OXFORD

BIOLOGY
Methods & Protocols

Biology Methods and Protocols, 2024, bpae048
https://doi.org/10.1093/biomethods/bpae048
Advance Access Publication Date: 2 July 2024
Methods Article

Machine learning of cellular metabolic rewiring

Joao B. Xavier [1],*

[1]Program for Computational and Systems Biology, Memorial Sloan Kettering Cancer Center, New York, NY, USA

*Correspondence address. Program for Computational and Systems Biology, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, Box 460, New York, NY 10065, USA. Tel: +1(646) 888-3195; Fax: +1(646) 422-0717. E-mail: xavierj@mskcc.org

Abstract

Metabolic rewiring allows cells to adapt their metabolism in response to evolving environmental conditions. Traditional metabolomics techniques, whether targeted or untargeted, often struggle to interpret these adaptive shifts. Here, we introduce *MetaboLiteLearner*, a lightweight machine learning framework that harnesses the detailed fragmentation patterns from electron ionization (EI) collected in scan mode during gas chromatography/mass spectrometry to predict changes in the metabolite composition of metabolically adapted cells. When tested on breast cancer cells with different preferences to metastasize to specific organs, *MetaboLiteLearner* predicted the impact of metabolic rewiring on metabolites withheld from the training dataset using only the EI spectra, without metabolite identification or pre-existing knowledge of metabolic networks. Despite its simplicity, the model learned captured shared and unique metabolomic shifts between brain- and lung-homing metastatic lineages, suggesting cellular adaptations associated with metastasis to specific organs. Integrating machine learning and metabolomics paves the way for new insights into complex cellular adaptations.

Introduction

Cells dynamically rewire their intracellular metabolism in response to changing nutrients, signals, and environmental cues [1]. These adjustments allow cells to maintain homeostasis, optimize energy production, and fulfill the biosynthetic demands for growth and repair. From unicellular organisms like yeast adapting to changing nutrient availability [2] to the cells of multicellular organisms during development [3], disease [4], or environmental stress [5], metabolic rewiring is a universal feature of cellular adaptation and survival.

Metabolic rewiring shifts the biochemical composition of cells, and the changes observed in the intracellular metabolome provide a window into the underlying cellular adaptation [6]. Traditional approaches to metabolomics are either targeted or untargeted. Targeted metabolomics focuses on a subset of metabolites of interest, validated among the set of detectable compounds; it aims to detect and quantify accurately that predefined set of metabolites but risks missing novel or unexpected ones [7, 8]. Untargeted metabolomics allows for the comprehensive survey of all detectable metabolites within a biological sample but has challenges regarding reproducibility, identifying unknown compounds and finding a sensitive technique for all relevant compounds [9].

Despite the potential to apply machine learning methods to metabolic rewiring, the field has also faced significant limitations. Unsupervised learning methods like clustering and principal component analysis are useful for identifying patterns among biological samples without predefined labels. Yet, they struggle with the high dimensionality and complexity of metabolic data [10]. Supervised learning approaches aim to predict the sample type from labeled data, but their effectiveness is often hampered by the scarcity of sample sizes in metabolic datasets [11]. Reinforcement learning, which optimizes metabolic pathways by modeling dynamic interactions, faces challenges due to their high computational demands and difficulty defining appropriate reward structures in biological contexts [12]. Additionally, most models in this domain lack interpretability [13], making it difficult to derive actionable biological insights from their predictions.

To address these limitations, we present *MetaboLiteLearner*. This novel computational method uses a simple machine learning algorithm to investigate metabolic rewiring using metabolomic data without relying on prior knowledge of the metabolic network or having to identify the compounds from their spectra. *MetaboLiteLearner* deploys the extensive molecular fragmentation achieved through electron ionization (EI) in gas chromatography/mass spectrometry (GC/MS) and acquired in scan mode. The fragmentation patterns are treated as input features in a lightweight supervised learning model, which can associate these features with the changes in metabolite abundance observed in cells undergoing metabolic rewiring. Despite the simplicity of the machine learning algorithm at its core, *MetaboLiteLearner* can effectively use information about the molecular structure of metabolites to predict how their abundances change in cells adapted metabolically to diverse circumstances.

As proof of concept, we utilized clones of the MDA-MB-231 breast cancer cell line that home to the lung and brain [14]. These derivatives, developed through *in vivo* selection in mice [15, 16], show pronounced transcriptomic shifts in laboratory cultures. In a previous study, we employed targeted metabolomics to analyze a comprehensive panel of 645 metabolites from these cells. That analysis unveiled distinct intracellular metabolic profiles between the brain- and lung-homing cells and their parental counterparts—distinct metabolic rewiring patterns that were

Received: 22 April 2024. **Revised:** 14 June 2024. **Editorial decision:** 24 June 2024. **Accepted:** 01 July 2024

posteriorly validated through direct metabolic exchange measurements and indirect assessments via stable isotope tracing [17]. Given this detailed characterization, the MDA-MB-231 system serves as a well-studied and oncological-relevant model to benchmark the efficacy of *MetaboLiteLearner*.

MetaboLiteLearner identified unique metabolic-fragment features in brain- and lung-homing cells that may be adaptations to the specific challenges of their target organs. This insight aligns with the findings from our earlier study [17], which relied on a large panel of metabolites validated and measured through liquid chromatography/mass spectrometry, a method that is more laborious and expensive. On the other hand, *MetaboLiteLearner* draws insights using all the data produced from GC/MS in scan mode, without relying on prior information such as spectral libraries or extensive validation. *MetaboLiteLearner* can be applied to various cell types and conditions, allowing researchers in different fields to explore new aspects of cellular metabolism. *MetaboLiteLearner* paves the way for studying metabolic rewiring by finding associations between molecular structure and metabolic adaptability.

Methods

Cell culture, metabolite extraction and derivatization

Cell lines were cultured in Dulbecco's modified Eagle medium (DMEM, Fisher 11965118) supplemented with 10% fetal bovine serum, produced in the Memorial Sloan Kettering Cancer Center media core facility, and 1% penicillin/streptomycin (Fisher 15140122). The cultivation conditions included a 37°C incubator with regulated humidity and a 5% CO2 atmosphere. Authenticated cell lines were procured from the Massague lab and were developed as described in previous studies [15, 16]. Cells were subjected to metabolite extraction using 1 ml of ice-cold 80% methanol, followed by overnight storage at −80°C. Subsequent to this, the extracts underwent a drying process using an evaporator. Resuspension was achieved by incubation with shaking at 30°C for 2 hours in a solution containing 50 µl of 40 mg ml^{-1} methoxyamine hydrochloride in pyridine. Derivatization was performed by adding 80 µl of N-methyl-N-(trimethylsilyl) trifluoroacetamide (with or without 1% trimethylchlorosilane from Thermo Fisher Scientific) and 70 µl of ethyl acetate (sourced from Sigma-Aldrich). This mixture was then incubated at 37°C for 30 min. The derivatization introduces specific changes in the EI spectra, such as the presence of Si-related isotopic patterns. These patterns are typically considered beneficial for identifying functional groups and improving volatility, thus enhancing the detection and quantification of metabolites in GC/MS analysis.

