Personalized Metabolic Analysis of Diseases

Ali Cakmak and M. Hasan Celik

Abstract—The metabolic wiring of patient cells is altered drastically in many diseases, including cancer. Understanding the nature of such changes may pave the way for new therapeutic opportunities as well as the development of personalized treatment strategies for patients. In this paper, we propose an algorithm called Metabolitics, which allows systems-level analysis of changes in the biochemical network of cells in disease states. It enables the study of a disease at both reaction- and pathway-level granularities for a detailed and summarized view of disease etiology. Metabolitics employs flux variability analysis with a dynamically built objective function based on biofluid metabolomics measurements in a personalized manner. Moreover, Metabolitics builds supervised classification models to discriminate between patients and healthy subjects based on the computed metabolic network changes. The use of Metabolitics is demonstrated for three distinct diseases, namely, breast cancer, Crohn's disease, and colorectal cancer. Our results show that the constructed supervised learning models successfully differentiate patients from healthy individuals by an average f1-score of 88 percent. Besides, in addition to the confirmation of previously reported breast cancer-associated pathways, we discovered that Biotin Metabolism along with Arginine and Proline Metabolism is subject to a significant increase in flux capacity, which have not been reported before.

Index Terms—Systems biology, biomedical informatics, classification algorithms, metabolomics, supervised learning

1 Introduction

The phenotype of diseases has often reflections on the metabolism of patients [1], [11], [18], [20], [42]. Certain pathways may be boosted, while some others may experience activity decrease. Collectively, such changes may explain the etiology of a disease [18], [44]. In this paper, we propose an algorithm, Metabolitics, which quantifies the changes in the activity levels of pathways given concentration fold changes for a set of metabolites. It later employs the computed metabolic activity changes as features to build classification models that predict whether an individual has a disease of interest or not.

A number of past studies (e.g., [13], [24], [56], [71], [77]) focused on the pathway/reaction level analysis of high-throughput biological data. The initial set of studies mainly employed pathway enrichment analysis ([56], [59], [68], [71], [72], [73], [74], [75], [76]). Briefly, these methods first identify significantly changing genes/metabolites in a given omics dataset. Then, the identified metabolites/genes are mapped to the pathways that they participate in. Based on the number of changing metabolites/genes included in each pathway, statistical significance analysis is performed to identify those pathways that are overrepresented within the measured genes or metabolites. Accordingly, each pathway is assigned a score based on the computed statistical test values. iOmicsPass [67] extends these approaches from a node-level view to an edge-level view in the context of biological networks. More specifically, it does not consider the measured entities by themselves but focuses on the direct interactions of the measured biological entities. To do this, it first computes co-expression scores for interactions based on their source and target entities' Z-scores. Next, using the interaction scores, it computes the significantly differing subnetworks. Then, on these computed subnetworks, pathway enrichment analysis is performed. XCMS Online [78] combines statistical significance-based filtering at both metabolite and pathway levels. The next set of methods ([13], [33], [50]) directly transfers the measured metabolite/gene changes to their corresponding pathways without any filtering. Then, statistical significance analysis is performed on the pathways and their change levels to assign them deregulation scores. These methods consider each pathway as a collection of genes/metabolites and ignore the interactions between these entities. An extension of the above set of methods also considers pathway topology ([12], [28], [52], [54], [69]). In particular, some measured genes/metabolites are given more weights in the statistical significance analysis based on the centrality of each gene/metabolite in the pathway topology. The centrality may be computed with different measures, e.g., betweenness, eigenvector centrality, neighborhood size, etc. [27].

The state of the art in this particular field is represented by Pathifier [13] and Paradigm [54]. Pathifier considers each pathway as a metabolite vector by matching the measured metabolites with the pathways that they participate. In this vector, each entry represents the concentration change measured for a metabolite. Then, Pathifier computes the principal curve that best fits the region where each individual is represented as a point in this high-dimensional space. On this curve, the mean distance from the point representing a person to the points representing healthy individuals is called the pathway dysregulation score which represents how

- *Ali Cakmak is with the Department of Computer Engineering, Istanbul Technical University, Maslak, 34467 Istanbul, Turkey. E-mail: ali.cakmak@gmail.com.*
- *M. Hasan Celik is with the Department of Computer Science and Engineering, Istanbul Sehir University, Dragos, 34865 Istanbul, Turkey. E-mail: hasancelik@std.sehir.edu.tr.*

Manuscript received 29 Jan. 2019; revised 3 June 2020; accepted 12 June 2020. Date of publication 9 July 2020; date of current version 3 June 2021. (Corresponding authors: Ali Cakmak.)
Digital Object Identifier no. 10.1109/TCBB.2020.3008196

much the pathway activity has altered in that individual. For each pathway, Pathifier fits a distinct curve.

Paradigm [54] is another successful approach in the pathway-based analysis domain. It employs probabilistic graph models to compute probability values for each pathway. A pathway is turned into what is called a *factor graph*. A factor graph contains a set of nodes representing different biological entities and their states. As an example, for each gene in a pathway, the factor graph contains the following nodes: gene, mRNA, protein, and active protein. The factor graph contains an edge for any type of information flow or state change between these nodes. For instance, there is an edge from a gene to its mRNA representing its transcription, an edge from mRNA to protein representing its translation, etc. Each node in a factor graph represents a variable that can have one of the following values: 1 (activation), 0 (normal), and -1 (de-activation). The values of variables are learned from the provided omics data.

None of the above-summarized pathway analysis methods considers the fact that pathways are part of a large biological network, and they interact with each other. The main novelty of the proposed method in this paper is that the analysis is performed on the whole pathway network in a holistic manner, rather than considering each pathway in an isolated manner. The core advantage of such an approach is that, for a given disease, it allows identifying those key player pathways for which there may be few or no associated gene/metabolite measurements in the analyzed omics data. Our method is not specific to the metabolomics domain. It may be easily extended to be used for the analysis of other types of omics data as well (e.g., [79] presents our preliminary results with gene expression data). Such an extension usually requires customizing the constraints and the objective function of an optimization model as also illustrated in several other works (e.g., [80], [81], [82], [83]).

In brief, Metabolitics assigns a score for each pathway/reaction in a patient. This score represents how much the activity of the corresponding pathway/reaction differs from that of healthy individuals. In order to achieve this, Metabolitics works on the whole network of metabolic pathways. In particular, it turns the analysis task into an optimization problem [38], where the objective is dynamically set to maximize the flux for increasing metabolites' reactions and minimize the flux for decreasing metabolites' reactions in proportion to their fold changes. The metabolic network is assumed to be in steady-state [35], that is, the amounts of consumption and production are equal for all the metabolites. The steady-state requirements are represented as constraints. Then, the optimization problem is solved using linear programming [14]. Since the optimization problem is under-determined [38], there are usually multiple solutions. In order to accommodate multiple solutions with a single score, we employ flux variability analysis [31], which identifies the lower and upper flux bounds for each pathway. The average lower and upper flux values of healthy individuals are considered as reference values. Then, for each pathway, Metabolitics computes how much the lower and upper flux values differ from the reference values in a given patient. Finally, pathway diff scores are derived from their reactions' flux boundary values.

Once "diff" scores are computed, Metabolitics builds supervised-learning models to predict whether an individual carries a disease of interest based on the computed metabolic changes. In particular, each individual is represented as a vector of the computed pathway diff scores.

Several other works combine machine learning and metabolomics analysis ([3], [8], [70], and [85] for recent reviews). Heckman *et al.* [40] employ regression models to predict the catalytic turnover rates of enzymes in a genome-scale metabolic model based on features such as metabolite concentrations, average flux, enzyme structure, etc. MFlux [37] constructs supervised learning models to estimate the fluxome of bacterial metabolism as a low-cost alternative to ^{13}C metabolic flux analysis. To this end, it utilizes the strain of bacteria, types of substrates, growth rate, and environmental conditions such as oxygen levels as features. GEESE [32] is an FBA approximator for bacteria that employs gene expression data as well as external glucose and oxygen concentrations as input. The estimation algorithm is based on a deep generative model, namely, a variational autoencoder. Similarly, Guo and Feng [25] study the use of deep learning to predict the phenotype based on transcriptomic data enriched with flux balance analysis results over metabolic networks. Yaneske and Angione [21] propose a metabolic age predictor based on multi-omics-integrated constraint-based models. The underlying model employs elastic net regression to estimate the metabolic predictors of aging. Toubinana *et al.* [17] combine metabolite correlation-based network analysis with machine learning models to predict unknown pathways. Costello and Martin [15] develops a supervised learning model that predicts metabolite concentrations based on a time series of proteomics and metabolomics measurements used as training data.

In order to evaluate the Metabolitics algorithm, we apply it to breast cancer, Crohn's disease, and colorectal cancer. We demonstrate that Metabolitics (i) captures biologically relevant information, (ii) accurately predicts disease status of subjects using supervised learning models, (iii) is robust to decrease in the amount of measured metabolite data, and (iv) provides more metabolic network coverage than the state of the art.

