Journal of proteome research

Article

Cite This: J. Proteome Res. 2018, 17, 337−347

pubs.acs.org/jpr

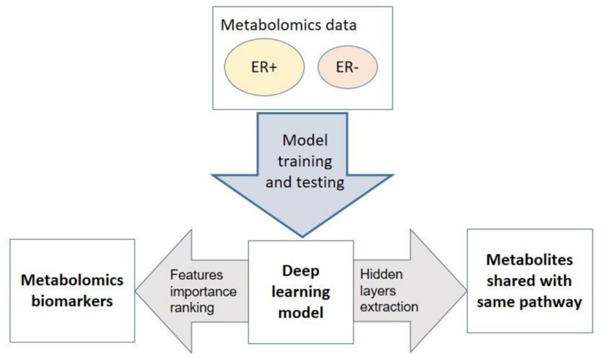
Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data

Fadhl M. Alakwaa,[†] Kumardeep Chaudhary,[†] and Lana X. Garmire*,[†,‡]

[†]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii 96813, United States

[‡]Molecular Biosciences and Bioengineering Graduate Program, University of Hawaii at Manoa, Honolulu, Hawaii 96822, United States

Supporting Information


Metabolomics data
ER+
ER-
Model training and testing
Metabolomics biomarkers
Features importance ranking
Deep learning model
Hidden layers extraction
Metabolites shared with same pathway

ABSTRACT: Metabolomics holds the promise as a new technology to diagnose highly heterogeneous diseases. Conventionally, metabolomics data analysis for diagnosis is done using various statistical and machine learning based classification methods. However, it remains unknown if deep neural network, a class of increasingly popular machine learning methods, is suitable to classify metabolomics data. Here we use a cohort of 271 breast cancer tissues, 204 positive estrogen receptor (ER+), and 67 negative estrogen receptor (ER−) to test the accuracies of feed-forward networks, a deep learning (DL) framework, as well as six widely used machine learning models, namely random forest (RF), support vector machines (SVM), recursive partitioning and regression trees (RPART), linear discriminant analysis (LDA), prediction analysis for microarrays (PAM), and generalized boosted models (GBM). DL framework has the highest area under the curve (AUC) of 0.93 in classifying ER+/ER− patients, compared to the other six machine learning algorithms. Furthermore, the biological interpretation of the first hidden layer reveals eight commonly enriched significant metabolomics pathways (adjusted *P*-value <0.05) that cannot be discovered by other machine learning methods. Among them, protein digestion and absorption and ATP-binding cassette (ABC) transporters pathways are also confirmed in integrated analysis between metabolomics and gene expression data in these samples. In summary, deep learning method shows advantages for metabolomics based breast cancer ER status classification, with both the highest prediction accuracy (AUC = 0.93) and better revelation of disease biology. We encourage the adoption of feed-forward networks based deep learning method in the metabolomics research community for classification.

KEYWORDS: *breast cancer, metabolomics, estrogen receptor, deep learning, bioinformatics*

INTRODUCTION

According to Global Health Estimates (WHO), more than one-half million women died due of breast cancer worldwide.[1] Breast cancer is the second leading cause of cancer-related deaths among women in the United States.[2] On the basis of human epidermal growth factor receptor 2 (Her2), progesteron receptor (PR), and estrogen receptor (ER), breast cancer can be categorized into four molecular subtypes:[3] Luminal A (ER+, PR±, and Her2-), Luminal B (ER+, PR±, and Her2 ±), Her2-enriched (ER−, PR−, and Her2+), and triple negative (ER−, PR−, and Her2−).[4] The survival outcomes differ significantly among these subtypes. Luminal A and B subtypes have a relatively good prognosis; however, triple negative tumors and Her2 tumors have very poor prognosis.[5] Identification of molecular subtypes is crucial in determining cancer prognosis and therapeutic selection. Recently, many studies used metabolomics data to segregate molecular subtypes, given that breast cancer is manifested as a metabolic disease.[6,7] For example, glutamate-to-glutamine ratio and aerobic glycolysis were proposed as biomarkers of ER and Her2 status, respectively.[8,9]

Metabolomics studies are usually done by three major platforms: gas chromatography−mass spectrometry (GC−MS), liquid chromatography (LC−MS), and nuclear magnetic resonance (NMR). The parallel use of these instruments allows detecting more metabolites for the same sample. Coupling with the development in the instrumentations, state-of-the-art data analysis tools are much needed to handle the large amount of metabolite data generated. For problems of metabolomics data classification and regression, machine learning algorithms have been applied.[10] For example, random forest (RF) is a widely used machine learning algorithm based on decision tree theory. It works with high-dimensional data and can deal with unbalanced and missing values in the data.[11] Support vector machine (SVM) is another machine learning algorithm that separates the metabolomics data with N data

Received: August 22, 2017
Published: November 7, 2017

points into (N-1) dimensional hyperplane.[12] SVM was used to classify healthy and pneumonia patients based on nuclear magnetic resonance (NMR) metabolomics data.[12]

DL or deep neural network, is a new class of machine learning methods that have been successfully applied to various areas of genomics research,[13,14] including predicting the intrinsic molecular subtypes of breast cancer,[15] inferring expression profiles of genes[16] and predicting the functional activity of genomic sequence.[17] In a recent study, denoising autoencoder (DAs), a type of DL algorithm, was applied to gene expression data of the breast cancer.[15] It successfully extracted features that stratify normal/tumor samples, ER+/ER− status, and intrinsic molecular subtypes. In another study based on gene expression data, DL outperformed linear regression in inference of the expression of target genes from the expression of landmark genes.[16] Moreover, an open source conventional neural networks (CNNs) package "Basset" was developed to learn the functional activity of 164 cell types DNA sequences from genomics data and to annotate the noncoding genome.[17] Compared to the flourishing applications of DL in genomics, it remains unknown if deep neural network is suitable to classify metabolomics data, especially when the samples are of medium size (i.e., several hundred).

Here we applied feed-forward networks, a type of DL framework, as an alternative to the machine learning methods such as those listed earlier, to classify metabolomics data. We examined the predictive accuracy of the DL and other machine learning algorithms to predict ER status from a public metabolomics data set.[18] We demonstrated this DL method performs better than a wide cluster of machine learning methods, including RF, SVM, recursive partitioning and regression trees (RPART), linear discriminant analysis (LDA), prediction analysis for microarrays (PAM), and generalized boosted models (GBM). Furthermore, the biological interpretation of the hidden layers reveals eight breast cancer related pathways such as central carbon metabolism in cancer and glutathione metabolism. Moreover, we further analyzed the extracted features from our DL model by mapping the biosynthetic enzymes involved in the metabolomics pathways.

■ MATERIALS AND METHODS

Data Set

The metabolomics data used in this study consists of 271 breast cancer samples (204 ER+ and 67 ER−) collected from a biobank at the Pathology Department of Charité Hospital, Berlin, Germany.[18] Metabolomics profiles of these BC patients can be downloaded from the Supplementary Material of this study.[19] A total of 162 metabolites with known chemical structure were measured using gas chromatography followed by time-of-flight mass spectroscopy (GC-TOFMS) for all tissue samples. A detailed description of the protocols and the platforms used in this study were described in ref 18. For validation, we downloaded gene expression data set GSE59198[20] from the Gene Expression Omnibus (GEO) database, which is composed of 154 samples, a subset of the 271 samples. In this data set, the gene expression profiles of BC tumor tissues (122 ER+ and 32 ER−) were analyzed using the cDNA-mediated annealing, selection, extension and ligation (DASL) assay. A total of 15 927 genes were detected ($p < 0.01$) in at least 10% of the samples after applying spline normalization. Data can be downloaded from GEO repository http://www.ncbi.nlm.nih.gov/geo.