GC/MS analysis

Analytical procedures utilized the Agilent 7890A gas chromatograph paired with an Agilent 5977C mass selective detector. The gas chromatograph operated in splitless injection mode, maintaining a constant helium gas flow at 1 ml/min. The injection involved introducing 1 µl of the derivatized metabolites onto an HP-5ms column. The temperature of the gas chromatograph oven was ramped from 60°C to 290°C over a 25-min interval, following the Fiehn method protocol provided by Agilent [18]. Samples comprised four distinct types: blank media, parental cells, lung-homing cells, and brain-homing cells. Each type was cultured in triplicate groups over a span of three days, resulting in nine replicates for each sample category.

Data processing

GC/MS raw data, stored in the Agilent .D format, underwent processing to generate the *MetaboLiteLearner* Open Dataset (MLOD). The raw .D files were initially converted into Comma-Separated Values (CSV). By creating a “virtual bulk sample” from this CSV data, we could detect and extract spectra from the total ion chromatogram.

Peak detection was performed directly on the virtual bulk sample, the Matlab function *mspeaks*, without any deconvolution step. The spectra of each detected peak were then extracted, and their peak areas were calculated for each sample. The absence of a deconvolution step simplifies the data processing pipeline and avoids potential issues associated with deconvolution quality, which can be critical for annotation via library matching.

An integration process yielded a peak area table, ensuring the relevance of the dataset by removing compounds that did not exhibit significant differences compared to blank media samples, as determined using analysis of variance. The refined data formed the MLOD dataset, comprising 153 unique spectra labeled with abundance alterations, represented as $\log_2$ fold changes for both brain-homing and lung-homing cells.

Using raw spectra and peak detection directly on a composite sample is advantageous as it avoids potential biases and errors introduced during deconvolution. Deconvolution relies heavily on the quality of the algorithm and may affect the accuracy of peak identification and quantification. By using raw spectra and focusing on peak detection and area calculation, our method ensures a robust and straightforward analysis pipeline, enhancing the reproducibility and reliability of the results.

Machine learning approach

Supervised learning attempts to learn from training data containing inputs matched to their correct outputs, a model $Y = f(X)$ that can generalize to a new dataset [19]. During the learning phase, the model is a “student” given a set X_{train}, Y_{train} of training. For each example x_i in the training dataset X_{train} the model presents the answer $f(x_i) = y\hat{\ }_i$. The supervisor or “teacher” compares $\hat{y}_i$ with the correct answer, y_i, and gives an error associated with the “student's” answer. The examples' errors are used to calculate a loss, $L(y_i, \hat{y}_i)$, and the model learns by adjusting its parameters to minimize the loss. After the training, the model is presented with a test dataset (X_{test}) containing a new set of examples. The model's accuracy in predicting the output for the new inputs shown is evaluated by comparing these predictions $\hat{Y}_{test}$ with the actual outputs Y_{test}.

MetaboLiteLearner model

MetaboLiteLearner models a linear function f that takes a feature vector representation of a metabolite as input. It predicts, as output, how the intracellular level of that metabolite is impacted by metabolic rewiring. Each metabolite is represented by an input vector x_i, a p-dimensional array of the features of metabolite i. In our case, x_i is 550-dimensional, representing the abundances of ionic fragments of metabolite i of sizes between 50 and 600 mass-to-charge ratio units (m/z) binned at unit intervals. This vector is obtained from the EI mass spectrum. The EI-spectrum is a mass-to-charge histogram of the ion fragments produced when a metabolite undergoes breakdown via EI. It offers a direct insight into the metabolite's structure—and potentially its function—without needing to identify that metabolite first.

The output vector y_i is two-dimensional, representing the log2 fold change of that metabolite in brain-homing and lung-homing cells compared to the parental lineage.

Partial least squares regression

MetaboLiteLearner's learning algorithm uses the partial least squares regression (PLSR) [20]. This algorithm has similarities with artificial neural networks [21] but uses only linear functions, which makes it more stable and less computationally demanding to run. The following equations illustrate PLSR:

Input Matrix X: The matrix of predictor variables (spectra).

Output Matrix Y: The matrix of response variables (log2 fold changes).

Latent Variables T and U:

$$T = XW$$

$$U = YC$$

where W and C are weights matrices.

Regression on Latent Variables:

$$T = XB + E$$

$$U = YC + F$$

where B is the regression coefficient matrix, and E and F are error matrices.

Predicting Y from X:

$$Y = XW(C^TC)^{-1}C^T$$

To transform the high-dimensional input (a 550-dimensional EI spectrum) and output (log2 fold changes) into a latent space, *MetaboLiteLearner* uses two matrices: W for the spectra and C for the log2 fold changes. The latent space dimensionality, N, dictates the model's complexity. An optimal number of dimensions, N_{opt}, is determined through cross-validation to prevent under- and overfitting. When applied to metabolomics data, the loadings derived from PLSR—coefficients that describe the relationship between the original predictors and the new latent factors—provide insights into the underlying cellular adaptations. These loadings are directly tied to the EI-fragmentation spectra of metabolites and indicate molecular structural features linked to their abundance changes in metabolically rewired cells. Therefore, once the model is trained, the loadings provide insights into the relationships between metabolite-fragment composition and metabolic rewiring, shedding light on cellular adaptation mechanisms. We used the MATLAB implementation of PLSR encoded in function *plsregress*.

Hyperparameter tuning and challenges

Hyperparameter tuning for *MetaboLiteLearner* primarily involved selecting N_{opt}, the optimal number of latent components which sets the dimensionality of the latent space. Using leave-one-out cross-validation, we systematically varied the number of latent components and selected the number that minimized prediction error. To ensure the robustness of the analytical outcomes, a shuffling test was administered, wherein the order of observed data was randomly rearranged a thousand times. A key challenge during training was avoiding overfitting, particularly given the high-dimensional feature space relative to the sample size. The regularization—implicit in the choice of a low-dimensional latent space—selected through cross-validation helped mitigate this risk.