We have also implemented Metabolitics as a web tool, MetaboliticsDB, which is described in another work [7].

2 Methods

In this section, the details of the Metabolitics algorithm are presented. Metabolitics employs Recon2 [53] as the metabolic network data. Recon2 is a genome-scale reconstructed human metabolic network model that includes 5324 metabolites and 7785 reactions. It features 100 pathways which are non-overlapping subgraphs of the metabolic network; thus, each has a unique set of reactions. Metabolitics is not specific to Recon2, and it is generic enough to adapt to other network models. In essence, the Metabolitics pipeline consists of the following steps. Fig. 1 pictorially depicts these steps which (except for data splitting) are explained below. Data splitting is discussed in Section 3 as part of the cross-validation strategy.

1. Matching the names of metabolites in the input metabolomics data to the metabolites in Recon.

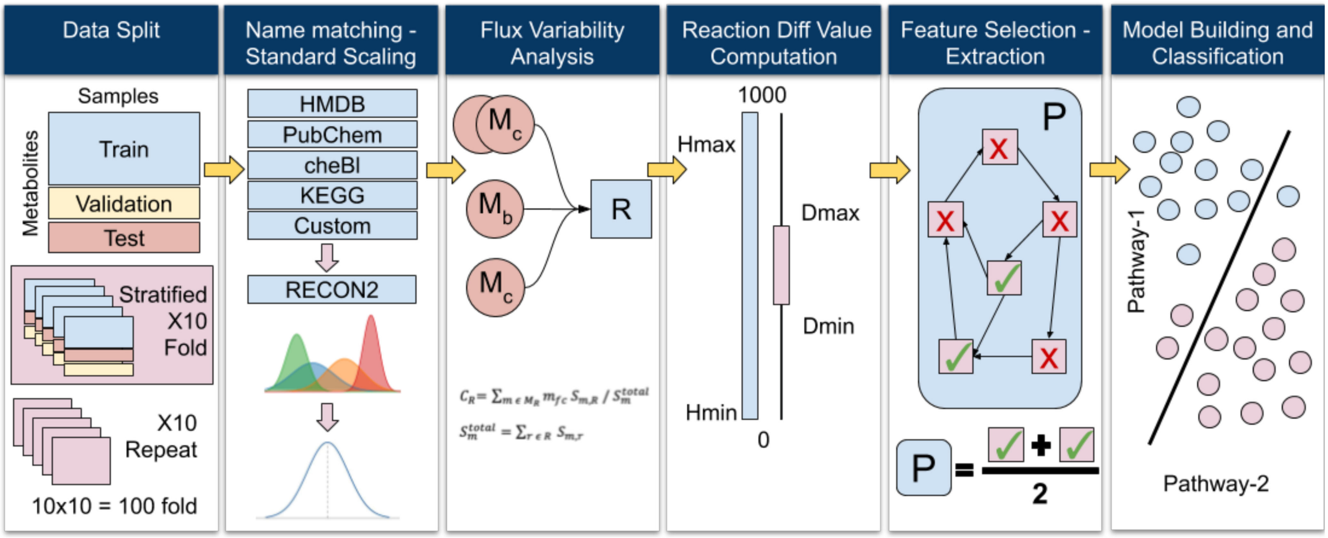
Data Split
Samples
Metabolites
Train
Validation
Test
Stratified
X10
Fold
X10
Repeat
10x10 = 100 fold
Name matching - Standard Scaling
HMDB
PubChem
cheBI
KEGG
Custom
RECON2
Flux Variability Analysis
M_c
M_b
M_c
R
$C_R = \sum_{m \in M_R} m_{fc} S_{m,R} / S_m^{total}$
$S_m^{total} = \sum_{r \in R} S_{m,r}$
Reaction Diff Value Computation
1000
Hmax
Dmax
Dmin
Hmin
0
Feature Selection - Extraction
P
P = (✓ + ✓) / 2
Model Building and Classification
Pathway-1
Pathway-2

Fig. 1. Metabolitics pipeline steps.

2. Dynamic creation of a linear programming model.
3. Flux variability analysis.
4. Calculation of reaction and pathway diff values.
5. Statistical significance analysis.
6. Feature extraction.
7. Feature selection.
8. Building a supervised classification model

2.1 Matching Metabolite Names

It may not be possible to match all the metabolite names in a given metabolomics dataset to those in Recon2 [41]. The main reason is that many metabolites do not have a standard common name used by all researchers. Therefore, as a first step, one needs to find the equivalents of user-provided metabolites in the Recon2 data. In this study, the set of all known alternative names for each metabolite is compiled by the cross-integration of different metabolite data sources. More specifically, the names of the metabolites that are included in two widely used public data sources, namely, HMDB [58] and CheBI [23], are combined. In particular, a pair of metabolite entries from two different sources are considered the same, if they share at least one synonym. Next, each metabolite of Recon2 is searched in this combined name dataset, and all known synonyms are transferred to Recon2. Finally, the names in the input metabolomics dataset are matched to the compiled set of synonyms.

2.2 Personalized Linear Programming Models

Metabolic analysis studies in the literature generally assume that metabolic networks are in steady-state [30], [34]. The steady-state hypothesis considers that the total production of each metabolite in the metabolic network is equal to its total consumption. This allows a metabolic network to be expressed mathematically as a set of linear equations. More specifically, a metabolic network may be represented as a matrix in which the rows correspond to the metabolites and the columns correspond to the reactions in the network. In this matrix, a cell located at row i and column j contains the stoichiometry of metabolite m_i in reaction r_j. If metabolite m_i does not participate in reaction r_j, then 0 is placed in the corresponding cell. With this representation, flux balance analysis may be expressed as a linear program under the steady-state assumption as follows [38]:

$$\begin{aligned}&maximize\ C^T\ V\\&subject\ to\ S\ \times\ V\ =\ 0\ and\ v_{lower}\ <\ V\ <\ v_{upper}\end{aligned} \tag{1}$$

In eq. (1), S is the matrix representation of the metabolic network, V is a vector of variables which represent reaction fluxes, C contains coefficients for reactions, v_{lower} and v_{upper} are the upper and lower boundaries of reaction fluxes. By solving the above linear program, the values of the reaction flux variables in vector V are determined. At this point, the steady-state assumption is represented as constraints. In single-cell organisms, the objective function is often set to maximize the fluxes of reactions that produce cell building blocks (amino acids, nucleotides, lipids, etc.) to amplify the cell biomass. However, for multi-cellular organisms, e.g., humans, there is no agreed-upon standard for the objective function structure. In this study, the objective function is dynamically set in a personalized manner. More specifically, the objective function (i.e., $C^T\ V$) is constructed as follows: V includes flux variables for all reactions that produce at least one metabolite in the input metabolomics dataset. C is a vector of coefficients. Each entry C_R in C is associated with the flux variable of a unique reaction R from V. Eq. (2) shows how C_R is computed.

$$\begin{aligned}C_R &= \sum_{m\ \epsilon\ M_R} m_{fc}\ S_{m,R}\ /\ S_m^{total}\\S_m^{total} &= \sum_{r\ \epsilon\ R} S_{m,r},\end{aligned} \tag{2}$$

where

- C_R is the coefficient of reaction R's flux variable,
- M_R is the set of metabolites that are produced by R,
- m_{fc} is the concentration fold change for metabolite m, as computed in Eq. (3),
- $S_{m\ ,r}$ is the stoichiometry of metabolite m in R,
- S_m^{total} is the total stoichiometry of m over all producer reactions of m in the metabolic network.

The objective function is heuristically built based on the intuition that, by incorporating m_{fc} as a coefficient, the higher the change in the measured concentration of a metabolite, the more effect it will have on the objective function value. In doing so, the total stoichiometry of the metabolite is used as a normalization term to prevent the popular (currency) metabolites from artificially dominating the objective function. This is because, in the metabolic network, there are several hub (currency) metabolites that participate in many reactions. H_2O, ATP, and NAD are some examples of currency metabolites. Many studies [64] remove these metabolites from their analysis, but the removal of the currency metabolites may damage the stoichiometric balance of the metabolic network. Moreover, the removal of currency metabolites is a strong assumption; therefore, it should be avoided. In order to accommodate the above considerations, the currency metabolites are kept, and each metabolite is normalized with its total stoichiometry. Moreover, the objective function is defined completely based on the measured metabolite concentration changes which will be different for each person. This way of modeling allows our approach to produce personalized results. Besides, metabolite concentrations need to be normalized as fold-changes since the scale of each metabolite is different, and a change is only meaningful when compared to a reference healthy value. That is,

$$m_{fc} = \log m^c - \log \mu_m^{healthy}, \tag{3}$$

where

- m^c is the concentration measurement of metabolite m in an individual.
- m_{fc} is the concentration fold change of metabolite m for the individual.
- $\mu_m^{healthy}$ is the mean concentration of metabolite m in all healthy individuals.