Data Preprocessing

We used K-Nearest Neighbors (KNN) method to impute missing metabolomics data.[21] To adjust for the offset between high and low-intensity features, and to reduce the heteroscedasticity, the logged value of each metabolite was centered by its mean ($\overline{x}$) and autoscaled by its standard deviation (s) as described in eq 1.[22] We used quantile normalization to reduce sample-to-sample variation:[23]

$$\hat{x}_{ij} = \left(\frac{\log_2(x_{ij}) - \overline{x}_i}{s} \right) \tag{1}$$

Deep Learning

DL refers to deep neural network framework, which is widely applied in pattern recognition, image processing, computer vision, and recently in bioinformatics.[13,24,25] Similar to other feed-forward artificial neural networks (ANNs), DL employs more than one hidden layer (y) that connects the input (x) and output layer (z) via a weight (W) matrix as shown in eq 2. Here we used sigmoid function as the activation function:

$$y = sigmoid(Wx + b) \tag{2}$$

Activation value of the hidden layer (y) can be calculated by sigmoid of the multiplication of the input sample x with the weight matrix W and bias b. The transpose of the weight matrix W and the bias b can then be used to construct the output (z) layer, as described in eq 3:

$$z = sigmoid(W'y + b') \tag{3}$$

The best set of the weight matrix W and bias b is expected to minimize the difference between the input layer (x) and the output layer (z). The objective function is called cross-entropy in eq 4 below, in which the optimal parameters are obtained by stochastic gradient descent searching:

$$L_H(x, z) = -\sum_{k=1}^{d} [x_k \log z_k + (1 - x_k) \log(1 - z_k)] \tag{4}$$

To train the model, we first supplied sample input (x) to the first layer and obtained the best parameters (W, b) and the activation of the first hidden layer (y), and then used y to learn the second layer. We repeated this process in subsequent layers, updating the weights and bias in each epoch. We then used back-propagation to tune the parameters of all layers. Finally, we fed the output of the last hidden layer to a softmax classifier which assigned new labels to the samples.[26] We used *h2o* R package to tune the parameters of the DL model.[27]

Other Machine Learning Algorithms

We selected a representative set of six machine-learning algorithms that are highly recommended by the metabolomics community and applied widely in the literature reports: RF, SVM, RPART, LDA, PAM, and GBM. To get the optimal predictions, we used the *caret* R package[28] to tune the parameters in the models.

Modeling and Evaluation

We randomly split metabolomics samples into 80% training set and 20% testing set. The 80/20 split is a common practice of splitting ratio for samples of a moderate size in the machine learning applications. We chose this ratio to having enough

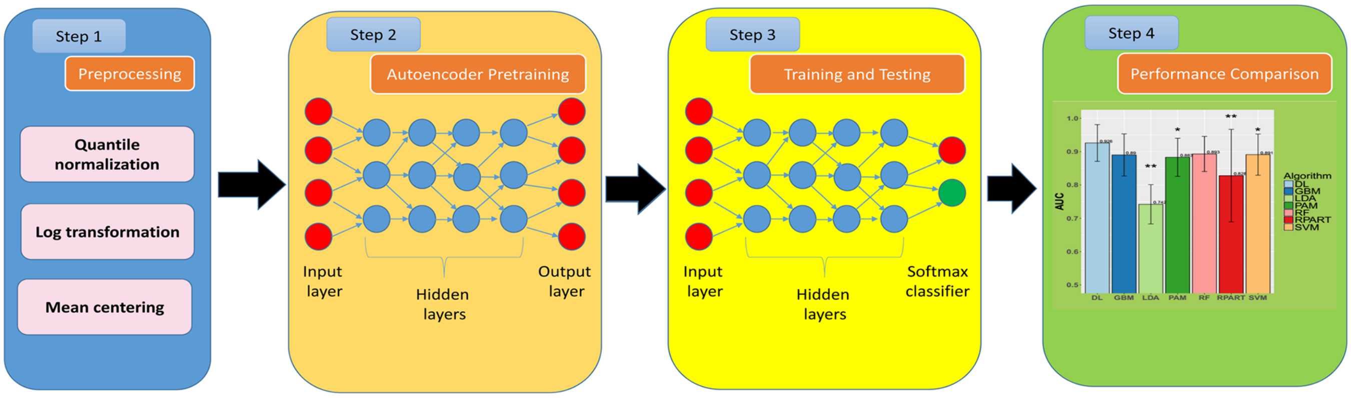


Figure 1. Block diagram of the proposed system. The first step is the preprocessing (log transformation, centering, autoscaling, and quantile normalization). We used Autoencoder pretraining (unsupervised step) to initial model weights and select model architecture. Model used the 80% of data split to train the model and the remaining 20% to measure model performance. The data were split 10 times to avoid the bias of data sampling, and the average AUC was calculated on the 10 hold out test sets.

training samples to build a good model and sufficient testing samples to evaluate the model. We performed 10-fold cross-validation on the 80% training data during the model construction process, and tested the model on the hold out 20% of data. We used pROC R package[29] to compute area under the curve (AUC) of a receiver-operating characteristic (ROC) curve to assess the overall performance of the models. To avoid sampling bias, we repeated the above splitting process ten times and calculated the average AUC on the 10 hold out test sets. To control overfitting, we used two regularization parameters: *L*1, which increases model stability and causes many weights to become 0 and *L*2, which prevents weights enlargement.

We tuned DL model and other machine learning algorithms, on the following parameters: DL model, Epochs (number of passes of the full training set), *l*1 (penalty to converge many weights to 0) and *l*2 (penalty to prevent weights enlargement), and input dropout ratio (ratio of ignored neurons in the input layer during training), number of hidden layers; RPART model, complexity parameters (cost of adding node to the tree); GBM model, number of trees and interaction depths; SVM model, cost of classification; RF model, number of trees to fit; PAM model, threshold amount by for each of the class's centroid shrinking toward the all classes' centroid.

Feature Importance

Features importance was estimated based on model based approach.[28] In other words, a feature is considered important if it contributes to the model performance.[30] We used the variable importance functions varimp in *h2o* and varImp in *caret* R packages to rank models' features.

Identifiers Standardization and Differentially Expressed Genes

We used the PubChem Identifier Exchange Service[31] to convert metabolites into their corresponding KEGG compound IDs; we then used KEGG API[32] to get the compound pathways and enzyme IDs. We used *limma* R package[33] to find enzymes with high fold changes as well as significant adjusted *p*-values between ER+ and ER− samples.

Metabolomics Enzymes Network Reconstruction and Visualization

We used MetScape[34] v3.1.3, a Cytoscape plug-in to generate gene-metabolite network that integrates reaction and pathway information from KEGG and Edinburgh human metabolic network (EHMN) databases. To build enzyme-metabolite network, we selected a pathway based network from MetScape analysis options. The input of this step were two files. The first file included the compound KEGG IDs, *p*-value and the fold change values of the top 20 metabolites extracted from the DL model. The second file included the enzyme KEGG IDs, *p*-value and the fold change values of the 898 genes whose expression values were statistically significantly different between ER− and ER+ samples.

Metabolites Enzymes Correlation

We calculated the correlations between the intensity levels of the metabolites and enzymes using Spearman's Correlation Coefficient in R. We plot the Circos plot of the strongest correlation using *Circlize* R package v0.4.0.

Joint Significant Pathway Analysis

To perform joint significant pathway analysis on metabolomics and gene expression data from the same samples, we considered a comprehensive list of pathways from Reactome, EHMN, and KEGG databases, using online web tool IMPaLA,[35] and calculated hypergeometric *p*-values of genes (P_G) and metabolites (P_M). The joint *p*-value (P_j) between metabolites and genes for pathway *i* was calculated as $P_{ji} = P_{Gi} P_{Mi}$.[36] This value was adjusted to control for multiple testing with the false discovery rate method.

Code Availability

We include all preprocessing and the learning steps of the DL method as an R script in the Supporting Information.