Evaluation metrics

Hold-out cross-validation was used to determine the optimal number of latent components, with the mean squared error calculated for both the training and test sets to quantify prediction accuracy. We also computed the correlation coefficient (ρ) between the predicted and actual log2 fold changes to assess predictive accuracy. The percentage of variance explained by each latent component was evaluated for both predictor and response variables. Additionally, a randomization test was performed to ensure the model captured biologically relevant patterns, comparing the performance of the model with 1,000 shuffled data to the performance with the original data.

Data preprocessing

Each spectrum was normalized to account for differences in sample ionization efficiency. Fragmentation spectra encoding involved converting the mass spectrometry data into a consistent format: binned m/z ratios into unit intervals, resulting in a 550-dimensional feature vector for each metabolite. This ensured uniform input dimensions across samples and preserved the relative abundances of ion fragments.

Biological context and data sources

To contextualize our results biologically, we referenced a table of compounds with known biological roles from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [22]. Additionally, we utilized spectra for trimethylsilyl (TMS)-derivatized compounds from the Fiehn Library [23]. These spectra were processed through *MetaboLiteLearner* to compute fold changes in brain- and lung-homing cells and derive values for the five latent components.

Data and code availability

The raw data used in this study are available in the Zenodo repository [24]. The code for *MetaboLiteLearner*, including scripts for data preprocessing, model training, and validation, is available on GitHub at https://github.com/joaobxavier/learn_metabolic_rewiring_matlab. This GitHub repository is permanently archived on Zenodo [25], ensuring long-term access and reproducibility of the code. The repository includes a detailed README file with a step-by-step tutorial to guide users through the entire workflow, from downloading raw data to performing analysis and validating results.

Results

MetaboLiteLearner: theory and application

MetaboLiteLearner capitalizes on GC/MS data from intracellular metabolome extracts. The extracts are primed for GC/MS analysis using methoximation (MeOX) followed by trimethylsilyl to derivatize the metabolites containing functional groups with active hydrogens (such as hydroxyl, carboxyl, amino, and thiol groups) and carbonyl groups (like aldehydes and ketones). The derivatization enhances metabolite volatility. The derivatization, combined with GC/MS and EI fragmentation, produces reproducible spectra and retention times for each metabolite that was successfully derivatized with TMS [26]. Conventional targeted metabolomics often measures specific ion peak areas (Fig. 1A). Here, *MetaboLiteLearner* harnesses the entirety of the data available in scan mode (Fig. 1B). This spans ion fragment abundances from 50 to 600 mass-to-charge ratio units (m/z) and all GC retention times.

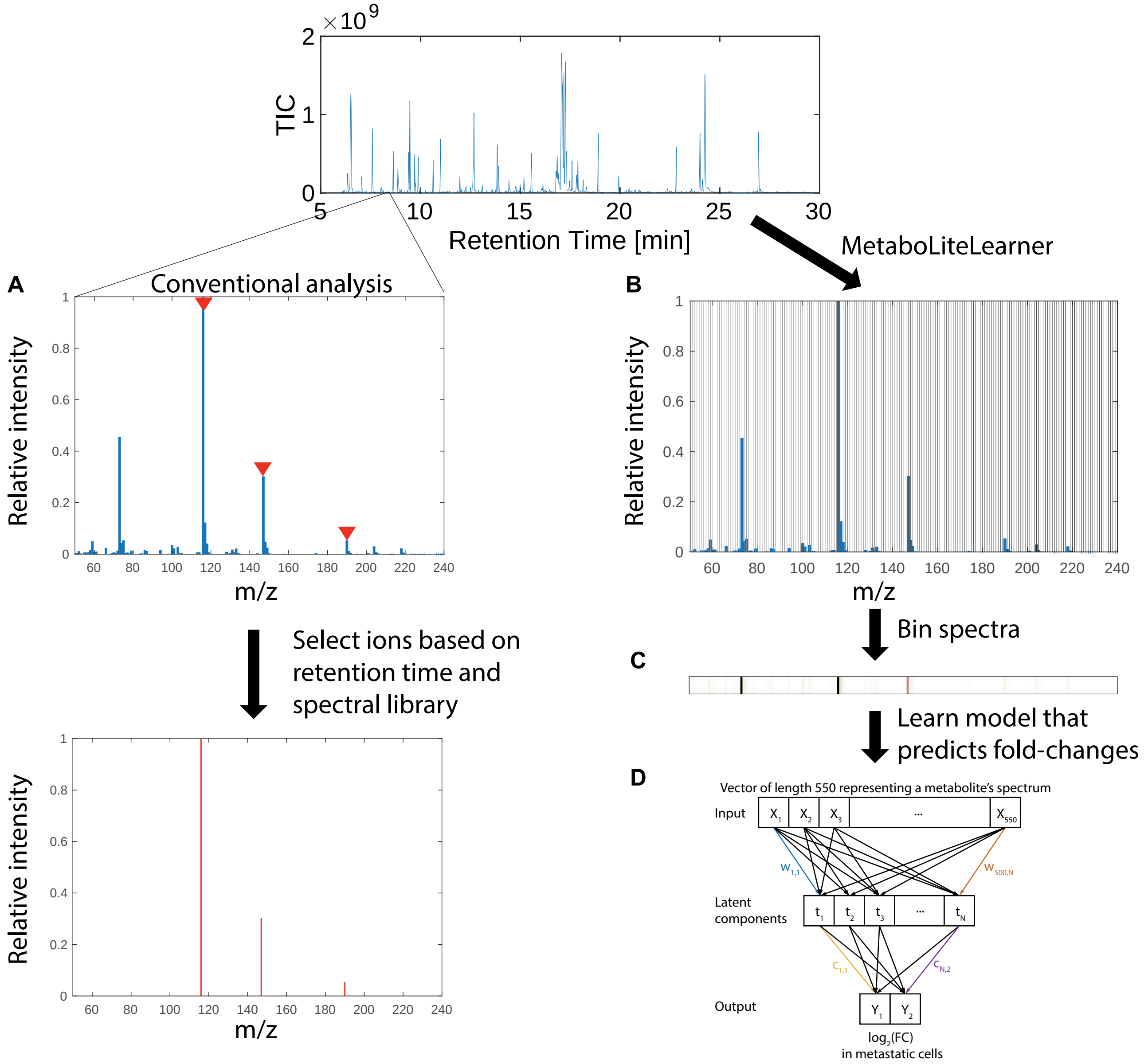

Figure 1. Workflow and Functionality of *MetaboLiteLearner*. **(A)** Traditional targeted metabolomics uses specific ion peak areas. **(B)** *MetaboLiteLearner* uses GC/MS data acquired in scan mode in *MetaboLiteLearner*, which encompasses ion fragment abundances ranging from 50 to 600 m/z captured at all GC retention times. **(C)** Each molecule's electron impact (EI) fragmentation spectrum is depicted as a 550-dimensional vector. These high-dimensional vectors are paired with their corresponding log2 fold change values, which serve as training labels. **(D)** The transformation matrices ('w' for spectra and 'c' for log2 fold changes) are the PLSR loadings used to map data into the *N*-dimensional latent space. This enables *MetaboLiteLearner* to learn the relationship between metabolite-fragment composition and metabolic rewiring.