In order to compute the fold-changes as in eq. (3), first, the reference mean concentration of each metabolite is computed based on the measurements in healthy individuals. Then, for all individuals (including healthy individuals and patients), the metabolite fold changes are computed relative to the above computed mean values. As a result, though usually smaller than patients, even healthy individuals are assigned metabolite fold changes. These fold changes are employed for each individual in the corresponding objective function, as described above.

2.3 Flux Variability Analysis

Typically, the number of metabolic reactions is greater than the number of metabolites in a metabolic network [38]. Hence, the optimization problem presented in the previous section is underdetermined. Therefore, many alternative solutions (i.e., reaction flux value assignment) are usually possible. In order to cover all of these alternative solutions, flux variability analysis (FVA) [36] is employed. FVA allows determining the minimum and maximum flux values for each reaction. In brief, FVA works as follows (Fig. 2): First, the optimal value of the objective function is determined by solving the optimization problem described in the previous section. Then, the computed optimal value of the objective function is added to the model as an additional constraint. For each reaction R in the metabolic network, the objective function is set to maximize the flux of R, and the optimization problem is solved. The computed objective function value is stored as the upper boundary for the flux of R. Next, the objective function is set to minimize the flux of R, and the optimization problem is solved once again. The computed objective function value is stored as the lower boundary for the flux of R. This process is repeated for each reaction in the metabolic network.

```
Flux variability analysis:
   where solveFBA returns $C^T V$

   input:   FBA problem as $P$
   output:  minimum and maximum feasible flux for each
            reaction of FBA problem as $Rmax$ and $Rmin$

   $solution$ = solveFBA($P$) with current objective
   $P$ also subject to additional constraint, $C^T V = solution$

   for each $r$ in $P.reactions$ do
       set maximization of $r$ as a new objective
       $Rmax_r$= solveFBA($P$)

       set minimization of $r$ as a new objective
       $Rmin_r$= solveFBA($P$)

   return $Rmax$, $Rmin$
```

Fig. 2. FVA algorithm [63].

2.4 Diff Value Computation

Pathways are defined by biologists as a set of closely related reactions that work together, often in a particular order, to achieve a common cellular goal (e.g., fatty acid synthesis). Analyzing the perturbations caused by diseases in metabolic networks in terms of changes in known pathways is a commonly used summarization method [13], [26]. In this study, a similar approach is adopted as well. To this end, for each pathway, we compute a “diff” value that represents the differentiation of the pathway activity for an individual in comparison to healthy individuals. The pathway diff values are computed as follows. First, the average lower and upper flux boundary values for each reaction in healthy individuals are computed as described above. These values are recorded as “reference” values. Then, during the analysis of a given set of metabolomics data of an individual, the overall differences between the lower and upper flux values obtained from this dataset and the reference values are computed for each reaction as its “diff value”. More formally, for a reaction R, let R_{min} and R_{max} be R’s lower and upper boundary flux values, respectively, computed for an individual; R_{Hmin} and R_{Hmax} be R’s average lower and upper boundary flux values, respectively, computed for healthy subjects (in the training/reference dataset). First, each flux interval is split into two subintervals, one for the positive side with boundary values R^+_{min} and R^+_{max}, R^+_{Hmin} and R^+_{Hmax}, and the other for the negative side with boundary values R^-_{min} and R^-_{max}, R^-_{Hmin} and R^-_{Hmax} for both patient and healthy subjects, respectively. For a flux interval, $[R_{min}, R_{max}]$, the boundaries for the negative and positive subintervals are computed as follows:

$$R^-_{min} = \begin{cases} -R_{max}, & if\ R_{max} < 0 \\ 0 & \text{otherwise} \end{cases}$$

$$R^-_{max} = \begin{cases} -R_{min}, & if\ R_{min} < 0 \\ 0, & \text{otherwise} \end{cases}$$

$$R^+_{min} = \begin{cases} R_{min}, & if\, R_{min} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$R^+_{max} = \begin{cases} R_{max}, & if\ R_{max} > 0 \\ 0, & \text{otherwise} \end{cases},$$

R^-_{Hmin}, R^-_{Hmax}, R^+_{Hmin}, and R^+_{Hmax} are computed similarly. Note that the boundaries of the subintervals are non-negative. Then, the diff scores are computed for positive and negative subintervals, R^+_{diff} and R^-_{diff}, respectively, as follows:

$$R^+_{diff} = \left(R^+_{min} - R^+_{Hmin}\right) + \left(R^+_{max} - R^+_{Hmax}\right) \tag{4}$$

$$R^-_{diff} = \left(R^-_{min} - R^-_{Hmin}\right) + \left(R^-_{max} - R^-_{Hmax}\right). \tag{5}$$

Finally, the diff value for reaction R, denoted as R_{diff}, is computed as the average of the diff values for negative and positive subintervals:

$$R_{diff} = \frac{R^-_{diff} + R^+_{diff}}{2}. \tag{6}$$

Eqs. (4), (5), and (6) may be rewritten in combination as follows:

$$R_{diff} = \frac{(|R_{min}| + |R_{max}|) - (|R_{Hmin}| + |R_{Hmax}|)}{2}. \tag{7}$$

A similar alternative approach to splitting reaction flux intervals is to convert all reversible reactions into two irreversible reactions (one in forward direction and the other in backward direction), and then, calculate diff scores based on only the positive intervals of newly introduced reactions (i.e., R^+_{min}, R^+_{max}, R^+_{Hmin}, and R^+_{Hmax}). In this new setting, the negative and positive subintervals of an originally reversible reaction will be represented by the positive subintervals of the newly added irreversible reactions. There is no difference regarding the irreversible reactions between these two alternative approaches.

A diff score represents the change in the range of possible flux values for a reaction. That is, a positive diff score indicates that the actual activity level of a reaction would be drawn from a range that includes mostly larger values, while a negative score indicates that the actual activity level would be drawn from a range that includes mostly smaller values. Hence, in positive diff cases, the likelihood of having larger flux values is higher compared to healthy individuals, while in negative diff cases, the likelihood of having smaller flux values is higher for the corresponding reactions/pathways. The magnitude of a diff score represents the expectation of how much higher (for positive scores) or lower (for negative scores) the flux magnitude of a reaction would be in a patient compared to the average healthy reference values. For simplicity, we assume that flux values within each flux interval are uniformly distributed. As an example, for a reaction R, let [2], [4] be the reference flux interval obtained from healthy individuals, and [3], [5] be the flux interval computed for a patient. R may have the following possible flux value combinations in a healthy individual and the patient (only integer flux values are considered in the example for brevity, even though real flux values are also possible in practice): [(2, 3), (2, 4), (2, 5), (3, 3), (3, 4), (3, 5), (4, 3), (4, 4), (4, 5)]. Then, the set of corresponding flux differences between the patient and a healthy individual would be [1, 2, 3, 0, 1, 2, -1, 0, 1]. The average (i.e., expectation) of these differences would be 1, which is also the diff score for R computed using Eq. (7) (i.e., $[(3-2)+(5-4)]/2 = 1$). Fig. 3 illustrates several example scenarios and the associated diff scores. In each case, the left-hand side shows the original flux intervals as computed by FVA, while the right-hand side shows the corresponding subintervals and the associated diff values. Fig. 3 presents examples of positive diff scores. The symmetric cases (where patient and reference-healthy labels are switched) would provide examples of negative diff scores.