■ RESULTS

Workflow of Autoencoder Based Classification

We aim to assess the predictive ability of the DL framework to separate breast cancer patients based on their ER status, using metabolomics data. Toward this goal, we implemented the workflow of DL framework as in Figure 1. We applied preprocessing steps (log transformation, centering, autoscaling, and quantile normalization) before constructing the DL model, as recommended by others.[18,22] Before training the model, we pretrained the model using autoencoder and the whole data without labels. This step improves the model performance, avoids random initialization of the weights, and selects the best model architecture.[37] Then we trained the DL model using a

DOI: 10.1021/acs.jproteome.7b00595
J. Proteome Res. 2018, 17, 337−347

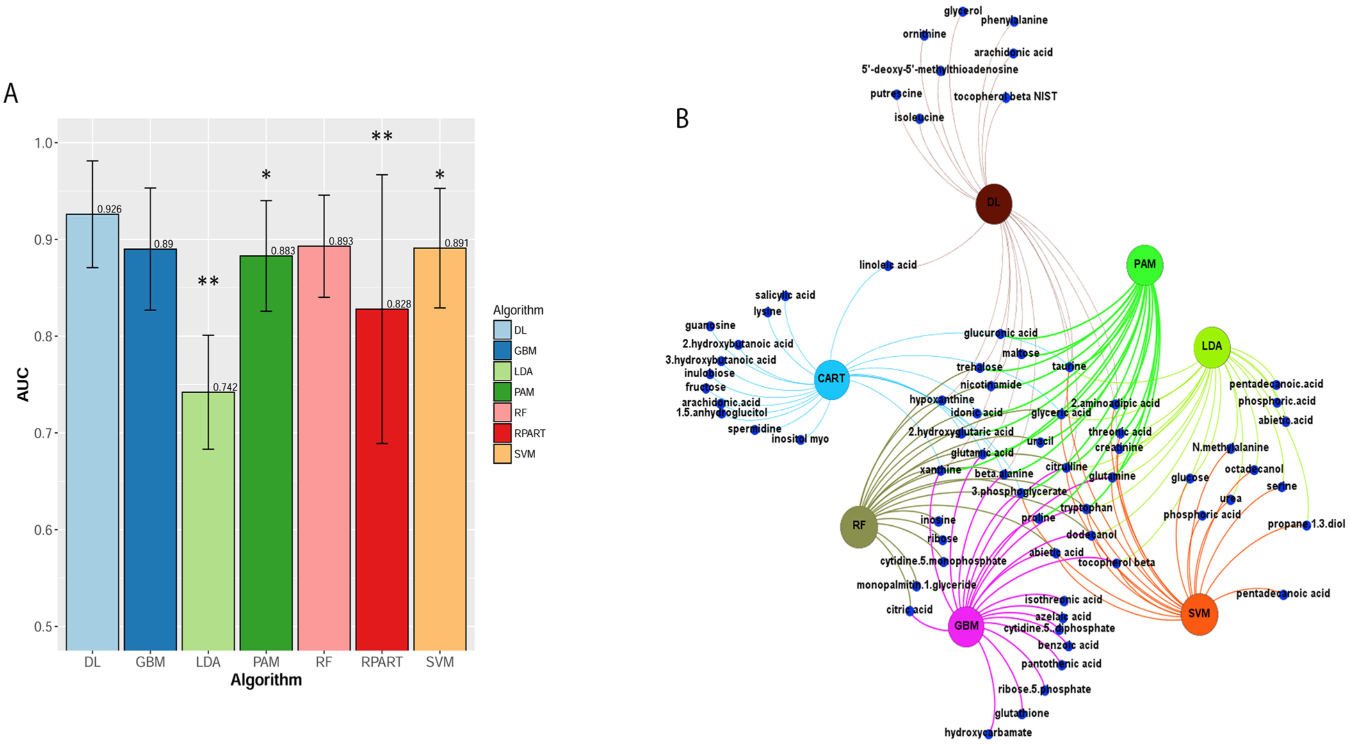


Figure 2. (A) Average AUC on 10 hold out test sets of the DL framework against six machine learning algorithms for prediction of ER status from metabolomics data: recursive partitioning and regression trees (RPART) (0.83), linear discriminant analysis (LDA) (0.74), support vector machine (SVM) (0.89), deep learning (DL) (0.93), random forest (RF) (0.89), generalized boosted models (GBM) (0.89), and prediction analysis for microarrays (PAM) (0.88). The above algorithms were run 10 times on different train/test splits. We used pairwise Wilcoxon signed-rank test to estimate the statistical significance of the difference in performance between DL and other methods ($** p < 0.01$, $* p < 0.1$). (B) Bipartite graph of the top 20 important metabolites extracted from DL model and other machine learning algorithms. Large nodes represent the models and small nodes are metabolites. A connection between metabolite and the model means this metabolite is one of the top 20 high importance metabolites extracted by this model.

wide range of parameters and selected the best model with the minimum mean square error (see Materials and Methods).

Performance of Autoencoder Based Deep Learning Classification

We compared DL with six other machine-learning methods commonly used in the community: RF, SVM, RPART, LDA, PAM, and GBM. To assess the predictive power of the models, we partitioned the data into 80% training and 20% testing subsets. We performed 10-fold cross-validation on the 80% training data, and tested the model on the hold out 20% of data. To avoid sampling bias, we performed 10 independent splitting of training and testing subsets. We reported the averaged AUCs calculated on the hold out test sets. As shown in Figure 2A, the average AUC of DL yields the best AUC of 0.93, compared to other six classification methods. The superiority of DL accuracy is statistically significant (Wilcoxon signed-rank test $P < 0.05$) than other methods, except RF and GBM. LDA and RPAT had the worst accuracy, likely due to their sensitivity to overfitting and being unfit to the nonlinear problems.[38]

DL, as other machine learning algorithm, needs more samples to achieve high accuracy.[39] To assess the effect of sample size on various models, we randomly removed $^1/_4$, $^1/_2$, and $^3/_4$ of the data sets (Figure S1). As expected, decreasing in sample size decreases the averaged AUCs of all classification methods in general except LDA on $^1/_4$ samples due to overfitting. Notably, the reduction of average AUC in DL is most pronounced among all methods, from the full to $^3/_4$ data set (Figure S1). While DL loses the best average AUC status when the sample size is around 255, GBM, SVM, and RF have the highest AUC for small sample sizes of 203, 136 and 68, respectively. Similarly, we also experimented the effect of metabolite size on various models (Figure S2). We randomly removed $^1/_8$, $^1/_4$, and $^1/_2$ of the 162 metabolites. Even with reduced numbers of metabolites, deep learning and the robust machine learning method SVM still have fairly good predictions, compared to other algorithms tested. This suggests that, due to colinearlity, much of the information still exists in the remaining metabolites. Together, the drop-out experiments (Figures S1 and S2) demonstrate that DL method is sensitive to sample size, but much less sensitive to metabolite size.

Important Features from DL

To relate the importance of metabolites to ER status directly, we ranked the metabolites extracted from DL model based on their functional contributions to the outputs. In this approach, features that provide unique information to the trained network are ranked more importantly than those giving redundant information.[40] We listed the top 20 metabolites from DL in Table S1, and presented their heatmap and boxplots in Figure S3. Note that the choice of 20 metabolite is guided by the original study, in which 19 out of 162 metabolites were claimed to change significantly among training and validation samples.[19] The original author divided the 271 samples into two parts, the training ($^2/_3$) and the validation ($^1/_3$) set. Among the training set, 65 metabolites were different in ER− and ER+ and only 19 metabolites were validated in the validation set.