MetaboLiteLearner integrates full spectra with the corresponding $\log_2$ fold changes in abundance, reflecting their abundance changes due to metabolic rewiring. Each molecule's EI fragmentation spectrum is encoded as a 550-dimensional vector (Fig. 1C), with corresponding $\log_2$ fold changes (obtained by comparing baseline vs. rewired cell metabolite abundances) as the training data labels. Following the supervised learning paradigm, the model's efficacy is gauged by predicting $\log_2$ fold changes on a new compound set and comparing these predictions with actual data [19].

MetaboLiteLearner uses PLSR to construct a linear model by projecting the predictors and the response variables onto a new *N*-dimensional space [20]. PLSR generates a model that balances complexity and interpretability without sacrificing the power of linear combinations of original variables. This approach is analogous to an Artificial Neural Network that employs a hidden layer with *N* neurons, except that PLSR employs linear functions instead of networks of non-linear activation functions [21].

To transform the high-dimensional input (a 550-dimensional EI spectrum) and output ($\log_2$ fold changes) into a latent space, *MetaboLiteLearner* uses two matrices: W for the spectra and C for the $\log_2$ fold changes (Fig. 1D). The latent space dimensionality, *N*, dictates the model's complexity. An optimal number of dimensions, N_{opt}, is determined through cross-validation to prevent overfitting [19]. When applied to metabolomics data, the loadings derived from PLSR—coefficients that describe the relationship between the original predictors and the new latent factors—provide insights into the underlying cellular adaptations. These loadings are directly tied to the EI-fragmentation spectra of metabolites and indicate molecular structural features linked to their abundance changes in metabolically rewired cells. Therefore, once the model is trained, the loadings provide insights into the relationships between metabolite-fragment composition and metabolic rewiring, shedding light on cellular adaptation mechanisms.

Breast cancer cell data integration into *MetaboLiteLearner*

We collected the data for *MetaboLiteLearner* from the MDA-MB-231 breast cancer cell line and its brain and lung-targeted derivatives. The derivatives, originating from *in vivo* mouse selection [15, 16], exhibited transcriptomic differences compared to the original MDA-MB-231 cells in lab cultures. Our prior work identified marked intracellular metabolome variances between the brain- and lung-homing cells and the parent cells, indicating that cells have undergone metabolic rewiring [17].

We used the MDA-MB-231 breast cancer cell line and its specialized derivatives to feed data into *MetaboLiteLearner*. The choice of these cells serves multiple purposes. First, our previous work with these cells has provided detailed data as a solid ground for validating *MetaboLiteLearner* [13]. This earlier research showed changes in the intracellular metabolome between brain- and lung-homing cells and their parent cells, indicating significant metabolic rewiring. Moreover, re-analyzing these cells with *MetaboLiteLearner* could reveal new metabolic changes and unidentified metabolites that were overlooked in the initial study. *MetaboLiteLearner* leverages alterations in all metabolites, even unidentified ones, allowing us to explore metabolic changes in areas of the metabolic network that are typically not covered by textbook metabolic models.

We cultured all three cell variants, and the metabolites were harvested during balanced growth, which is when cells divide at exponential growth and before any slowdown in growth due to confluency (Fig. 2A). We then extracted soluble metabolites, dried all samples and controls, and derivatized the metabolites. After collecting the data with GC/MS with EI in scan mode [23], all the samples were aligned to consistent retention (6–30 min at 0.01-min intervals) and m/z (50 to 600 m/z at 1 m/z steps) ranges. The resulting individual data matrices were joined into a singular "virtual bulk sample" matrix (Fig. 2B). Spectrum extraction from this composite data enabled us to determine the $\log_2$ fold change in metabolite abundance in the organ-homing cells relative to the parental cells through a linear mixed-effects model. Our input for *MetaboLiteLearner* consisted of spectra arrays from m/z intervals of 50 to 600, normalized to their norms. The dataset, encompassing 153 unique spectra alongside their respective $\log_2$ fold-changes, is now presented as the *MetaboLiteLearner* Open Dataset (MLOD).

Training and evaluating *MetaboLiteLearner*

Using the MLOD, we first determined the optimal number of latent components, N_{opt}. Through hold-out cross-validation, as we increased N from 1 (simplest model) to 30 (most complex), the training error dropped monotonically, suggesting a better fit of the training data with increased model complexity (Fig. 3A). The test error reached a minimum at $N = 11$ components before rising, indicating potential overfitting in the more complex models. Utilizing a conservative methodology from supervised learning [19], we selected $N_{opt} = 5$, which showcased the smallest test error within one standard error. The $N_{opt} = 5$ model predictions correlated robustly with the actual log2 fold changes ($\rho = 0.39$, P-value $\ll 0.01$). We should note that other datasets may require a different value of N_{opt}; N_{opt} should be empirically determined for every dataset.

The model with $N_{opt} = 5$, when trained on the entire MLOD, optimally adjusted its loadings to maximize covariance between the projections of both inputs (the spectra) and outputs (log2 fold changes) onto the five-dimensional latent space (Fig. 3B). Unlike unsupervised methods such as principal component analysis, which primarily focuses on variance within individual datasets, the PLSR emphasizes joint variance optimization [19]. The resultant transformations into the latent space accounted for 32% of the predictor variance and 68% of the response variance. The variance explained by each latent factor for the predictor and response can be analyzed separately (Fig. 3B' and Fig. 3B", respectively).

Next, we performed a randomization test to determine our model's ability to capture biologically relevant patterns, not statistical artifacts. By shuffling the log2 fold changes, we retained internal correlations but disrupted correlations with input spectra. Notably, *MetaboLiteLearner*'s error, when trained with these shuffled data, was consistently worse than with the original dataset (Fig. 3C). This reaffirms our model's capability to discern significant links between metabolite spectra and abundance shifts in rewired cells.

MetaboLiteLearner reveals metabolic changes in metastatic breast cancer cells

The optimal model with five latent components ($N_{opt} = 5$) trained on the entire MLOD can transform a spectrum (a 500-dimensional vector) into $\log_2$ fold changes for brain- and lung-homing cells (a two-dimensional vector). This transformation can be visualized using a biplot, with each m/z ionic fragment represented by a vector (Fig. 1A). The biplot shows that specific fragments, such as $m/z = 104$ which arises from the EI fragmentation of TMS derivatives of amino acids [26], correlate with increased levels in both cell types. In contrast, fragments like $m/z = 306$, associated with EI-fragmentation of certain TMS-derivatized sugars [26], relate to decreased levels in organ-homing compared to parental cells.