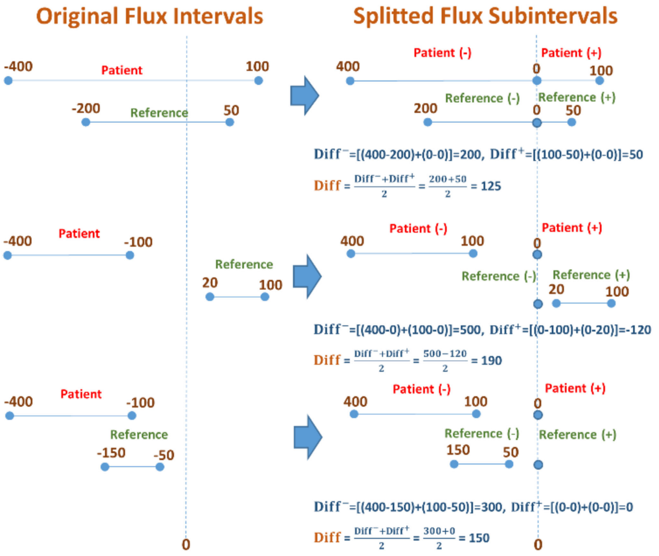
Original Flux Intervals
Splitted Flux Subintervals
-400
Patient
100
-200
Reference
50
Patient (-)
Patient (+)
400
0
100
Reference (-)
Reference (+)
200
0
50
$\text{Diff}^-=[(400\text{-}200)+(0\text{-}0)]=200$, $\text{Diff}^+=[(100\text{-}50)+(0\text{-}0)]=50$
$\text{Diff} = \frac{\text{Diff}^- + \text{Diff}^+}{2} = \frac{200+50}{2} = 125$
-400
Patient
-100
Reference
20
100
Patient (-)
Patient (+)
400
100
0
Reference (-)
Reference (+)
20
100
$\text{Diff}^-=[(400\text{-}0)+(100\text{-}0)]=500$, $\text{Diff}^+=[(0\text{-}100)+(0\text{-}20)]=\text{-}120$
$\text{Diff} = \frac{\text{Diff}^- + \text{Diff}^+}{2} = \frac{500-120}{2} = 190$
-400
Patient
-100
Reference
-150
-50
Patient (-)
Patient (+)
400
100
0
Reference (-)
Reference (+)
150
50
$\text{Diff}^-=[(400\text{-}150)+(100\text{-}50)]=300$, $\text{Diff}^+=[(0\text{-}0)+(0\text{-}0)]=0$
$\text{Diff} = \frac{\text{Diff}^- + \text{Diff}^+}{2} = \frac{300+0}{2} = 150$
0
0

Fig. 3. Various diff score examples.

Next, for each pathway, a diff value is computed as the mean diff value of its reactions. This is similar to pathway-level aggregation methods [62] in the literature. As an alternative, the mean of the top-k reactions with the highest ANOVA score is also considered.

2.5 Statistical Significance Analysis

It may be misleading to directly interpret a diff value solely based on its magnitude. In this study, a statistical significance analysis is carried out to evaluate the possibility of random occurrence of these scores (i.e., *null hypothesis*). To this end, ANOVA [9] is used to calculate the F- and p-values (corrected for multiple hypothesis testing using Benjamini-Hochberg), which indicate the statistical significance of pathway activity differences between patient and healthy groups. ANOVA is preferred because it is widely used in bioinformatics studies [61]. This way, essential pathway changes that characterize a disease may be statistically determined.

2.6 Machine Learning-Based Classification

Computed diff scores for healthy individuals and patients are used as training data to build machine learning-based classification models. These trained models are then employed to predict whether a new/unseen metabolomics data belongs to a patient. In this context, a metabolomics analysis result should be represented as a numerical vector. To this end, one needs to first determine the structure and content of this numeric vector.

Feature Extraction. Since our approach in this study is pathway-based, pathway diff values are used as features (i.e., entries in the numeric vector representation). There are 100 different pathways in the current version of the Recon2 data. Hence, in this study, the size of each vector representing an individual is 100.

Feature Selection. The employed reaction diff score features are correlated among themselves. This is because the steady-state constraint of linear programming leads to correlated min-max values among neighboring reactions. This multicollinearity may negatively affect the performance of the trained linear classification models [16]. Hence, as a preprocessing step, the highly correlated features are eliminated from the vector representation before building a classification model. We consider recursive feature elimination [22], feature importance based on feature weights, and ANOVA.

Classification. A binary classification model for a disease is constructed by using the vector representations of the metabolomics analysis results from healthy individuals and patients as training data. In this work, we consider the most commonly used algorithms, e.g., Support Vector Machines [22], Decision Trees [29], Logistic Regression [6], etc.

3 Experimental Evaluation

In this section, we present our results from an experimental evaluation of Metabolitics on a breast cancer dataset [26] that includes 160 metabolite measurements collected from 214 individuals' plasma samples (76 healthy individuals and 138 breast cancer patients). Metabolitics is also applied to Crohn's disease and colorectal cancer for further evaluation. First, an evaluation of the breast cancer dataset is presented in detail, and then, the results on Crohn's disease and colorectal cancer datasets are discussed. The implementation of the presented algorithms is done in Python. ScikitLearn [39] library is used for machine learning models and statistical significance analysis. Cameo [14] is used for the FVA.

Cross-Validation Strategy. In all experiments, repeated stratified K-fold (K = 10 and repeat count = 10) cross-validation is employed. The pipeline is re-run end-to-end for each iteration of the cross-validation. That is, in each fold, (i) the diff values are re-computed for all individuals in the train split based on the average fluxes computed for healthy people in the train split, (ii) a new classification model is built along with feature selection based on the current train split, (iii) diff values are recomputed for all individuals in the test split based on the average fluxes computed for healthy people in the train split, and (iv) the built model is tested on the corresponding test split. Then, the above steps are repeated. In each repeat, the data is re-shuffled, and test and train splits among healthy individuals and patients are re-determined in a stratified manner. The reported results are the average and standard deviation of the 100 folds (K = 10, repeats = 10). In each iteration, 8 folds are used for training, 1 fold is used for parameter tuning, and 1 fold is used for testing.

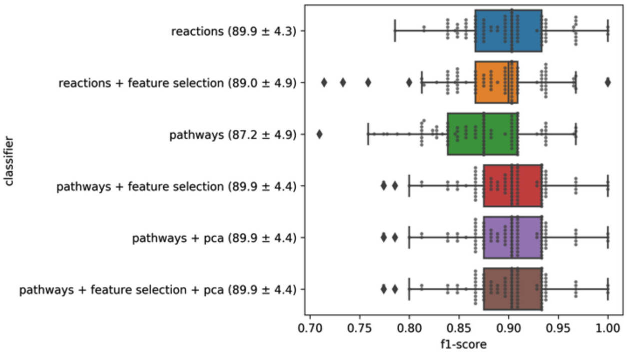

Fig. 4. Repeated stratified K-fold (10 fold and 10 repeat) cross-validation on breast cancer dataset (Feature selection is ANOVA-based).

3.1 Evaluation of Pipeline Steps of Metabolitics

Alternative approaches for each step of the pipeline is benchmarked to justify our choice in each step. The results are reported in Fig. 4. Higher classification performance is achieved using reaction-level diff-scores. However, reaction-level information alone is not interpretable due to a large number of reactions (about 7500 of them). Thus, we aim to obtain pathway-level scores that provide a similar classification performance. Averaging all reaction scores to obtain pathway scores slightly decreases the classification performance because few numbers of reactions for each pathway are significantly altered and probably have important signals, unlike the rest of the reactions whose change is insignificant. Thus, as an alternative approach, the top-k significant reactions are chosen based on their ANOVA score, and then the average pathway diff score is computed over the top-k reactions. The value of k is computed automatically as part of the parameter tuning fold. The tuned value for k is 100 in our reported experiments. Reaction-level feature selection slightly decreases the performance (by 0.9 percent). However, averaging at the pathway-level restrores the performance of classification back to the same level (i.e., 89.9 percent). Feature selection at the pathway-level contributes to pathway-level performance (by 2.2 percent), as it better represents pathway alteration characteristics. Recursive feature elimination (89.8 ± 4.8) [22] and feature importance based on feature weights (89.4 ± 4.5) in logistic regression are considered as alternatives to ANOVA. However, none of them provides better performance; thus, ANOVA is employed in feature selection step. Moreover, we also explore whether dimensionality reduction techniques, such as PCA, would improve the classification performance. Based on Fig. 4, there is no clear benefit that PCA provides. Moreover, alternative dimensionality reduction methods such as factor analysis (82.4 ± 1) [65] and truncated SVD (89.7 ± 4.4) [66] are considered, but they do not provide any better performance than PCA. Therefore, dimensionality reduction is not employed in this study. PCA results are still kept in Fig. 4 to illustrate the results of the exploratory analysis.

Classifiers. Several alternative classifiers are considered. Average and standard dev. of f1-score (i.e., 2^{*}precision*recall/(precision+recall)) for each classifier are presented in Fig. 5 (calculated over 10-folds that are repeated 10 times). We tune the hyper-parameters of each model, such as regularization parameter and different regularization

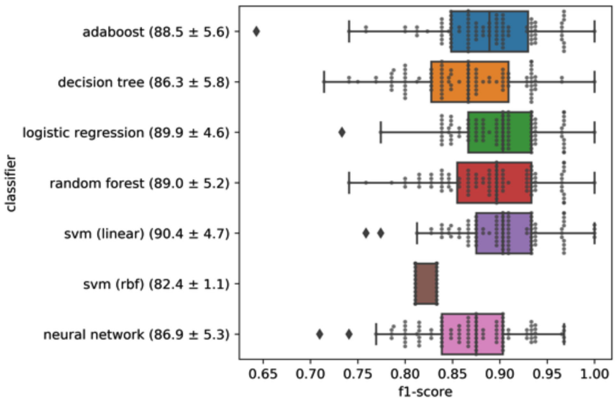

Fig. 5. f1-scores for different classification algorithms.

metrics (L1 and L2) for logistic regression and SVM, gini and entropy as criterion metric and different tree depths for decision tree and random-forest, different numbers of estimators for ensemble methods (adaboost and random-forest), etc. Also, we include a neural network model with one hidden layer, relu activation, and adam optimizer. Logistic regression (LR) and linear-svm (SVM) beat other methods, and they perform almost equally well (with f1 score ~90 percent). Hence, one could choose either SVM or LR for the dataset considered in this study. In our experiments, LR is employed (for no particular reason over SVM).