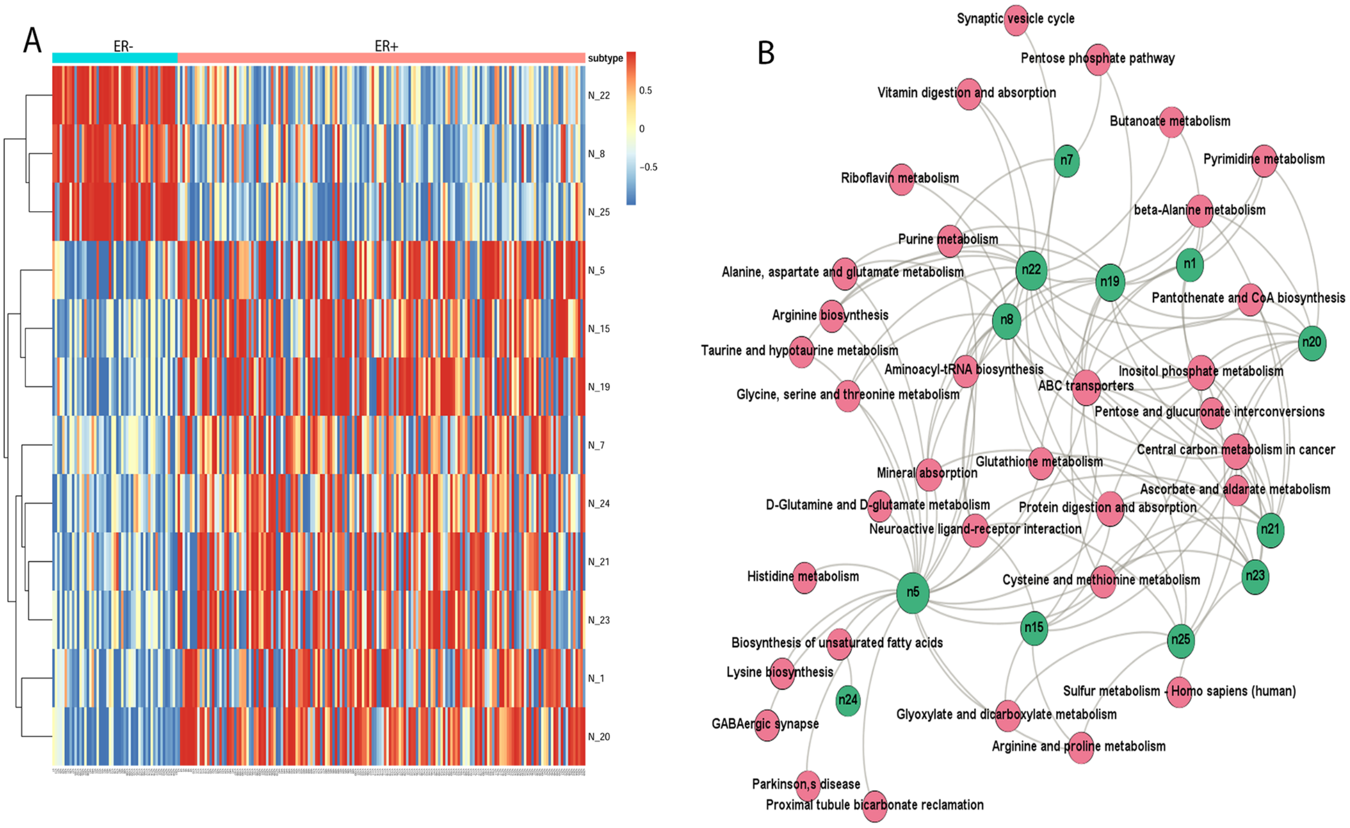


Figure 3. Biological relevance of the DL hidden layers. (A) Activation levels of the high variance nodes extracted from the layer 1 of the DL model. Columns are samples and rows are the top 12 nodes with high variance >0.1. (B) Bipartite graph of enriched significant metabolomics pathways and top hidden nodes. The nodes represent enriched pathways common to all top 12 nodes (green color) in the first hidden layer of DL in KEGG pathway enrichment analysis (FDR< 0.05).

Among the 20 features, the top five features are beta-alanine, xanthine, isoleucine, glutamate, and taurine. These five metabolites have been either proposed as breast cancer biomarkers or associated with breast cancers in the original metabolomics report[19] and/or other studies.[6,8,41−43] For instance, Budczies et al.[19] found that beta-alanine had the most significant and largest fold changes between ER−(n = 67) and ER+ (n = 204) tumor tissues. In another study, Glutamate was suggested as markers to segregate ER− from ER+ in the training (n = 186) as well as validation data set (n = 88).[8] Glutamate to glutamine ratio (GGR) was significantly increased in the ER− tumors as compared to ER+. Overall survival analyses suggested GGR as a positive prognostic marker for BC.[8] In another study, Fan et al. classified BC plasma samples into subtypes, that is, ER+ versus ER− and HER2+ versus HER2-, based on a training set (n = 51) and another test set (n = 45).[6] They found isoleucine had significant differential level between ER+ (lower) and ER− (higher) samples. Similarly, a study among female breast cancer patients (n = 50) suggested serum taurine as an early marker, where its level was significantly lower than the normal (n = 20) and high risk samples (n = 15).[42] In a cell line based study, xanthine was suggested as potential biomarker of breast cancer metastasis,[43] as it had the highest variable influence on projection (VIP) in the three pairwise comparisons among MCF-7/MCF-10A, MDA-MB-231/MCF-10A, and MDA-MB-231/MCF-7.[43]

Further, we compared DL-based top 20 features with the same number of top features from all other methods in a bipartite graph (Figure 2B). Twelve metabolites are shared between DL and one or more algorithms. Among them, one (xanthine) is shared by six methods, two (glyceric acid and citrulline) are shared by five methods, four (glutamine, taurine, glutamic acid, and beta-alanine) are shared by four methods, one (2-aminoadipic acid) is shared by three methods, and two (nicotinamide acid and trehalose) are shared by two methods (Table S1). Additionally, DL has identified eight unique metabolites: isoleucine, putrescine, glycerol, 5′-deoxy-5′-methylthioadenosine, ornithine, tocopherol beta, phenylalanine, and arachidonic acid,

Biological Relevance of Hidden Layers

To understand the high performance of the DL model, we probed into the hidden layer and analyzed the 25 activation nodes from the first hidden layer. Among the top 12 nodes with the variances >0.1, node 8, 22, and 25 are significantly correlated with the samples' ER− status (P = 1.14e−12), whereas all other top nine nodes are associated with the ER+ status (Figure 3A). These results confirm that the nodes in DL have significant biological meaning.

We identified a total of 129 metabolites which contribute most to the activation values of the top 12 nodes. Their relationships between the 129 metabolites and 12 nodes are shown in Figure S4. We define that metabolite x contributes to the activation value (y) of node n, if the aboslute value of the weight connecting metabolite x and node n is greater than 0.1. Beta-alanine and xanthine are the most common metabolites from all top 12 nodes. Among nodes 8, 22, and 25 which are highly correlated with ER− (Figure 3A), four common metabolites are shared: inositol, glutamate, xanthine, and uracil.

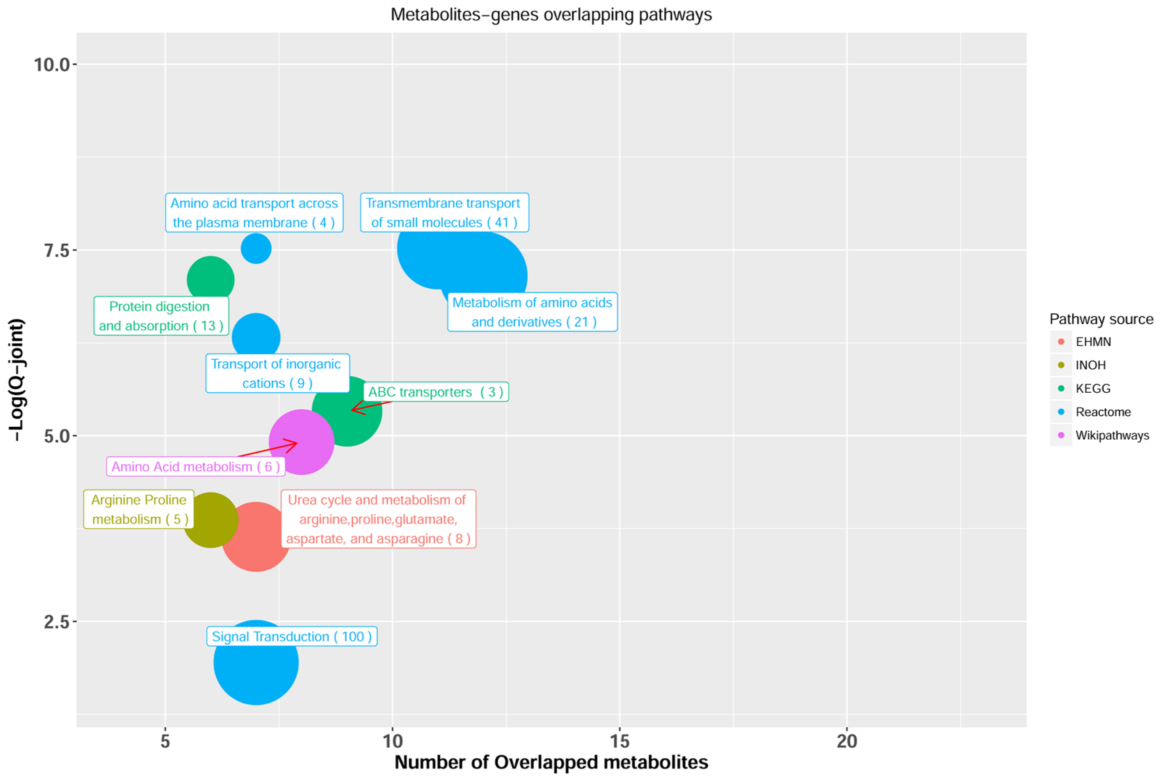


Figure 4. Joint pathway analysis between the top 20 DL metabolites and the highly differentiated enzymes. Only significant pathways with at least five overlapping metabolites are shown. *X*-axis shows the number of overlapped metabolites with the number of genes (number in parentheses) involved in the same pathway, *y*-axis shows the adjusted joint *P*-value calculated from IMPALA tool.[42] The size of the nodes represents the size of metabolomic pathway (number of metabolites involved in that pathway). The color of the nodes represents the database source of these pathways.