Our previous study used an extensive panel of metabolites measured in targeted mode complemented with stable isotope tracing to analyze these cells [17]. In that study, lung-homing cells displayed a pronounced Warburg effect, underscored by a high lactate dehydrogenase to pyruvate dehydrogenase expression ratio, a potential biomarker for lung metastasis.

To determine whether *MetaboLiteLearner* was discerning patterns of metabolic rewiring consistent with our previous study [17], we analyzed a set of 263 EI-fragmentation spectra for TMS-derivatized metabolites [23] which we grouped into seven categories from the KEGG database of compounds with biological roles [22]. We fed the spectra into our trained model to predict fold changes. While most metabolites showed concurrent changes in both cell types, specific carbohydrates, and nucleic acids increased only in lung-homing cells (Fig. 4A).

The five latent component model explains 67.7% of the variance in $\log_2$ fold changes (Fig. 3B). This indicates that the model fits the data reasonably well. But beyond the quality of the fit, one of PLSR's strengths lies in the possibility of dissecting the

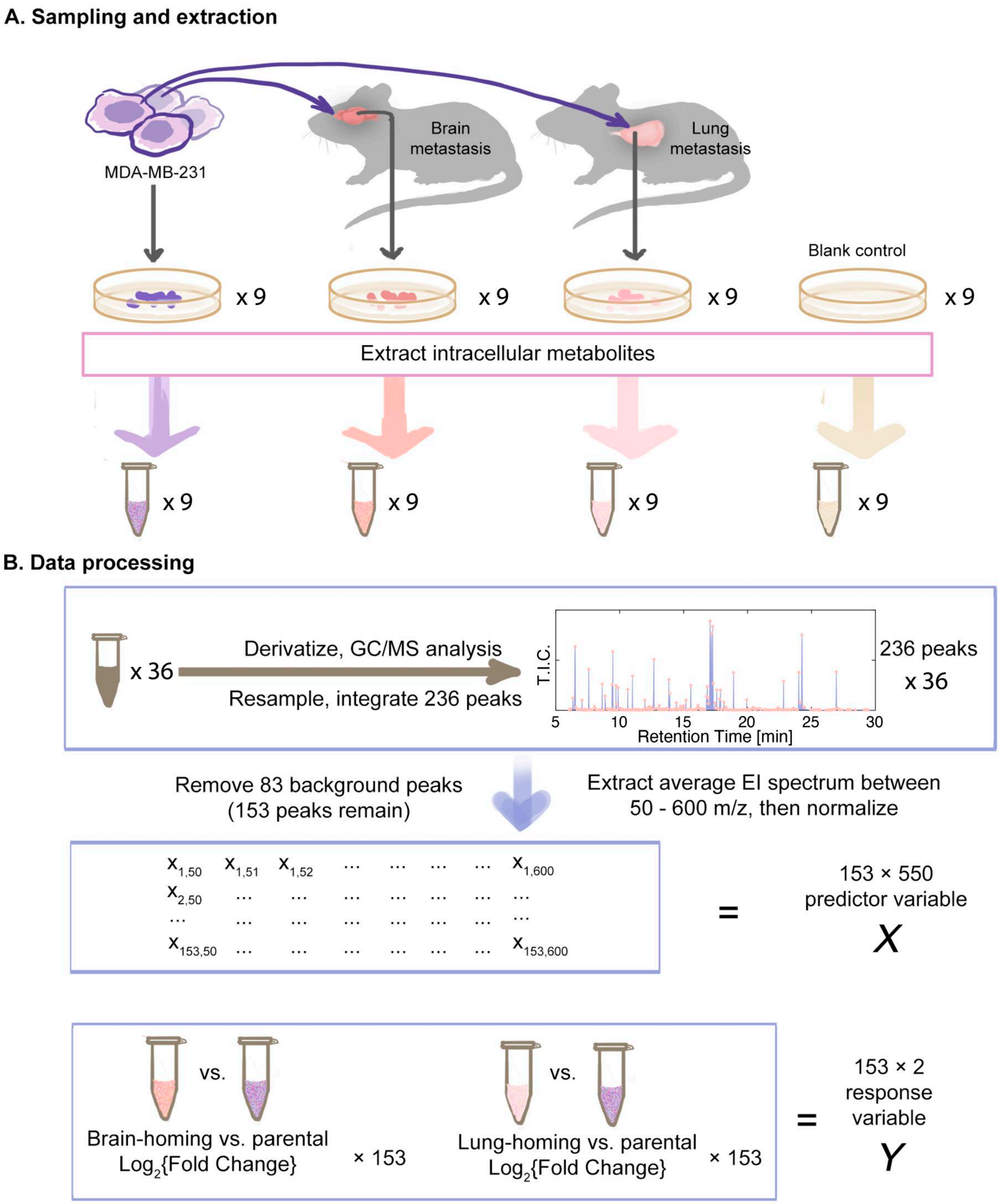
A. Sampling and extraction
MDA-MB-231
Brain metastasis
Lung metastasis
Blank control
x 9
x 9
x 9
x 9
Extract intracellular metabolites
x 9
x 9
x 9
x 9
B. Data processing
x 36
Derivatize, GC/MS analysis
Resample, integrate 236 peaks
T.I.C.
5
10
15
20
25
30
Retention Time [min]
236 peaks
x 36
Remove 83 background peaks (153 peaks remain)
Extract average EI spectrum between 50 - 600 m/z, then normalize
x1,50 x1,51 x1,52 … x1,600
x2,50
x153,50 … x153,600
=
153 × 550 predictor variable
X
vs.
Brain-homing vs. parental Log2{Fold Change}
× 153
vs.
Lung-homing vs. parental Log2{Fold Change}
× 153
=
153 × 2 response variable
Y

Figure 2. Data Acquisition and Processing for *Metabolitelearner* from Breast Cancer cell derivatives. **(A)** Cell culture and metabolite extraction: Breast cancer cell lines, including the parental MDA-MB-231 cells and its brain- and lung-homing derivatives, were cultivated. These derivatives were procured through *in vivo* selection using mice. Under consistent media conditions *in vitro*, intracellular metabolites from these cells were extracted to ensure a reliable data source for subsequent processing. **(B)** GC/MS processing and data aggregation: Samples underwent GC/MS analysis following the TMS derivatization protocol. The generated data matrices, unique for each sample, were amalgamated to create a virtual "bulk" sample. Peaks were identified, and their spectra were extracted from this consolidated matrix. The input (X) for *MetaboLiteLearner* encompasses the mass spectra of each peak. The output data (Y) indicate the comparative abundance shift of each peak in brain- and lung-homing cells relative to the parental cells.

model's coefficients to shed light on the molecular shifts underpinning cellular adaptations.