3.2 Metabolitics Captures Biologically Relevant Information in Breast Cancer

Table 1 lists the top-10 significantly changing pathways in breast cancer patients. In the Alanine and Aspartate Metabolism pathway, the reaction that experiences most flux capacity increase is *asparagine synthase* (encoded by *ASNS*). This reaction converts *glutamine* into *glutamate* and *asparagine*. *Glutamate* is a precursor that leads to the biosynthesis of other amino acids which are needed by breast cancer tumor cells to proliferate [57].

The association of Arginine and Proline Metabolism with breast cancer is reported for the first time in this study. De Ingeniis *et al.* [10] suggested that *NADP+*, which is produced during the synthesis of *proline* from *arginine*, may be directed to *PP-ribose-P* synthesis. *PP-ribose-P* is later metabolized in nucleotide synthesis. Breast cancer tumors may also use the same mechanism to enhance nucleotide synthesis.

Breast cancer cells often experience high oxidative stress [4]. *Hypotaurine* has a significant effect on the reactive oxidative stress states of cells [5], [19].

Fatty acids are used for lipid synthesis, and lipids form the main structure of the cell membrane. In order to form new tumor cells, lipids have a vital role as a building block of cell membranes [2]. The first committed step of fatty acid synthesis is the carboxylation of acetyl-CoA to malonyl-CoA. This reaction is heavily dependent on Biotin to take place [51]. Hence, the increase in flux capacity of Biotin Metabolism is another indicator of increased fatty acid synthesis.

As related to the above phenomena, the CoA metabolism (both synthesis and catabolism) exhibits a positive flux capacity change. Acetate is an important input for lipid synthesis in breast cancer tumors [46]. More specifically, Acetyl CoA is first produced from Acetate (ACOT12) and then converted to Malonyl CoA (ACC-alpha) [47]. Malonyl CoA is the primary input metabolite to the first steps of fatty acid synthesis.

TABLE 1
Significantly Changing Pathways in Breast Cancer

Pathway	F-val	p-val	diff	Pathifier	Paradigm	Met Cnt
Alanine and aspartate metabolism	120	4.5e-21	550	0.16	0.0012	7
Arginine and proline metabolism	100	1.5e-18	305	0.46	-0.0190	8
Taurine and hypotaurine metabolism	99	1.8e-18	475	-	0.1800	1
CoA catabolism	79	1.4e-15	285	-	-	0
Nucleotide interconversion	76	4.4e-15	-235	-	-	1
Biotin metabolism	73	1.0e-14	480	-	0.0570	1
Glycolysis/gluconeogenesis	66	1.4e-13	-310	0.11	-0.0190	5
Eicosanoid metabolism	62	4.9e-13	-340	0.50	-	4
CoA synthesis	56	4.6e-12	225	-	-	0
Aminosugar metabolism	52	1.7e-11	-280	-	-0.0011	1

The statistical significance values (i.e., F-val and p-val) are computed through ANOVA analysis based on computed diff values. The last column reports the measured metabolite counts for the reported pathways. Columns 5-6 provide scores for Pathifier and Paradigm.

NTPs and *dNTPs* form the main input metabolites of nucleotide synthesis. Nucleotide interconversion performs the transformations *(d) NMP* ↔ *(d) NDP* ↔ *(d) NTP* by transferring phosphate to produce the critical metabolites necessary for nucleotide metabolism which is essential for breast cancer tumor cells to proliferate [48].

N-acetylgalactosamine and *N-acetylglucosamine* of Aminosugar metabolism are associated with the invasiveness capability of breast cancer cells [84].

Eicosanoids are divided into 3 different subgroups: *prostanoids* (*COX*), *lipoxygenases* (*LOX*), and *ω-hydroxylases/epoxygenases*. *COX* and *LOX* eicosanoids have a supporting role in breast cancer development and metastasis [55]. *COXs* are more active in ER-positive breast cancer tumors, while *LOX's* are more effective in ER-negative tumors. In the ER-positive breast cancer subtype, tumor cells have estrogen-sensitive receptors. Such tumors bind to the estrogen hormone, and estrogen accelerates the development and spread of these cells. COXs boost estrogen synthesis by increasing the activity of the rate-limiting enzyme in estrogen synthesis, *aromatase* [55].

Finally, Glycolysis is associated with what is known as the Warburg effect [43]. That is, tumor cells consume and convert glucose into lactate even though plenty of oxygen is available and mitochondrial oxidation is possible.

3.3 Comparison With the State of the Art

We compare Metabolitics with Pathifier [13] and Paradigm [54] in terms of (i) accuracy (AUC-ROC), (ii) the coverage of metabolism, and (iii) robustness to data loss.

Accuracy. Fig. 6 presents AUC (Area under the Curve) ROC (Receiver Operating Characteristics) curve that is obtained with repeated stratified K-fold (10 folds and 10 repeats) cross validation. Metabolitics has better AUC value than both Paradigm and Pathifier. The default parameters

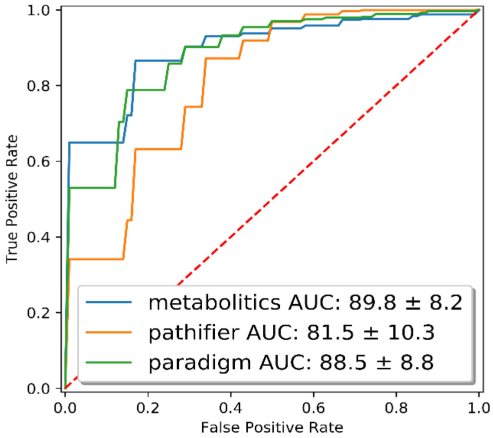

Fig. 6. AUC – ROC curve for metabolitics, pathifier, and paradigm.

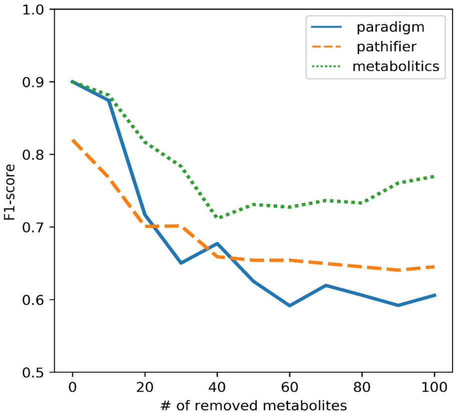

Fig. 7. Classification performance (average f1-score) of metabolitics, paradigm, and pathifier with less measurements.

of Pathifier and Paradigm are used, since they provide the best performance. Pathifier, in its original implementation, employs the labels of the entire dataset to learn and extract pathway features. The original implementation is adapted to the k-fold cross-validation setting so that in each fold, it uses only the training split for that fold to fit its principal curve, and later, the same curve is used to compute the pathway dysregulation scores for the test split. The same train-test splits as Metabolitics are used in a 10-fold cross-validation setting (repeated 10 times). Likewise, Paradigm learns its model parameters based on the entire dataset. However, a similar adaptation for Paradigm was not possible, as it is not open-source, and we had to use the binary executable of Paradigm to perform the comparison experiments in this study. Nevertheless, Paradigm's AUC values are still included in Fig. 6 to show that Metabolitics outperforms Paradigm, even though Paradigm uses the whole dataset to learn its parameters.

Coverage of the Metabolic Network. In this section, we demonstrate that Metabolitics provides higher coverage of the metabolism than Paradigm and Pathifier. The coverage of the metabolic network is important to discover the complete set of metabolic changes caused by a disease. According to Table 1, 6 of the 10 significant pathways reported by Metabolitics cannot be detected by Pathifier (i.e., no dysregulation score is computed). Similarly, Paradigm misses 4 of the top-10 pathways. The reason for this drawback of the Pathifier and Paradigm is explained by the last column of Table 1. That is, Pathifier and Paradigm are capable of evaluating and scoring only pathways with at least several metabolites included in the analyzed metabolomics data set. Therefore, they may miss several important key pathways for which no measured metabolite is available in the analyzed metabolomics dataset. On the other hand, since Metabolitics considers the metabolic network as a whole, and the pathways are interconnected in this network, it can produce analysis results for pathways for which no or only a few measurements are available.