Xanthine was among the panel of prognostic markers of breast cancer metastasis based on the metabolic profiling of the three breast cancer cell lines.[43] Glutamate has been reported as biomarker to segregate ER− from ER+ in the training as well as validation data set, as described earlier.[8] Inositol phosphate metabolism pathway was previously reported to be associated with breast cancer, but not between ER+ and ER− cancers.[44] Uracil is, however, a potencial new marker for ER− breast cancer that was not reported previously, according to our knowledge.

To link the metabolites in Figure S4 with biological functions, we conducted pathways enrichment analysis using online web tool IMPaLA.[35] The pathways are taken from Reactome, EHMN, and KEGG databases. Eight significant breast cancer related pathways (Figure 3B) are enriched in all nodes: protein digestion and absorption, central carbon metabolism in cancer, neuroactive ligand receptor interaction, ABC transporters, mineral absorption, inositol phosphate metabolism, glutathione metabolism, and cysteine and methionine metabolism. Albeit the name of "neuroactive ligand-receptor interaction", this pathway is significantly enriched (q-value = 0.001) and it was shown changed in breast cancer cell lines[45] and naked mole rat.[46] Aspartate, glycine, taurine, and glutamate are metabolites associated with this pathway in the metabolic data set. Another interesting pathway with the name "mineral absorption" also shows significance (q-value = 7.51 × 10^{-06}), attributed by five metabolites tryptophan, alanine, glycine, phosphoric acid, and glutamine. All these five metabolites were found related with breast cancer previously.[47−49]

Integration of DL Metabolites and Enzymes

We further aimed to validate the important metabolite features of DL model by integrating metabolomics and gene expression data from the same patients. Toward this, we first conducted a joint pathway analysis between 20 metabolites extracted from DL model and 898 significantly differentiated enzymes between ER+ and ER− samples using IMPALA (Figure 4). Most of the top significant pathways are related to metabolism of amino acids or protein digestion and absorption. Two pathways remain significant in joint pathway analysis by comparing to metabolomics based pathway analysis in Figure 3B: protein digestion and absorption and ABC transporters, with six and nine metabolites over-represented, respectively. Specifically, urea, inositol allo-, phosphoric acid, glucose, glutamine, isoleucine, and glutathione are the associated metabolites in ABC transporters. For protein digestion, glutamine, lysine, isoleucine, and beta-alanine are associated metabolites. Some literature evidences show that protein digestion and ABC transporters are related to breast cancer. For example, humans have 49 members of the ATP-binding cassette (ABC) membrane proteins.[50] Several of them, such as ABCB1 and ABCC1, have developed "multidrug resistance" (MDR) in breast cancer, when they are overexpressed over a period of time.[51]

To gain insights at individual metabolite/enzyme level, we then calculated Spearman's correlations between the intensity levels of the top 20 metabolites and enzymes whose gene expression levels are significantly different between ER+/ER− for the same patients.[20] The Circos plot in Figure 5 shows the names of metabolomics and enzymes that have correlations ($|r|$ > 0.35). Impressively, beta-alanine, the top ranked metabolite in

DOI: 10.1021/acs.jproteome.7b00595

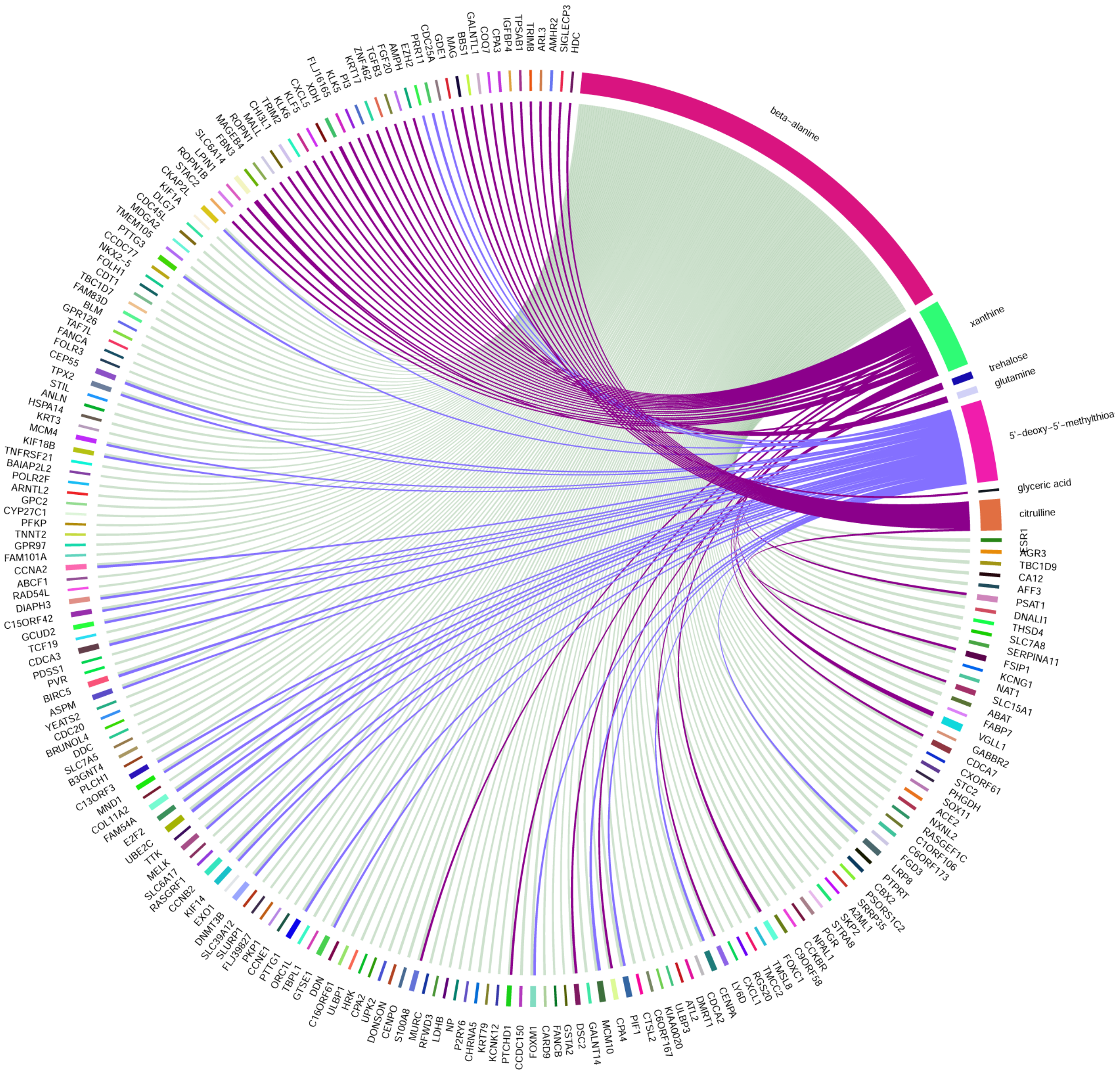


Figure 5. Circos plot of Spearman's correlation values between top 20 DL metabolites and highly differentiated enzymes with cutoff = |0.35|.

DL, is the single most connected metabolite, correlated to more than 100 significantly differentially expressed enzymes. Pathway analysis of these enzymes correlated with beta-alanine shows strikingly significant enrichment (adjusted *p*-value = 3.84e−05) with FOXM1 transcription factor network pathway. FOXM1 is highly expressed in ER− samples and with a correlation coefficient $r = 0.5$ with beta-alanine.