Components 1 and 3 showcase metabolites with decreased levels in both cell types, indicating overlapping metabolic shifts (Fig. 4B). Components 2 and 5 indicate metabolites with increased levels in both cell types. Component 4 highlights differences between the cell types—some metabolites decrease in brain-homing but increase in lung-homing cells. Component 1, which explains 27% of the response variance, captures a common trend—reduced levels in both cell types. Evaluating spectra from various compound classes, most amino acids follow this trend, whereas carbohydrates deviate from it (Fig. 4C). Analyzing all

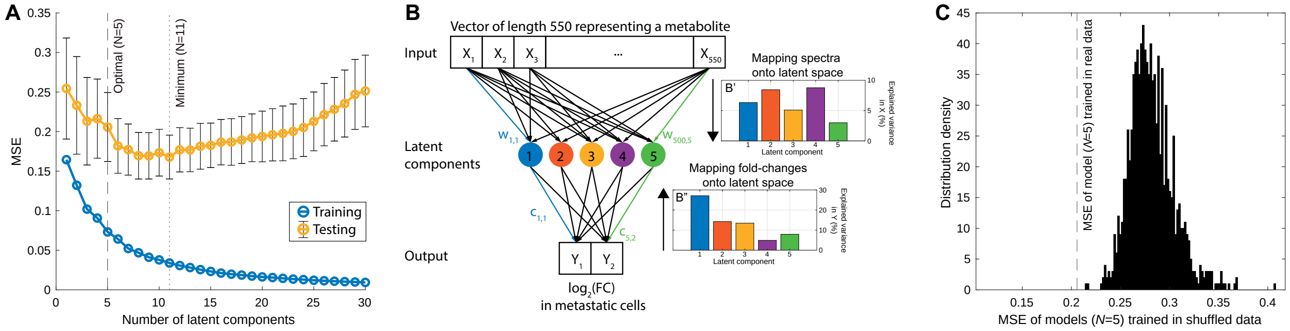

Figure 3. Optimization and Evaluation of *MetaboLiteLearner*'s Predictive Performance. **(A)** Model Complexity vs. Error: Through hold-out cross-validation, the training error consistently decreased as latent components (N) increased from 1 to 30. Test error reached its lowest at $N = 11$ before it began to rise, highlighting potential overfitting with more complex models. The chosen optimal model had $N_{opt} = 5$ components, correlating strongly with true log2 fold changes ($\rho = 0.39$, P-value $\ll 0.01$). **(B)** Schematic of the model trained with five components. **(B' and B")** Variance explained by latent factors: Using the $N_{opt} = 5$ model, the transformations into the five-dimensional latent space covered 32% of the predictor variance and 68% of the response variance. The variance explained by each latent factor for the predictor and response datasets can be viewed separately. **(C)** Randomization test results: After shuffling the log2 fold changes, disrupting their correlations with input spectra, *MetaboLiteLearner*'s error with shuffled data was consistently higher than with the original dataset, confirming its ability to identify genuine relationships between metabolite spectra and abundance changes in rewired cells.

latent components (Fig. 4D–G) revealed component 4's unique role. Despite accounting for just 4.9% of the response variance, it distinctly captures the variation between brain- and lung-homing cells. Carbohydrates and deoxyribonucleosides dominate this component, indicating potential metabolic shifts between the two cell types. These observed metabolic changes could reflect adaptations to the unique environments of the brain and lungs. With its rich blood supply, the lung might favor carbohydrate-utilizing cells. Conversely, the elevated deoxyribonucleosides in lung-homing cells could suggest more robust DNA repair mechanisms than in brain-homing cells. These findings align with our previous work [17] and suggest potential pathways for further exploration.

Discussion

In this paper, we showcase the capabilities of *MetaboLiteLearner*, a supervised learning method for analyzing metabolomic changes in metabolically rewired cells. We conducted experiments on intracellular metabolites from a breast cancer cell line and its derivatives that are known to migrate to the brain and lungs. This dataset is relevant for oncology research and demonstrates the effectiveness of our approach. Through computational experiments involving cross-validation and data shuffling, we have shown that *MetaboLiteLearner* can identify patterns even in a relatively small dataset consisting of 153 unidentified metabolites that are significantly altered in organ-homing cells when compared to their parent cells, without needing prior knowledge of metabolic networks.

A defining feature of *MetaboLiteLearner*'s is its use of PLSR—a computationally lightweight and robust statistical algorithm, especially useful for smaller datasets. As noted in previous work, PLSR can handle internal correlations within predictors and responses, which arise due to shared ion fragments produced by naturally abundant isotopes and functional groups shared among certain metabolite classes and the universally witnessed metabolic shifts in disseminated cells [17]. PLSR handles these internal correlations by generating a reduced-dimensional latent space that maximizes the joint variance between predictors and responses. Here, the latent space represents the leading associations between a metabolite's molecular features (the ions produced by EI fragmentation) and its abundance change in rewired cells. While the current implementation of PLSR in our model is efficient, there is potential for further refinement, such as using regularization and Laplacian constraints [27]. Incorporating regularization might improve model interpretability. Meanwhile, Laplacian constraints could allow the model to incorporate existing knowledge from comprehensive spectral libraries and metabolic network models.

Other supervised approaches used for metabolomics analysis, such as Partial Least Squares Discriminant Analysis (PLS-DA), typically consider each sample as a data point in the supervised analysis with sample types as the labels [28]. *MetaboLiteLearner* operates differently: each metabolite is treated as a data point, and the labels are the changes in metabolite abundances between different conditions. Focusing on metabolite-level data points allows *MetaboLiteLearner* to associate molecular fragments to the observed changes, which would be overlooked in traditional PLS-DA methodologies.

Machine learning tools generally require sufficient training data to perform properly, which is true for *MetaboLiteLearner*. By treating each metabolite as a data point rather than each sample, *MetaboLiteLearner* increases the number of data points, thereby enhancing model robustness. Our study has 36 samples but detected 153 metabolites, providing a larger dataset for training the model. This approach, combined with cross-validation and shuffling tests outlined in our code repository [25], helps ensure the model is well-validated and avoids overfitting.