Robustness to Data Loss. In this section, the robustness of Metabolitics to that of the Pathifier and Paradigm is compared. We informally define "robustness" as the resistance of an algorithm's performance to keep its initial value in case of a decrease in the number of measured metabolites. In other words, more robust algorithms experience lower performance loss, when the number of measured metabolites decreases. In order to simulate the worst-case scenario, first, ANOVA is performed to identify the most significantly differing metabolites between healthy individuals and patients. Next, the metabolites are sorted by their F-values (as computed by ANOVA). Then, in an iterative way, the top-10 metabolites with the highest F-values are removed from the metabolomics dataset. Next, the classification models are rebuilt in each iteration on the reduced dataset. Fig. 7 shows the average f1-score of stratified K-fold ($K = 10$) cross-validation for Metabolitics, Pathifier, and Paradigm, after each metabolite removal iteration. As shown, Metabolitics is more robust to decrease in measured data than Pathifier and Paradigm, since it provides insights on parts of the metabolic network with no measurements.

3.4 Pathway Significance-Based Prediction

In this section, as an alternative method, we compute pathway-wise significance that considers all positive and negative diffs in a pathway to determine whether the pathway is indeed up- or down-regulated (similar to gene set enrichment analysis, e.g., [60]). More specifically, the averages of healthy and patient samples are computed separately. As a result, for each pathway, two reaction vectors are obtained, one for the healthy group and one for cancer patients. Then, Fisher exact test is performed for each pathway based on these reaction vectors. The contingency matrix has "healthy" and "patient" as columns, and the numbers of increasing, decreasing, and unchanged reactions as rows, respectively. The resulting significance values are employed as features in the classification task. The f1-score of this classification is 80 percent which is lower than Metabolitics' performance.

3.5 Stability Analysis

In this section, the stability of the proposed linear programming objective function is evaluated under systematically introduced noise in the input data. To this end, the effect of noise in metabolomics data on the f1-score is measured on a real dataset, i.e., the breast cancer dataset. More specifically, some uniform noise proportional to metabolite fold-changes is introduced into the original breast cancer dataset with increasing amounts. Then, Metabolitics is run for each case. Fig. 8 shows the change in f1-score as the noise amount

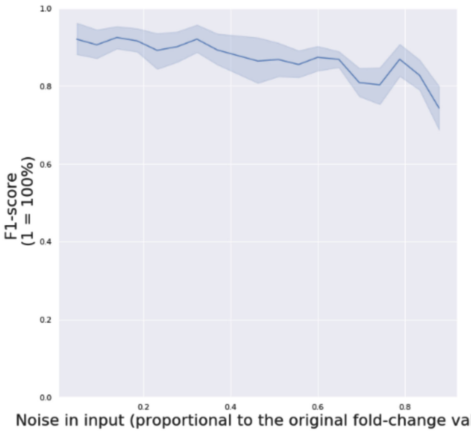

Fig. 8. The effect of noise in input metabolite measurements on the breast cancer diagnosis f1-score.

increases. We conclude that Metabolitics performs reasonably well under increasing noise.

3.6 Applying Metabolitics on Other Diseases

In this section, to further validate the proposed methodology, we demonstrate the application of Metabolitics on two additional disease datasets, namely, Crohn's disease [49] and colorectal cancer [45]. Fig. 9 presents the overall f1-score of Metabolitics in predicting disease status of subjects, separately for each disease.

Crohn's disease (CD) is an autoimmune disorder that causes inflammation in the gastroenterological tract of patients. Metabolitics is run on a recent dataset [49] that is obtained from the serum measurements of 40 subjects (20 with CD and 20 healthy). With the Metabolitics assigned pathway diff scores as features, a logistic regression-based classifier (with C = 0.3e-6) discriminates between subjects with CD and healthy ones with a mean f1-score of 77 percent.

Colorectal cancer (CRC) originates from the epithelial cells of the colon or rectum, and it is one of the most common cancers worldwide. Metabolitics is run on a metabolomics dataset [45] that are obtained from visceral fat tissues of 55 subjects (49 CRC patients, 6 healthy controls). On this dataset, logistic regression achieves a 98 percent f1-score.

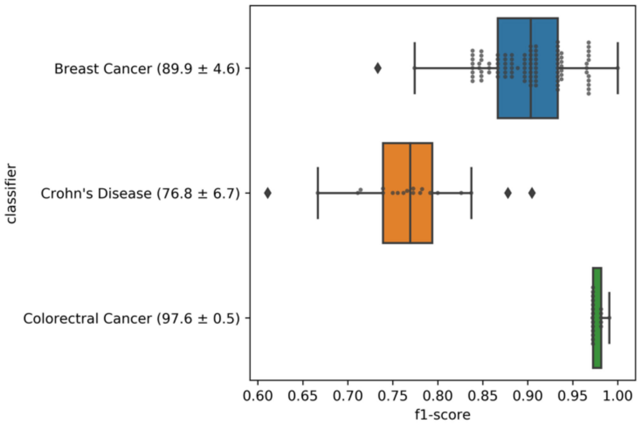

Fig. 9. Diagnosis f1-score of Metabolitics in different diseases.

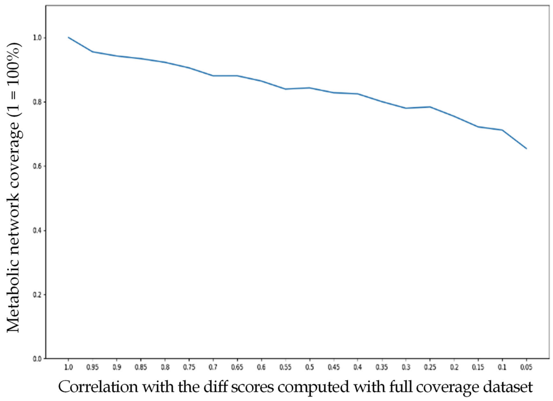

Fig. 10. Simulation study results with the breast cancer dataset employed as a seed.

3.7 Sparsity Robustness Analysis

In this section, a simulation study is performed to analyze the effect of sparsity in terms of the coverage of metabolomics measurements of the metabolic network on the stability of Metabolitics analysis results. To this end, a new metabolite fold-change dataset is generated by using the original breast cancer data set as a seed, and expanding it into a full-coverage metabolomics data set as follows: For each unmeasured metabolite m, within the seed metabolomics data, we identify the metabolite that is most similar to m in terms of the reactions that it participates. That is, given two metabolites, the number of common reactions that they both participate in is employed as a measure of their similarity. Suppose that the most similar metabolite to m is x. Then, the fold change of m is assigned as that of x plus a certain noise which is generated randomly under a normal distribution with a mean of 0 and a standard deviation of 0.1. In order to perform the simulation, the network coverage is systematically decreased by 5 percent by randomly removing metabolites from the initial dataset at each iteration. At each step, the proposed linear programming objective function is run on the current data set, and the correlation between the reaction min-max flux values of the current results and that of the full coverage data-set is computed. The results are presented in Fig. 10. The figure shows that even with very sparse measurements, Metabolitics provides results that are highly correlated with those of the full coverage.

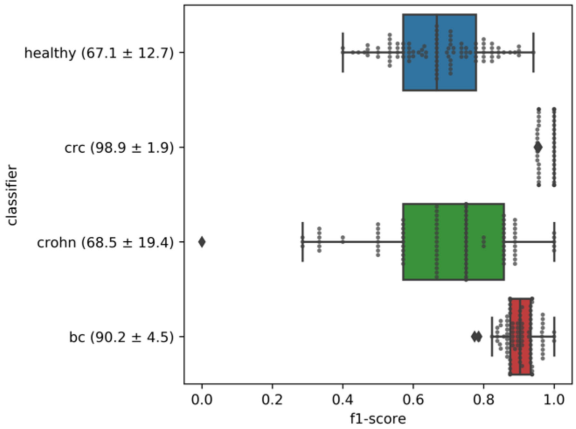

Fig. 11. Multi-class classification results.

3.8 Multi-Class Classification

In this section, we experiment with the multi-class classification of multiple diseases based on the computed diff values. Metabolitics is run independently on each disease dataset to obtain pathway-level diff-scores. Repeated stratified K-fold (10 fold and 10 repeats) cross-validation is employed. Then, a multi-class logistic regression model is trained to predict diseases. The results (Fig. 11) show that the average performance in the multi-class setting is worse than that of binary classification.

4 Conclusion

Metabolic networks in patients often experience dramatic perturbations in comparison to healthy individuals. These perturbations carry important information regarding the root cause of the underlying condition. In this paper, we present the Metabolitics algorithm. Given some metabolomics measurements of an individual and a database containing metabolic network data, Metabolitics computes the metabolic network configuration that may lead to the measured changes in the provided metabolomics data. More specifically, it computes a personalized "diff" score for each pathway that represents the amount of activity change in the pathway of that individual. Our approach is based on (i) dynamically creating a linear programming model of the metabolic network according to the given metabolite measurements, and (ii) performing a customized flux variability analysis on these models to quantify the differentiation of pathway flux capacities in reference to healthy people. After computing the diff scores, classification models that employ the computed diff scores as features are built to differentiate patients from healthy individuals. We extensively evaluated Metabolitics on three different disease datasets, namely, breast cancer, Crohn's disease, colorectal cancer, and show that it (i) can identify patients with around 88 percent f1-score on the average, (ii) provides a biologically relevant metabolic analysis of diseases, (iii) is robust to noise and (iv) enables a larger coverage of the metabolic network in the analysis result.