Complementary to the correlation based analysis, we also used MetScape (Cytoscape plug-in) for gene-metabolite network analysis, by combining the ER+/ER− metabolomics data[18] and gene expression (from GSE59198)[20] for the same patients. ABAT, the enzyme that catalyzes beta-alanine to malonate semialdehyde (Figure 6B), is highly correlated with beta-alanine ($r = -0.62$, Figure 6A). To understand better the connection between beta-alanine and FOX genes family, we performed motif enrichment analysis for the enzymes interacted with beta-alanine in Figure 6B using PASTAA tool.[52] Interestingly, FOXO1 was one of most significant transcription factors ($p = 5.89e{-}04$) that targeted the promoter regions of beta-alanine interacting enzymes.

DISCUSSION

Metabolomics has become a new platform for biomarker discovery. Accompanying this technology, robust and accurate classification methods to predict sample labels are in critical need. Recently, DL methods have gained much attention in domains such as genomics and imaging analysis. However, there has not been any systematic investigation of DL methods in the metabolomics space. In this report, we aimed to fill this void and assessed the performance of feed-forward network, a widely used DL framework, on classifying ER+/ER− breast cancer metabolomics data.

There are many advantages of DL over shallow machine learning algorithms, which are beyond the scope of this study. The conventional machine learning algorithms require engineering domain knowledge to create features from raw data, whereas DL automatically extracts simple features from the input data using general purpose learning procedure. These simple features are mapped into outputs using a complex architecture composed of a series of nonlinear functions "hierarchical representations," to maximize the predictive

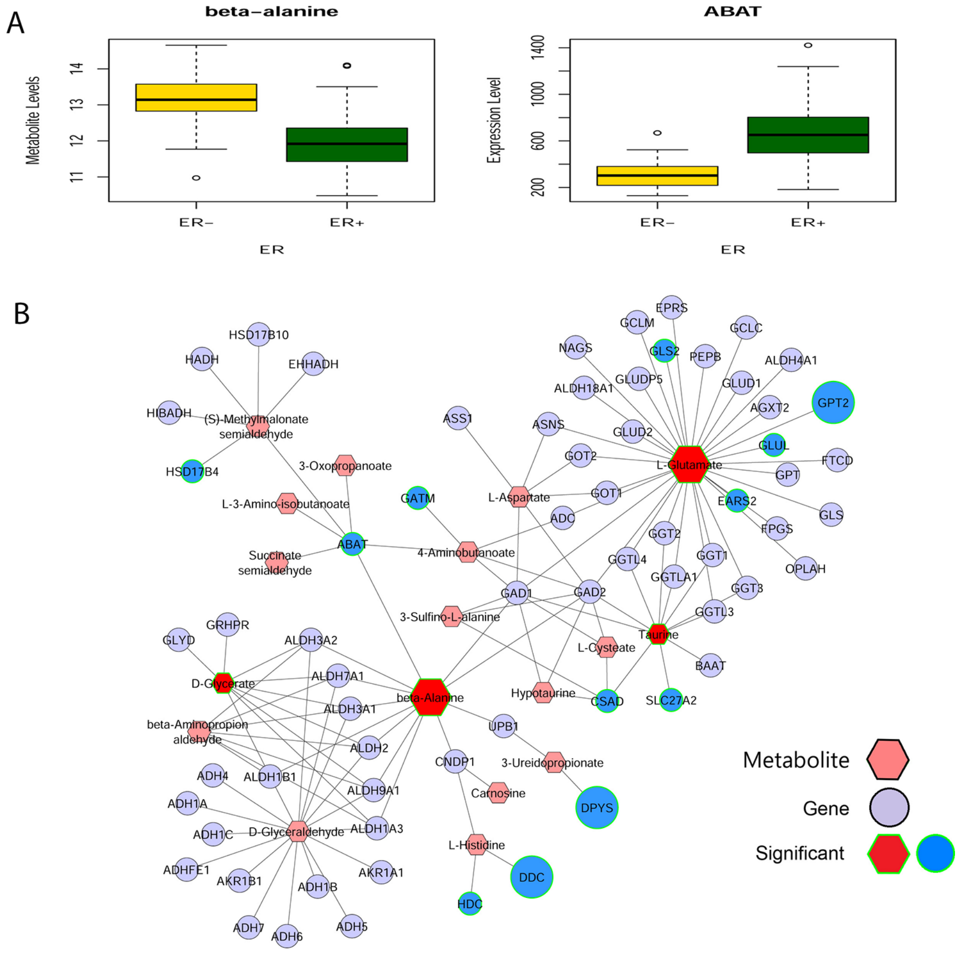


Figure 6. Beta-alanine and ABAT interaction network. (A) Metabolite level of beta-alanine and expression of ABAT. (B) Beta-alanine-ABAT interaction network in ER− breast cancer tissues compared to ER+ breast cancer tissues. MetScape, a Cytoscape plug-in, was used to integrate ER+/ER− metabolomics and gene expression data (GSE59198) of the same patients. Fold change of metabolites (hexagon nodes) or enzymes (circle nodes) are represented by the size of the nodes. The input of MetScape are the top 20 metabolites from the DL model and the 898 genes whose expression values are statistically significantly different between ER− and ER+ samples. Enzymes and metabolites with significant difference are marked by green line(s) on the shapes.

accuracy of the model optimally. By increasing number of layers and neurons per layers, robust features may be constructed, and error signals can be diminished as they pass through multiple layers.[13] Therefore, DL succeeds to construct high-level transformed features from input data, making it more desirable than shallow machine learning algorithms in this respect.[14]

We demonstrated that DL has a higher predictive accuracy over the other six popular machine learning methods in detecting ER status from metabolomics data. DL exploits the idea that the higher "succeeding" layer is learned from the lower "preceding" layer and selects the essential metabolites from DL model. These metabolites are useful for the learning process and explain the high predictability of DL compared to conventional machine learning algorithms. DL extracted features that could be considered as novel biomarkers, such as uracil, which were not previously reported as breast cancer. Also, unlike other machine learning methods, DL method offers additional insights on eight KEGG pathway being significantly different due to ER status. All these new observations warrant further investigation.

An interesting new link we discover lies between FOXM1 family and beta-alanine. A recent study showed FOXM1 to be a major cause for resistance to various chemotherapeutics,[53] and reduction of FOXM1 levels induced apoptosis of breast cancer cells.[54] The motif enrichment analysis of the beta-alanine interacted enzymes indicates that the transcription factor FOXO1 targeted the promoter regions of these enzymes. Thus, the relationships among beta-alanine, FOXM1, and FOXO1 are worth further investigation. In addition, we found many interesting involvement of DL-based unique metabolites in breast cancer diagnosis and treatment. For example, phenylalanine is found significantly elevated in the advanced metastatic breast cancer[55] and linoleic acid has been used to lower the risk of breast cancer.[56] Also, putrescine has been known to play a critical role in many metabolomics processes in breast cancer, such as apoptosis, and proliferation.[57] The knock-down experiments on ornithine decarboxylase (ODC), an enzyme which converts ornithine to putrescin, showed the growth inhibition in the ERα+ MCF7 and T47D and ERα- MDA-MB-231 breast cancer cells.[58] Arachidonic acid was previously shown to be integral part of the new signaling for the cell migrations in the MDA-MB-231 breast cancer cells.[59]

Despite the outstanding performance of DL methods, one should be mindful of several caveats in its application in metabolomics research. DL is time-consuming computation (Table S2), relative to some other machine learning methods.[40]

Also, metabolomics data sets are generally small, in comparison to imaging data. Thus, very small data sets may not be suitable for DL. We experimented with the effects of reducing sample size and metabolite size on the seven methods in comparison, and found that DL is indeed sensitive to the sample size of the study. On the contrary, due to colinearlity among metabolites, autoencoder has fairly robust predictions even when the number of metabolites are reduced. Another point of consideration is the reproducibility of the technology itself. A platform with better reproducibility is expected to yield biomarker models that predict more accurately in validation data sets (less overfitting). We thus speculate that DL models based on NMR metabolomics data (more metabolites and better reproducibility) will be more accurate than DL models based on LC−MS data, when other conditions are the same.