In this study, we also introduce the MLOD, an open dataset tailored for machine learning research focusing on the metabolic reconfigurations in cancer cells. The MLOD, enriched with meticulously captured spectra and corresponding abundance shifts, can be a standard for benchmarking future supervised learning endeavors in this domain. Our decision to use GC/MS spectra from TMS-derivatized samples within a specific interval was primarily driven by its alignment with prevalent practices [23]. However, we recognize the potential benefits of integrating data from high-resolution instruments like time-of-flight [29] or Orbitrap [30]. These devices could capture finer details, thus enhancing model predictions. Furthermore, harnessing data generated by tandem instruments, such as MS/MS or MS2, can lead to more sophisticated computational strategies, as seen in works external to cancer research [31–33].

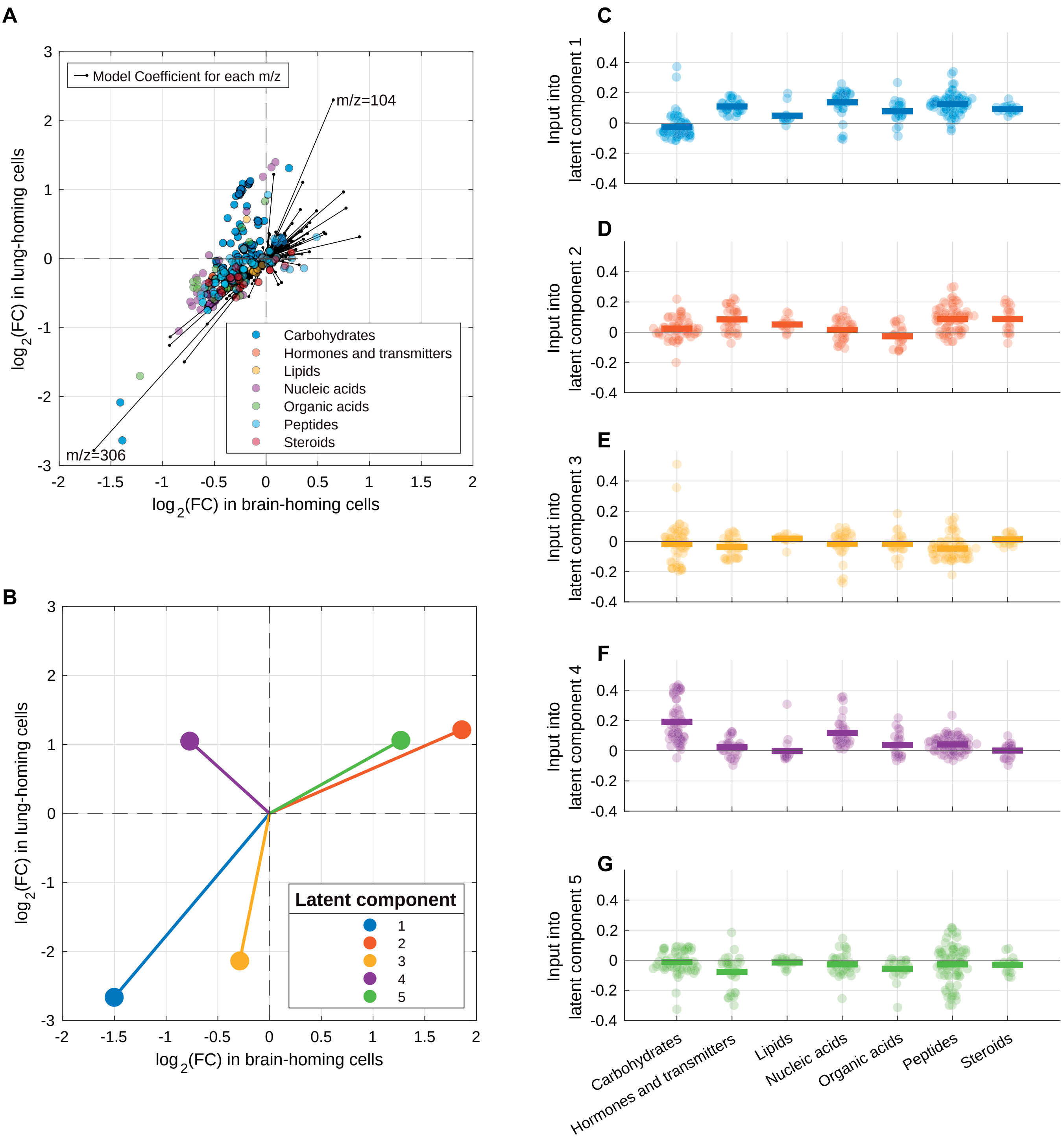

Figure 4. Interpretation of the Model for Metastatic Breast Cancer Cells. **(A)** Biplot representation of the m/z ionic fragment vectors. Specific fragments, such as m/z = 104, are associated with increased levels in both cell types in contrast to fragments like m/z = 306 which are associated with decreased levels. **(B)** The proportional contribution of the five latent components to the variance in log2 fold changes, with components 1 and 3 indicating overlapping metabolic shifts with decreased levels in both cell types, components 2 and 5 showcasing increases in both, and component 4 underscoring the divergence between the two cell types. **(C)** In component 1, accounting for 27% of the response variance, most amino acids follow the trend of reduced levels in both cell types, while carbohydrates differ. **(D–G)** Among the latent components, we see the distinctive role of component 4 which is dominated by carbohydrates and deoxyribonucleosides and highlights potential metabolic variances between brain- and lung-homing cells.

We utilized TMS derivatization because it enhances the volatility and reproducibility of metabolites detected by GC/MS. This derivatization introduces Si-related isotopic patterns in the EI spectra, which are characteristic of the derivatization process and can aid in identifying specific functional groups. This raises the question of whether *MetaboLiteLearner* performance depends on the spectra's nature, particularly whether it can be applied to non-derivatized samples. The feature selection process relies on the fragmentation patterns captured in the spectra and should in princple be applicable to non-derivatized samples. However, the specific features and patterns used by the model may differ. Future work should explore the application of

MetaboLiteLearner to non-derivatized samples to validate its performance and generalizability across different sample preparation methods.

Our dataset focused on $\log_2$ fold changes from unsynchronized intracellular metabolite concentrations. This has constraints, as the insights obtained mainly reflect consistent changes across a generalized cell population. However, integrating synchronized data acquisition at different cellular growth stages [34] or incorporating stable isotope tracing [35] can inject dynamic elements into future datasets, allowing our models to capture more nuanced metabolic variations. While this study centered on breast cancer, the potential uses of *MetaboLiteLearner* extend beyond: it may be applied to a diverse range of cell types, conditions, and even less homogeneous samples like tissues or tumors. In our case study, the response variable was bidimensional, capturing metabolite abundance changes in brain- and lung-homing cells relative to their parental lineage. Incorporating additional dimensions and more comprehensive datasets could require more sophisticated computational approaches, like deep neural networks, enabling a broader and deeper exploration of metabolic alterations across different cell conditions and types.

