fabris, D. Palmer, J.P. de Magalhães, A.A. Freitas, Comparing enrichment analysis and machine learning for identifying gene properties that discriminate between gene classes, Brief. Bioinform. 21 (3) (2020) 803–814.
[52] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[53] P.H. Dinh, Combining gabor energy with equilibrium optimizer algorithm for multi-modality medical image fusion, Biomed. Signal Process. Control 68 (2021) 102696.
[54] P.H. Dinh, A novel approach based on grasshopper optimization algorithm for medical image fusion, Expert Syst. Appl. 171 (2021) 114576.
[55] C. Angione, J. Costanza, G. Carapezza, P. Lió, G. Nicosia, Multi-target analysis and design of mitochondrial metabolism, PLoS One 10 (9) (2015) e0133825.

[56] J. Zhou, Y. Qiu, S. Zhu, D.J. Armaghani, C. Li, H. Nguyen, S. Yagiz, Optimization of support vector machine through the use of metaheuristic algorithms in forecasting TBM advance rate, Eng. Appl. Artif. Intell. 97 (2021) 104015.
[57] M. Ojala, G.C. Garriga, Permutation tests for studying classifier performance, J. Mach. Learn. Res. 11 (6) (2010).
[58] G. Valente, A.L. Castellanos, L. Hausfeld, F. De Martino, E. Formisano, Cross-validation and permutations in MVPA: Validity of permutation strategies and power of cross-validation schemes, NeuroImage (2021) 118145.
[59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[60] T.U. Consortium, UniProt: The universal protein knowledgebase in 2021, Nucleic Acids Res. 49 (D1) (2020) D480–D489, http://dx.doi.org/10.1093/nar/gkaa1100, arXiv:https://academic.oup.com/nar/article-pdf/49/D1/D480/35364103/gkaa1100.pdf.
[61] M.E. Brosnan, J.T. Brosnan, Hepatic glutamate metabolism: A tale of 2 hepatocytes, Am. J. Clin. Nutr. 90 (3) (2009) 857S–861S.
[62] R.J. Dunlop, C.W. Campbell, Cytokines and advanced cancer, J. Pain Symptom Manag. 20 (3) (2000) 214–232.
[63] F. Vidal-Vanaclocha, C. Amézaga, A. Asumendi, G. Kaplanski, C.A. Dinarello, Interleukin-1 receptor blockade reduces the number and size of murine B16 melanoma hepatic metastases, Cancer Res. 54 (10) (1994) 2667–2672.
[64] M. Shimonosono, T. Arigami, S. Yanagita, D. Matsushita, Y. Uchikado, Y. Kijima, H. Kurahara, Y. Kita, S. Mori, K. Sasaki, et al., The association of human endogenous retrovirus-H long terminal repeat-associating protein 2 (HHLA2) expression with gastric cancer prognosis, Oncotarget 9 (31) (2018) 22069.
[65] M. Karlsson, C. Zhang, L. Méar, W. Zhong, A. Digre, B. Katona, E. Sjöstedt, L. Butler, J. Odeberg, P. Dusart, et al., A single–cell type transcriptomics map of human tissues, Sci. Adv. 7 (31) (2021) eabh2169.
[66] Y. Dai, Y. Wang, Y. Cao, P. Yu, L. Zhang, Z. Liu, Y. Ping, D. Wang, G. Zhang, Y. Sang, et al., A multivariate diagnostic model based on urinary EpCAM-CD9-positive extracellular vesicles for prostate cancer diagnosis, Front. Oncol. 11 (2021).
[67] G. Lin, H. Ye, J. Wang, S. Chen, X. Chen, C. Zhang, Immune checkpoint human endogenous retrovirus-H long terminal repeat-associating protein 2 is upregulated and independently predicts unfavorable prognosis in bladder urothelial carcinoma, Nephron 141 (4) (2019) 256–264.
[68] A.A. Pramono, G.M. Rather, H. Herman, K. Lestari, J.R. Bertino, NAD-and NADPH-contributing enzymes as therapeutic targets in cancer: An overview, Biomolecules 10 (3) (2020) 358.
[69] Q. Zhao, Z. Zhang, J. Li, F. Xu, B. Zhang, M. Liu, Y. Liu, H. Chen, J. Yang, J. Zhang, Lysine acetylome study of human hepatocellular carcinoma tissues for biomarkers and therapeutic targets discovery, Front. Genet. 11 (2020).
[70] G. Vazquez Rodriguez, A. Abrahamsson, M.V. Turkina, C. Dabrosin, Lysine in combination with estradiol promote dissemination of estrogen receptor positive breast cancer via upregulation of U2AF1 and RPN2 proteins, Front. Oncol. 10 (2020) 2650.
[71] R. Zhang, L. Noordam, X. Ou, B. Ma, Y. Li, P. Das, S. Shi, J. Liu, L. Wang, P. Li, et al., The biological process of lysine-tRNA charging is therapeutically targetable in liver cancer, Liver Int. 41 (1) (2021) 206–219.
[72] T.Y. Hargrove, L. Friggeri, Z. Wawrzak, S. Sivakumaran, E.M. Yazlovitskaya, S.W. Hiebert, F.P. Guengerich, M.R. Waterman, G.I. Lepesheva, Human sterol 14α-demethylase as a target for anticancer chemotherapy: Towards structure-aided drug design1, J. Lipid Res. 57 (8) (2016) 1552–1563.
[73] A. Gorbatenko, C.W. Olesen, E. Boedtkjer, S.F. Pedersen, Regulation and roles of bicarbonate transport in cancer, Front. Physiol. 5 (2014) 130.
[74] I.F. Robey, N.K. Martin, Bicarbonate and dichloroacetate: Evaluating pH altering therapies in a mouse model for metastatic breast cancer, BMC Cancer 11 (1) (2011) 1–10.
[75] O.C. Yang, S.H. Loh, Acidic stress triggers sodium-coupled bicarbonate transport and promotes survival in A375 human melanoma cells, Sci. Rep. 9 (1) (2019) 1–12.
[76] D.R. Chhetri, Myo-inositol and its derivatives: Their emerging role in the treatment of human diseases, Front. Pharmacol. 10 (2019) 1172.
[77] L. Dearden, S.G. Bouret, S.E. Ozanne, Sex and gender differences in developmental programming of metabolism, Mol. Metab. 15 (2018) 8–19.
[78] L.A. Williams, J. Sample, C.C. McLaughlin, B.A. Mueller, E.J. Chow, S.E. Carozza, P. Reynolds, L.G. Spector, Sex differences in associations between birth characteristics and childhood cancers: A five-state registry-linkage study, Cancer Causes Control 32 (11) (2021) 1289–1298.
[79] S. Niedenführ, W. Wiechert, K. Nöh, How to measure metabolic fluxes: A taxonomic guide for 13C fluxomics, Curr. Opin. Biotechnol. 34 (2015) 82–90.
[80] J.D. Orth, I. Thiele, B.Ø. Palsson, What is flux balance analysis? Nature Biotechnol. 28 (3) (2010) 245–248.