Lastly, in this report we compared the ML versus DL under the topic of classification of metabolomics data. The advantages of DL on other nonclassification problems in metabolomics research are yet to be explored. For example, unsupervised machine learning algorithms such as PCA and hierarchical clustering were applied to the metabolomics data,[60] and our group is currently exploring using autoencoders for unsupervised learning in metabolomics data. As another example, we have also worked on prognosis prediction using shallow and deep neural network models in the genomics space.[61,62] We successfully used autoencoder to integrate multiple omics data sets (RNA-Seq, microRNA-Seq and DNA methylation) to predict patient survival robustly, exemplified by liver cancer.[62] Compared to genomics data, metabolomics data have higher multicolinearity and noise levels. Also the number of identifiable metabolites are lower than the identifiable genes in genomics assays. These issues pose potential challenges when extending genomics tools for metabolomics research. Nevertheless, it will be very interesting to test these DL and neural network models on appropriate metabolomics data sets alone or in combination with coupled genomics data.

CONCLUSIONS

We show evidence that DL outperforms other machine learning algorithms for ER status classification in breast cancer metabolomics data. The biological interpretation of the hidden layer of the DL model also reveals eight significant breast cancer related pathways, which are not able to obtain from the other machine learning algorithms in comparison.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.7b00595.

Effect of sample and metabolite size on performance of DL and other machine learning algorithms; DL 20 important metabolites; heatmap of the metabolites that most contribute to activation value of top hidden nodes (PDF)

List of top 20 important features (XLSX)

Running time of seven algorithms on metabolomics data set (XLSX)

R code of preprocessing, models training, and testing (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: lgarmire@cc.hawaii.edu. Phone: +1 (808) 441-8193.

ORCID

Fadhl M. Alakwaa: 0000-0001-5349-7960

Author Contributions

L.X.G. and F.M.A. envisioned the project and designed the work. F.M.A. coded the project and conducted the analysis. K.C. mapped metabolites and enzymes into KEGG pathway. F.M.A. wrote the manuscript with help from L.X.G. and K.C. L.X.G., F.M.A., and K.C. have read, revised, and approved the final manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was supported by Grant Nos. K01ES025434 awarded by NIEHS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (http://datascience.nih.gov/bd2k), P20 COBRE GM103457 awarded by NIH/NIGMS, R01 LM012373 awarded by NLM, and R01 HD084633 awarded by NICHD to L.X.G. We would like to thank Dr. Carsten Denkert and Dr. Oliver Fiehn for making their metabolomics data accessible. We thank all members in Garmire group for reviewing and commenting on the manuscript.

REFERENCES

(1) *Breast Cancer: Prevention and Control*; World Health Organization, 2017. http://www.who.int/cancer/detection/breastcancer/en/index1.html (accessed October 10, 2017).

(2) *About Breast Cancer*; American Cancer Society, 2017. https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html (accessed September 21, 2017).

(3) Carey, L. A.; Perou, C. M.; Livasy, C. A.; Dressler, L. G.; Cowan, D.; Conway, K.; Karaca, G.; Troester, M. A.; Tse, C. K.; Edmiston, S.; Deming, S. L.; Geradts, J.; Cheang, M. C.; Nielsen, T. O.; Moorman, P. G.; Earp, H. S.; Millikan, R. C. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA* **2006**, *295* (21), 2492−2502.

(4) O'Brien, K. M.; Cole, S. R.; Tse, C. K.; Perou, C. M.; Carey, L. A.; Foulkes, W. D.; Dressler, L. G.; Geradts, J.; Millikan, R. C. Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study. *Clin. Cancer Res.* **2010**, *16* (24), 6100−6110.

(5) Haque, R.; Ahmed, S. A.; Inzhakova, G.; Shi, J.; Avila, C.; Polikoff, J.; Bernstein, L.; Enger, S. M.; Press, M. F. Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades. *Cancer Epidemiol., Biomarkers Prev.* **2012**, *21* (10), 1848−1855.

(6) Fan, Y.; Zhou, X.; Xia, T. S.; Chen, Z.; Li, J.; Liu, Q.; Alolga, R. N.; Chen, Y.; Lai, M. D.; Li, P.; Zhu, W.; Qi, L. W. Human plasma metabolomics for identifying differential metabolites and predicting molecular subtypes of breast cancer. *Oncotarget* **2016**, *7* (9), 9925−9938.

(7) Tang, X.; Lin, C. C.; Spasojevic, I.; Iversen, E. S.; Chi, J. T.; Marks, J. R. A joint analysis of metabolomics and genetics of breast cancer. *Breast Cancer Res.* **2014**, *16* (4), 415.

(8) Budczies, J.; Pfitzner, B. M.; Gyorffy, B.; Winzer, K. J.; Radke, C.; Dietel, M.; Fiehn, O.; Denkert, C. Glutamate enrichment as new diagnostic opportunity in breast cancer. *Int. J. Cancer* **2015**, *136* (7), 1619−1628.

(9) Lien, E. C.; Lyssiotis, C. A.; Juvekar, A.; Hu, H.; Asara, J. M.; Cantley, L. C.; Toker, A. Glutathione biosynthesis is a metabolic

vulnerability in PI(3)K/Akt-driven breast cancer. *Nat. Cell Biol.* **2016**, *18* (5), 572−578.

(10) Truong, Y.; Lin, X.; Beecher, C. Learning a Complex Metabolomic Dataset Using Random Forests and Support Vector Machines. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004*; ACM, 2004; pp 835−840.

(11) Huang, J.-H.; Yan, J.; Wu, Q.-H.; Duarte Ferro, M.; Yi, L.-Z.; Lu, H.-M.; Xu, Q.-S.; Liang, Y.-Z. Selective of informative metabolites using random forests based on model population analysis. *Talanta* **2013**, *117*, 549−555.

(12) Mahadevan, S.; Shah, S. L.; Marrie, T. J.; Slupsky, C. M. Analysis of Metabolomic Data Using Support Vector Machines. *Anal. Chem.* **2008**, *80* (19), 7562−7570.

(13) Min, S.; Lee, B.; Yoon, S. Deep learning in bioinformatics. *Briefings Bioinf.* **2016**, bbw068.

(14) Angermueller, C.; Parnamaa, T.; Parts, L.; Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **2016**, *12* (7), 878.

(15) Tan, J.; Ung, M.; Cheng, C.; Greene, C. S. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. *Pac Symp. Biocomput* **2015**, 132−143.

(16) Chen, Y.; Li, Y.; Narayan, R.; Subramanian, A.; Xie, X. Gene expression inference with deep learning. *Bioinformatics* **2016**, *32* (12), 1832−1839.

(17) Kelley, D. R.; Snoek, J.; Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **2016**, *26* (7), 990−999.

(18) Budczies, J.; Denkert, C.; Muller, B. M.; Brockmoller, S. F.; Klauschen, F.; Gyorffy, B.; Dietel, M.; Richter-Ehrenstein, C.; Marten, U.; Salek, R. M.; Griffin, J. L.; Hilvo, M.; Oresic, M.; Wohlgemuth, G.; Fiehn, O. Remodeling of central metabolism in invasive breast cancer compared to normal breast tissue - a GC-TOFMS based metabolomics study. *BMC Genomics* **2012**, *13*, 334.

(19) Budczies, J.; Brockmoller, S. F.; Muller, B. M.; Barupal, D. K.; Richter-Ehrenstein, C.; Kleine-Tebbe, A.; Griffin, J. L.; Oresic, M.; Dietel, M.; Denkert, C.; Fiehn, O. Comparative metabolomics of estrogen receptor positive and estrogen receptor negative breast cancer: alterations in glutamine and beta-alanine metabolism. *J. Proteomics* **2013**, *94*, 279−288.

(20) Edgar, R.; Domrachev, M.; Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **2002**, *30* (1), 207−210.

(21) Chen, J.; Shao, J. Nearest Neighbor Imputation for Survey Data. *J. Official Statistics* **2000**, *16* (2), 113−131.

(22) van den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **2006**, *7*, 142.

(23) Jauhiainen, A.; Madhu, B.; Narita, M.; Narita, M.; Griffiths, J.; Tavare, S. Normalization of metabolomics data with applications to correlation maps. *Bioinformatics* **2014**, *30* (15), 2155−2161.

(24) Li, H. Deep learning for image denoising. *International Journal of Signal Processing, Image Processing and Pattern Recognition* **2014**, *7* (3), 171−180.

(25) LeCun, Y.; Kavukcuoglu, K.; Farabet, C. *Proceedings of 2010 IEEE International Symposium on Convolutional Networks and Applications in Vision, Circuits, and Systems (ISCAS), 2010*; IEEE, 2010; pp 253−256.