This study showcases the potential synergy between machine learning and metabolomics. With evolving datasets and improving computational methods, we can make new strides in unraveling the intricacies of metabolic rewiring—a fundamental aspect of cellular adaptation.

Acknowledgements

This work was supported by the National Institutes of Health (NIH) grant R01 CA266068 “Mathematical modeling of metabolism rewiring in cancer eco-evolution and metastasis tropism” to J.B.X. Acknowledgments to Chen Liao, Deepti Mathur, Heesoo Seong, Julia Simundza and Carlos Carmona Fontaine for their comments on the manuscript.

Author contributions

Joao Xavier (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal])

Conflict of interest statement. None declared.

References

1. Gomes AP, Blenis J. A nexus for cellular homeostasis: the interplay between metabolic and signal transduction pathways. *Curr Opin Biotechnol* 2015;**34**:110–7.
2. Moxley JF, Jewett MC, Antoniewicz MR *et al.* Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator Gcn4p. *Proc Natl Acad Sci U S A* 2009;**106**:6477–82.
3. Miyazawa H, Aulehla A. Revisiting the role of metabolism during development. *Development* 2018;**145**(19):dev131110.
4. Chapman NM, Chi H. Metabolic adaptation of lymphocytes in immunity and disease. *Immunity* 2022;**55**:14–30.
5. Sengupta N, Rose ST, Morgan JA. Metabolic flux analysis of CHO cell metabolism in the late non-growth phase. *Biotechnol Bioeng* 2011;**108**:82–92.
6. Schmidt DR, Patel R, Kirsch DG *et al.* Metabolomics in cancer research and emerging applications in clinical oncology. *CA Cancer J Clin* 2021;**71**:333–58.
7. Lu W, Su X, Klein MS *et al.* Metabolite measurement: pitfalls to avoid and practices to follow. *Annu Rev Biochem* 2017;**86**:277–304.
8. Wieder C, Frainay C, Poupin N *et al.* Pathway analysis in metabolomics: recommendations for the use of over-representation analysis. *PLoS Comput Biol* 2021;**17**:e1009105.
9. Dunn WB, Erban A, Weber RJM *et al.* Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* 2013;**9**:44–66.
10. Want EJ, Wilson ID, Gika H *et al.* Global metabolic profiling procedures for urine using UPLC-MS. *Nat Protoc* 2010;**5**:1005–18.
11. Goodacre R, Broadhurst D, Smilde AK *et al.* Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics* 2007;**3**:231–41.
12. Galal A, Talal M, Moustafa A. Applications of machine learning in metabolomics: disease modeling and classification. *Front Genet* 2022;**13**:1017340.
13. Camacho DM, Collins KM, Powers RK *et al.* Next-generation machine learning for biological networks. *Cell* 2018;**173**:1581–92.
14. Cailleau R, Olivé M, Cruciger QV. Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization. *In Vitro* 1978;**14**:911–5.
15. Minn AJ, Gupta GP, Siegel PM *et al.* Genes that mediate breast cancer metastasis to lung. *Nature* 2005;**436**:518–24.
16. Bos PD, Zhang XH-F, Nadal C *et al.* Genes that mediate breast cancer metastasis to the brain. *Nature* 2009;**459**:1005–9.
17. Mathur D, Liao C, Lin D *et al.* The ratio of key metabolic transcripts is a predictive biomarker of breast cancer metastasis to the lung. *Cancer Res* 2023;**83**(20):3478–91.
18. Agilent G1676AA Fiehn GC/MS Metabolomics RTL Library.
19. Hastie T, Tibshirani R, Friedman J. (2009) *The Elements of Statistical Learning* (Springer New York, New York, NY). 2nd Ed.
20. de Jong S. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* 1993;**18**:251–63.
21. Hsiao T-C, Lin C-W, Zeng M-T, *et al.* The implementation of partial least squares with artificial neural network architecture, in *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol.20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No.98CH36286)*, Hong Kong, China, 1998; 1341–3.
22. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
23. Kind T, Wohlgemuth G, Lee DY *et al.* FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal Chem* 2009;**81**:10038–48.
24. Xavier J. Raw GC/MS data used for MLOD (MetaboLiteLearner). *Zenodo* 2023; https://doi.org/10.5281/zenodo.8193580.
25. Xavier J. joaobxavier/learn_metabolic_rewiring_matlab: metab oLiteLearner08032023. *Zenodo* 2023; https://doi.org/10.5281/zenodo.8213019.
26. Lai Z, Fiehn O. Mass spectral fragmentation of trimethylsilylated small molecules. *Mass Spectrom Rev* 2016;**37**:245–57.

27. Li X, Panea C, Wiggins CH *et al.* Learning “graph-mer” motifs that predict gene expression trajectories in development. *PLoS Comput Biol* 2010;**6**:e1000761.
28. Hadi NI, Jamal Q, Iqbal A *et al.* Serum metabolomic profiles for breast cancer diagnosis, grading and staging by gas chromatography-mass spectrometry. *Sci Rep* 2017;**7**:1715.
29. Da Cunha PA, Nitusca D, Canto LMD *et al.* Metabolomic analysis of plasma from breast cancer patients using ultra-high-performance liquid chromatography coupled with mass spectrometry: an untargeted study. *Metabolites* 2022;**12**(5):447.
30. An R, Yu H, Wang Y *et al.* Integrative analysis of plasma metabolomics and proteomics reveals the metabolic landscape of breast cancer. *Cancer Metab* 2022;**10**:13.
31. Watrous J, Roach P, Alexandrov T *et al.* Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci U S A* 2012;**109**:E1743–52.
32. Tripathi A, Vázquez-Baeza Y, Gauglitz JM *et al.* Chemically informed analyses of metabolomics mass spectrometry data with Qemistree. *Nat Chem Biol* 2021;**17**:146–51.
33. Stravs MA, Dührkop K, Böcker S *et al.* MSNovelist: de novo structure generation from mass spectra. *Nat Methods* 2022;**19**:865–70.
34. Diehl FF, Miettinen TP, Elbashir R *et al.* Nucleotide imbalance decouples cell growth from cell proliferation. *Nat Cell Biol* 2022;**24**:1252–64.
35. Jang C, Chen L, Rabinowitz JD. Metabolomics and isotope tracing. *Cell* 2018;**173**:822–37.

Biology Methods and Protocols, 2024, 00, 1–10
https://doi.org/10.1093/biomethods/bpae048
Methods Article