Acknowledgments

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) Career Grant to Ali Cakmak [grant number: 114E115]. *Availability:* https://github.com/sehir-bioinformatics-database-lab/metabolitics.

References

[1] F. Ameer, L. Scandiuzzi, S. Hasnain, H. Kalbacher, and N. Zaidi, "De novo lipogenesis in health and disease," *Metabolism*, vol. 63, no. 7, pp. 895–902, 2014.

[2] J. Baumann, C. Sevinsky, and D. S. Conklin, "Lipid biology of breast cancer," *Biochimica et Biophysica Acta (BBA)-Mol. Cell Biol. Lipids*, vol. 1831, no. 10, pp. 1509–1517, 2013.

[3] B. K. Tiwary, "Computational medicine: Quantitative modeling of complex diseases," *Briefings Bioinf.*, vol. 21, no. 2, pp. 429–440, 2020.

[4] N. S. Brown and R. Bicknell, "Hypoxia and oxidative stress in breast cancer oxidative stress-its effects on the growth, metastatic potential and response to therapy of breast cancer," *Breast Cancer Res.*, vol. 3, no. 5, pp. 323–327, 2001.

[5] M. N. Bucak *et al.*, "Effects of hypotaurine, cysteamine and amino ac-ids solution on post-thaw microscopic and oxidative stress parameters of angora goat semen," *Res. Veterinary Sci.*, vol. 87, no. 3, pp. 468–472, 2009.

[6] G. C. Cawley and N. L. Talbot, "Gene selection in cancer classification using sparse logistic regression with Bayesian regularization," *Bioinformatics*, vol. 22, no. 19, pp. 2348–2355, 2006.

[7] M. H. Celik, T. Saleh, A. Dokay, and A. Cakmak, "MetabolitcsDB: A database of metabolomics analyses (under review)," *IEEE Trans. Comp. Biol. Bioinf.*, 2020.

[8] P. Rana, C. Berry, P. Ghosh, and S. S. Fong, "Recent advances on constraint-based models by integrating machine learning," *Curr. Opinion Biotechnol.*, vol. 64, pp. 85–91, 2020.

[9] X. Cui and G. A. Churchill, "Statistical tests for differential expression in cDNA microarray experiments," *Genome Biol.*, vol. 4, no. 4, 2003, Art. no. 210.

[10] J. De Ingeniis, M. D. Kazanov, K. Shatalin, M. S. Gelfand, A. L. Osterman, and L. Sorci, "Glutamine versus ammonia utilization in the NAD synthetase family," *PLoS One*, vol. 7, no. 6, 2012, Art. no. e39115.

[11] J. C. Dodge *et al.*, "Metabolic signatures of amyotrophic lateral sclerosis reveal insights into disease pathogenesis," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 26, pp. 10812–10817, 2013.

[12] S. Draghici *et al.*, "A systems biology approach for pathway level analysis," *Genome Res.*, vol. 17, no. 10, pp. 1537–1545, 2007.

[13] Y. Drier, M. Sheffer, and E. Domany, "Pathway-based personalized analysis of cancer," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 16, pp. 6388–6393, 2013.

[14] A. Fernández-Castané, T. Fehér, P. Carbonell, C. Pauthenier, and J. L. Faulon, "Computer-aided design for metabolic engineering," *J. Biotechnol.*, vol. 192, pp. 302–313, 2014.

[15] Z. Costello and H. G. Martin, "A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data," *NPJ Syst. Biol. Appl.*, vol. 4, no. 1, pp. 1–14., 2018.

[16] A. Alin, "Multicollinearity," *Wiley Interdisciplinary Rev.: Comput. Stat.*, vol. 2, no. 3, pp. 370–374, 2010.

[17] D. Toubiana *et al.*, "Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data," *Commun. Biol.*, vol. 2, no. 1, pp. 1–13, 2019.

[18] B. Ghesquiere, B. W. Wong, A. Kuchnio, and P. Carmeliet, "Metabolism of stromal and immune cells in health and disease," *Nature*, vol. 511, no. 7508, pp. 167–176, 2014.

[19] D. Gossai and C. A. Lau-Cam, "The effects of taurine, taurine homologs and hypotaurine on cell and membrane antioxidative system alterations caused by type 2 diabetes in rat erythrocytes," *Adv. Exp. Med. Biol.*, vol. 7, pp. 359–368, 2009.

[20] F. Guillaumond *et al.*, "Cholesterol uptake disruption, in association with chemotherapy, is a promising combined metabolic therapy for pancreatic adenocarcinoma," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 8, pp. 2473–2478, 2015.

[21] E. Yaneske and C. Angione, "The polyomics of ageing through individual-based metabolic modelling," *BMC Bioinf.*, vol. 19, no. 14, pp. 83–96, 2018.

[22] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1, pp. 389–422, 2002.

[23] J. Hastings *et al.*, "ChEBI in 2016: Improved services and an expanding collection of metabolites," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1214–D1219, 2016.

[24] L. M. Heiser *et al.*, "Subtype and pathway-specific responses to anticancer compounds in breast cancer," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 8, pp. 2724–2729, 2012.

[25] W. Guo, Y. Xu, and X. Feng, "DeepMetabolism: A deep learning system to predict phenotype from genome sequencing," *AIChE Annu. Meeting*, AIChE, 2017.

[26] S. Huang, N. Chong, N. E. Lewis, W. Jia, G. Xie, and L. X. Garmire, "Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis," *Genome Med.*, vol. 8, no. 1, 2016, Art. no. 34.

[27] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[28] P. Khatri, S. Sellamuthu, P. Malhotra, K. Amin, A. Done, and S. Draghici, "Recent additions and improvements to the onto-tools," *Nucleic Acids Res.*, vol. 33, no. suppl 2, pp. W762–W765, 2005.

[29] C. Kingsford and S. L. Salzberg, "What are decision trees?," *Nature Biotechnol.*, vol. 26, no. 9, pp. 1011–1013, 2008.

[30] S. Klamt *et al.*, "From elementary flux modes to elementary flux vectors: Metabolic pathway analysis with arbitrary linear flux constraints," *PLoS Comput. Biol.*, vol. 13, no. 4, Art. no. e1005409, 2017.

[31] P. Labhsetwar, J. A. Cole, E. Roberts, N. D. Price, and Z. A. Luthey-Schulten, "Heterogeneity in protein expression induces metabolic variability in a modeled Escherichia coli population," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 34, pp. 14006–14011, 2013.

[32] M. Barsacchi, H. Andres-Terre, and P. Lió, "GEESE: Metabolically driven latent space learning for gene expression data," 2018, bioRxiv, 365643

[33] E. Lee, H. Y. Chuang, J. W. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification," *PLoS Comput. Biol.*, vol. 4, no. 11, 2008, Art. no. e1000217.

[34] M. MacGillivray, A. Ko, E. Gruber, M. Sawyer, E. Almaas, and A. Holder, "Robust analysis of fluxes in genome-scale metabolic pathways," *Sci. Rep.*, vol. 7, 2017, Art. no. 268.

[35] M. L. Mo, B. Ø. Palsson, and M. J. Herrgård, "Connecting extracellular metabolomic measurements to intracellular flux states in yeast," *BMC Syst. Biol.*, vol. 3, no. 1, 2009, Art. no. 37.

[36] A. C. Müller and A. Bockmayr, "Fast thermodynamically constrained flux variability analysis," *Bioinformatics*, vol. 29, no. 7, pp. 903–909, 2013.

[37] S. G. Wu *et al.*, "Rapid prediction of bacterial heterotrophic fluxomics using machine learning and constraint programming," *PLoS Comput. Biol.*, vol. 12, no. 4, 2016, Art. no. e1004838.

[38] J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?," *Nature Biotechnol.*, vol. 28, no. 3, pp. 245–248, 2010.

[39] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[40] D. Heckmann *et al.*, "Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models," *Nature Commun*, vol. 9, no. 1, pp. 1–10, 2018.

[41] X. Qi, Z. M. Ozsoyoglu, and G. Ozsoyoglu, "Matching metabolites and reactions in different metabolic networks," *Methods*, vol. 69, no. 3, pp. 282–297, 2014.

[42] A. Raj, E. LoCastro, A. Kuceyeski, D. Tosun, N. Relkin, M. Weiner, and Alzheimer's Disease Neuroimaging Initiative (ADNI), "Network dif-fusion model of progression predicts longitudinal patterns of atrophy and metabolism in Alzheimer's disease," *Cell Rep.*, vol. 10, no. 3, pp. 359–369, 2015.