(26) Lee, H. Tutorial on deep learning and applications, *NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning*; NIPS, 2010.

(27) Candel, A.; Parmar, V.; LeDell, E.; Arora, A. *Deep Learning with H2O*; H2O.ai, 2015.

(28) Kuhn, M. Caret package. *Journal of Statistical Software* **2008**, *28* (5), 1−26.

(29) Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J. C.; Muller, M. pROC: an open-source package for R and S + to analyze and compare ROC curves. *BMC Bioinf.* **2011**, *12*, 77.

(30) Gedeon, T. D. Data mining of inputs: analysing magnitude and functional measures. *International Journal of Neural Systems* **1997**, *8* (02), 209−218.

(31) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases. *Nucleic Acids Res.* **2016**, *44* (D1), D1202−1213.

(32) Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **2012**, *40* (D1), D109−D114.

(33) Smyth, G. K. Limma: Linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*; Springer, 2005; pp 397−420.

(34) Karnovsky, A.; Weymouth, T.; Hull, T.; Tarcea, V. G.; Scardoni, G.; Laudanna, C.; Sartor, M. A.; Stringer, K. A.; Jagadish, H. V.; Burant, C.; Athey, B.; Omenn, G. S. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics* **2012**, *28* (3), 373−380.

(35) Kamburov, A.; Cavill, R.; Ebbels, T. M.; Herwig, R.; Keun, H. C. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics* **2011**, *27* (20), 2917−2918.

(36) Cavill, R.; Kamburov, A.; Ellis, J. K.; Athersuch, T. J.; Blagrove, M. S.; Herwig, R.; Ebbels, T. M.; Keun, H. C. Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. *PLoS Comput. Biol.* **2011**, *7* (3), e1001113.

(37) Pasa, L.; Sperduti, A. *Pre-training of Recurrent Neural Networks via Linear Autoencoders*; Advances in Neural Information Processing Systems, 2014; pp 3572−3580.

(38) Lee, C.; Nkounkou, B.; Huang, C. H. Comparison of LDA and SPRT on Clinical Dataset Classifications. *Biomed Inform Insights* **2011**, *4*, BII.S6935−7.

(39) Cho, J.; Lee, K.; Shin, E.; Choy, G.; Do, S. *How Much Data Is Needed To Train a Medical Image Deep Learning System To Achieve Necessary High Accuracy?* arXiv preprint arXiv:1511.06348, **2015**.

(40) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016.

(41) Fini, M. A.; Monks, J.; Farabaugh, S. M.; Wright, R. M. Contribution of xanthine oxidoreductase to mammary epithelial and breast cancer cell differentiation in part modulates inhibitor of differentiation-1. *Mol. Cancer Res.* **2011**, *9* (9), 1242−1254.

(42) El Agouza, I. M.; Eissa, S. S.; El Houseini, M. M.; El-Nashar, D. E.; Abd El Hameed, O. M. Taurine: a novel tumor marker for enhanced detection of breast cancer among female patients. *Angiogenesis* **2011**, *14* (3), 321−330.

(43) Kim, H. Y.; Lee, K. M.; Kim, S. H.; Kwon, Y. J.; Chun, Y. J.; Choi, H. K. Comparative metabolic and lipidomic profiling of human breast cancer cells with different metastatic potentials. *Oncotarget* **2016**, *7* (41), 67111−67128.

(44) Tan, J.; Yu, C. Y.; Wang, Z. H.; Chen, H. Y.; Guan, J.; Chen, Y. X.; Fang, J. Y. Genetic variants in the inositol phosphate metabolism pathway and risk of different types of cancer. *Sci. Rep.* **2015**, *5*, 8473.

(45) Huan, J.; Wang, L.; Xing, L.; Qin, X.; Feng, L.; Pan, X.; Zhu, L. Insights into significant pathways and gene interaction networks underlying breast cancer cell line MCF-7 treated with 17β-Estradiol (E2). *Gene* **2014**, *533* (1), 346−355.

(46) Yang, Z.; Zhang, Y.; Chen, L. *In Silico Identification of Novel Cancer-Related Genes by Comparative Genomics of Naked Mole Rat and Rat*; IEEE, 2012; pp 285−290.

(47) Hensley, C. T.; Wasti, A. T.; DeBerardinis, R. J. Glutamine and cancer: cell biology, physiology, and clinical opportunities. *J. Clin. Invest.* **2013**, *123* (9), 3678−3684.

(48) Amelio, I.; Cutruzzola, F.; Antonov, A.; Agostini, M.; Melino, G. Serine and glycine metabolism in cancer. *Trends Biochem. Sci.* **2014**, *39* (4), 191−198.

(49) Lyon, D. E.; Walter, J. M.; Starkweather, A. R.; Schubert, C. M.; McCain, N. L. Tryptophan degradation in women with breast cancer: a pilot study. *BMC Res. Notes* **2011**, *4*, 156.

(50) Sun, Y. L.; Patel, A.; Kumar, P.; Chen, Z. S. Role of ABC transporters in cancer chemotherapy. *Aizheng* **2012**, *31* (2), 51−57.

(51) Liu, Y.; Peng, H.; Zhang, J. T. Expression profiling of ABC transporters in a drug-resistant breast cancer cell line using AmpArray. *Mol. Pharmacol.* **2005**, *68* (2), 430−438.

(52) Thomas-Chollier, M.; Hufton, A.; Heinig, M.; O'Keeffe, S.; Masri, N. E.; Roider, H. G.; Manke, T.; Vingron, M. Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nat. Protoc.* **2011**, *6* (12), 1860−1869.

(53) Park, Y. Y.; Jung, S. Y.; Jennings, N. B.; Rodriguez-Aguayo, C.; Peng, G.; Lee, S. R.; Kim, S. B.; Kim, K.; Leem, S. H.; Lin, S. Y.; Lopez-Berestein, G.; Sood, A. K.; Lee, J. S. FOXM1 mediates Dox resistance in breast cancer by enhancing DNA repair. *Carcinogenesis* **2012**, *33* (10), 1843−1853.

(54) Bergamaschi, A.; Madak-Erdogan, Z.; Kim, Y. J.; Choi, Y. L.; Lu, H.; Katzenellenbogen, B. S. The forkhead transcription factor FOXM1 promotes endocrine resistance and invasiveness in estrogen receptor-positive breast cancer by expansion of stem-like cancer cells. *Breast Cancer Res.* **2014**, *16* (5), 436.

(55) Jobard, E.; Pontoizeau, C.; Blaise, B. J.; Bachelot, T.; Elena-Herrmann, B.; Tredan, O. A serum nuclear magnetic resonance-based metabolomic signature of advanced metastatic human breast cancer. *Cancer Lett.* **2014**, *343* (1), 33−41.

(56) Arab, A.; Akbarian, S. A.; Ghiyasvand, R.; Miraghajani, M. The effects of conjugated linoleic acids on breast cancer: A systematic review. *Adv. Biomed. Res.* **2016**, *5*, 115.

(57) Lessard, M.; Zhao, C.; Singh, S. M.; Poulin, R. Hormonal and feedback regulation of putrescine and spermidine transport in human breast cancer cells. *J. Biol. Chem.* **1995**, *270* (4), 1685−1694.

(58) Zhu, Q.; Jin, L.; Casero, R. A.; Davidson, N. E.; Huang, Y. Role of ornithine decarboxylase in regulation of estrogen receptor alpha expression and growth in human breast cancer cells. *Breast Cancer Res. Treat.* **2012**, *136* (1), 57−66.

(59) Navarro-Tito, N.; Soto-Guzman, A.; Castro-Sanchez, L.; Martinez-Orozco, R.; Salazar, E. P. Oleic acid promotes migration on MDA-MB-231 breast cancer cells through an arachidonic acid-dependent pathway. *Int. J. Biochem. Cell Biol.* **2010**, *42* (2), 306−317.

(60) Xia, J.; Wishart, D. S. D.S Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis. *Curr. Protoc. Bioinform* **2016**, *55* (14), 14.10.1.

(61) Ching, T.; Zhu, X.; Garmire, L. Cox-nnet: an artificial neural network Cox regression for prognosis prediction. *bioRxiv* **2016**, 1 DOI: 10.1101/093021.

(62) Chaudhary, K.; Poirion, O. B.; Lu, L.; Garmire, L. X. Deep Learning based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.* **2017**, 1.