[43] I. F. Robey, R. M. Stephen, K. S. Brown, B. K. Baggett, R. A. Gatenby, and R. J. Gillies, "Regulation of the warburg effect in early- passage breast cancer cells," *Neoplasia*, vol. 10, no. 8, pp. 745IN1–756, 2008.

[44] R. Rueedi *et al.*, "Genome-wide association study of metabolic traits reveals novel gene-metabolite-disease links," *PLoS Genetics*, vol. 10, no. 2, 2014, Art no. e1004132.

[45] I. Schlecht *et al.*, "Visceral adipose tissue but not subcutaneous adipose tissue is associated with urine and serum metabolites," *PLoS One*, vol. 12, no. 4, pp. e0175133, 2017, Art no. e0175133.

[46] Z. T. Schug *et al.*, "Acetyl-CoA synthetase 2 promotes acetate utilization and maintains cancer cell growth under metabolic stress," *Cancer Cell*, vol. 27, no. 1, pp. 57–71, 2015.

[47] Z. T. Schug, J. V. Voorde, and E. Gottlieb, "The metabolic fate of acetate in cancer," *Nature Rev. Cancer*, vol. 16, pp. 708–717, 2016.

[48] M. Schwab (Ed.) in *Encyclopedia of Cancer*. Berlin, Germany: Springer, 2008.

[49] E. A. Scoville *et al.*, "Alterations in lipid, amino acid, and energy metabolism distinguish Crohn's disease from ulcerative colitis and control subjects by serum metabolomic profiling," *Metabolomics*, vol. 14, no. 1, 2018, Art. no. 17.

[50] A. Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Nat. Acad. Sci. USA.*, vol. 102, no. 43, pp. 15545–15550, 2005.

[51] C. O. Rock, S. Jackowski, and J. E. Cronan, "Lipid metabolism in prokaryotes," *New Comprehensive Biochem.*, vol. 31, pp. 35–74, 1996.

[52] A. L. Tarca *et al.*, "A novel signaling pathway impact analysis," *Bioinformatics*, vol. 25, no. 1, pp. 75–82, 2009.

[53] I. Thiele *et al.*, "A community-driven global reconstruction of human metabolism," *Nature Biotechnol.*, vol. 31, pp. 419–425 (2013).

[54] C. J. Vaske *et al.*, "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM," *Bioinformatics*, vol. 26, no. 12, pp. i237–i245, 2010.

[55] L. Vona-Davis and D. P. Rose, "The obesity-inflammation-eicosanoid axis in breast cancer," *J. Mammary Gland Biol. Neoplasia*, vol. 18, no. 3-4, pp. 291–307, 2013.

[56] J. Wang, D. Duncan, Z. Shi, and B. Zhang, "WEB-based gene set analysis toolkit (WebGestalt): Update 2013," *Nucleic Acids Res.*, vol. 41, no. W1, pp. W77–W83, 2013.

[57] D. R. Wise and C. B. Thompson, "Glutamine addiction: A new therapeutic target in cancer," *Trends Biochem. Sci.*, vol. 35, no. 8, pp. 427–433, 2010.

[58] D. S. Wishart, R. Mandal, A. Stanislaus, and M. Ramirez-Gaona, "Cancer metabolomics and the human metabolome database," *Metabolites*, vol. 6, no. 1, 2016, Art. no. 10.

[59] B. R. Zeeberg *et al.*, "GoMiner: A resource for biological interpretation of genomic and proteomic data," *Genome Biol.*, vol. 4, no. 4, 2003, Art. no. R28.

[60] L. Väremo, J. Nielsen, and I. Nookaew, "Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods," *Nucleic Acids Res.*, vol. 41, no. 8, pp. 4378–4391, 2013.

[61] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[62] S. Hwang, "Comparison and evaluation of pathway-level aggregation methods of gene expression data," *BMC Genomics*, vol. 13, no. 7, 2012, Art. no. S26.

[63] S. Gudmundsson and I. Thiele, "Computationally efficient flux variability analysis," *BMC Bioinf.*, vol. 11, no. 1, 2010, Art. no. 489.

[64] P. Gerlee, L. Lizana, and K. Sneppen, "Pathway identification by network pruning in the metabolic network of Escherichia coli," *Bioinformatics*, vol. 25, no. 24, pp. 3282–3288, 2009.

[65] Z. Ghahramani, and G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," Tech.Report, University of Toronto, 1996.

[66] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions," 2009, *arXiv:0909.4061*.

[67] H. W. Koh, D. Fermin, C. Vogel, K. P. Choi, R. M. Ewing, and H. Choi, "iOmicsPASS: A novel method for integration of multi-omics data over biological networks and discovery of predictive subnetworks," *NPJ Syst. Biol. Appl. 5*, no. 1, pp. 1–10, 2019.

[68] P. Mertins *et al.*, "Proteogenomics connects somatic mutations to signaling in breast cancer," *Nature*, vol. 534, no. 7605, pp. 55–62, 2016.

[69] J. Chong *et al.*, "MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W486–W494, 2018.

[70] M. Cuperlovic-Culf, "Machine learning methods for analysis of metabolic data and metabolic pathway modeling," *Metabolites*, vol. 8, no. 1, 2018, Art. no. 4.

[71] A. Noronha *et al.*, "The virtual metabolic human database: Integrating human and gut microbiome metabolism with nutrition and disease," *Nucleic Acids Res.*, vol. 47, pp. D614–D624, 2019.

[72] N. S. Kale *et al.*, "MetaboLights: An open-access database repository for metabolomics data," *Curr. Protocols Bioinf.*, vol. 53, no. 1, pp. 14–13, 2016.

[73] J. A. Vizcaíno *et al.*, "2016 update of the PRIDE database and its related tools," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D447–D456, 2016.

[74] A. Fabregat *et al.*, "The reactome pathway knowledgebase," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D649–D655, 2017.

[75] D. N. Slenter *et al.*, "WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research," *Nucl. Acids Res.*, vol. 46, no. D1, pp. D661–D667, 2017.

[76] C. I. Cruickshank-Quinn *et al.*, "Metabolomics and transcriptomics pathway approach reveals outcome-specific perturbations in COPD," *Sci. Rep.*, vol. 8, no. 1, 2018, Art. no. 17132.

[77] D. G. Brown *et al.*, "Metabolomics and metabolic pathway networks from human colorectal cancers, adjacent mucosa, and stool," *Cancer Metab.*, vol. 4, no. 1, 2016, Art. no. 11.

[78] E. M. Forsberg *et al.*, "Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS online," *Nature Protocols*, vol. 13, no. 4, pp. 633–651, 2018.

[79] A. Rasid, A. Aboudakika, and A. Cakmak, "Genobolitics: Extending metabolitics to include gene expression data," Technical Report, Istanbul Sehir University, 2018. [Online]. Available: http://bit.ly/genobolitics

[80] J. S. Cho, C. Gu, T. Han, J. Y. Ryu, and S. Y. Lee, "Reconstruction of context-specific genome-scale metabolic models using multi-omics data to study metabolic rewiring," *Curr. Opinion Syst. Biol.*, vol. 15, pp. 1–11, 2019.

[81] M. Tian and J. L. Reed, "Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis," *Bioinformatics*, vol. 34, no. 22, pp. 3882–3888, 2018.

[82] C. Angione, "Integrating splice-isoform expression into genome-scale models characterizes breast cancer metabolism," *Bioinformatics*, vol. 34, no. 3, pp. 494–501, 2017.

[83] J. Nielsen, "Systems biology of metabolism: A driver for developing personalized and precision medicine," *Cell Metab.*, vol. 25, no. 3, pp. 572–579, 2017.

[84] T. H. More *et al.*, "Metabolomic alterations in invasive ductal carcinoma of breast: A comprehensive metabolomic study using tissue and serum samples," *Oncotarget*, vol. 9, no. 2, pp. 2678–2696, 2018.

[85] G. Zampieri, S. Vijayakumar, E. Yaneske, and C. Angione, "Machine and deep learning meet genome-scale metabolic modeling," *PLoS Comput. Biol.*, vol. 15, no. 7, 2019, Art. no. e1007084.

Ali Cakmak received the BSc degree from the Computer Engineering Department at Bilkent University, in Ankara, Turkey, in 2003, and the PhD degree from the Electrical Engineering and Computer Science Department at Case Western Reserve University, in Cleveland, OH, in 2008. His research interests include bioinformatics, machine learning, data mining, databases, and data science. He is a recipient of the TUBITAK CAREER Grant.

M. Hasan Celik received the BSc degree from the Computer Science and Engineering Department at Istanbul Sehir University, in Istanbul, Turkey, in 2016. He is currently working toward the graduate degree in the Technical University of Munich, Munich, Germany. His research interests are machine learning, bioinformatics, and systems biology.