Medical Multi-omics

Liu Yifeng, Pan Yunzhou & Tian Yuchen
Instructors: Yuan Yang, Hu Zeping, Zhao Yizi*, Li Biao*

December 26, 2022

Abstract

Medical multi-omics is a comprehensive approach for exploring the interaction of multiple substances in the body, including genomics, proteomics, metabolomics, etc. It has the potential to discover new strategies for prevention and treatment of diseases. Based on KEGG database, we developed and implemented 4 different methods to analyze variant multi-omics datasets related to Zika virus(ZIKV), COVID-19 and Three Negative Breast Cancer(TNBC), including rule-based non-AI method, feed-forward network for partially unmatched samples, SHAP analysis on XGBoost model and improved P-Net model. Our results have revealed some significant disturbances for these diseases. Particularly, our method discovered ENTPD1, which encodes CD39, as a potentially crucial gene for TNBC, which is consistent with the experimental results as previously reported [1]. However, due to lack of big datasets from large cohort samples, especially consistent multi-omic dataset, our analyses are limited. It is promising that more discoveries can be made from large cohort datasets by using our developed methods. See `https://github.com/lauyikfung/multiomics` for all the codes.

1 Introduction

Multiomics is a comprehensive biological analysis approach on multiple omics including the genomics, proteomics, transcriptomics, metabolomics etc. Synthesis of different omics can provide researchers with a better understanding of the flow of information [2], and thus discovering more associations between biological entities, new biomarkers and therapeutic targets for precise treatment of diseases. Now analysis on the level of multi-omics is a burgeoning method for researching diseases. However, although there are lots of high-throughput multi-omics data, most of the previous researches are limited to analysis on the statistics [3, 4], and some researches so far are just simple superposition of monomics analysis [5].

Artificial Intelligence is a set of technologies which can capture the latent patterns and inexplicit relationships among features of a great amount of samples. In recent years, more and more medical scientists are applying AI methods to make breakthroughs in their fields [6, 7]. Our project is focused on applying AI algorithms to break through the bottlenecks of traditional analysis strategies on integration of transcriptomics, proteomics and metabolomics data.

We are based on KEGG database[1], which contains 50,650 genes, 19,418 compounds and 25,252 orthologs. They are also related by 11,776 reactions, 459 modules as well as 690 pathways. Near 100 thousand entities on KEGG database form a huge network of metabolism. And we are also based on three different datasets:

- ZIKV dataset from Hu-Lab, including proteomics, metabolomics and genomics information of healthy people and ZIKV patients;

*Zhao Yizi from School of Pharmaceutical Science and Li Biao from Shanghai Qi Zhi Institute have helped construct the KEGG database on Neo4j. Zhao also provided some ideas of these methods from a biological perspective.

[1]See `https://www.kegg.jp` for details.

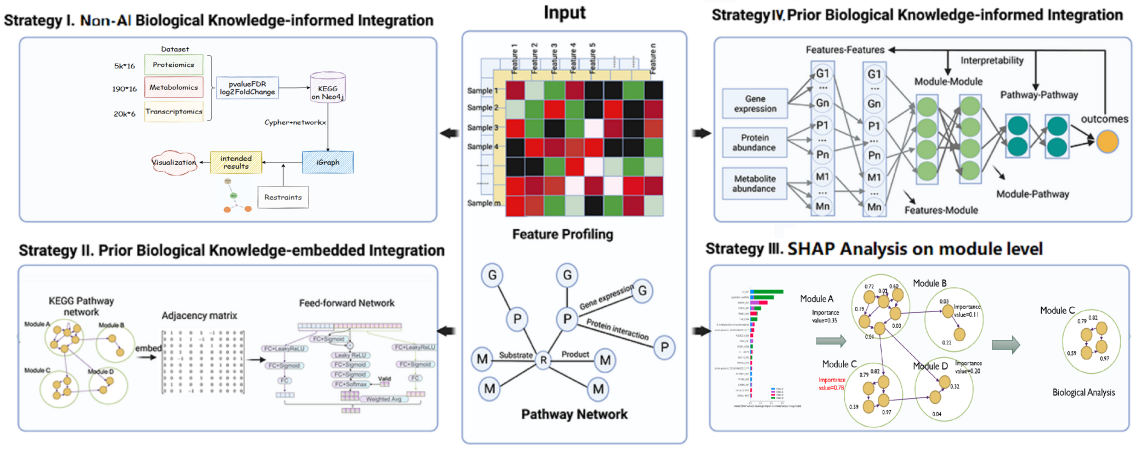

Figure 1: Overview of our project. Strategy I is non-AI method, and other strategies are all AI methods. All the methods are implemented by codes.

- COVID-19 dataset from [5], including proteomics and metabolomics information of healthy people and COVID-19 patients; and
- TNBC dataset from [8], including metabolomics and transcriptomics information of different subtypes of Triple Negative Breast Cancer.

However, all the datasets available are with some or all of the following defects:

- Lack of samples;
- Non-matching multi-omics data;
- Inexistence of many features in KEGG.

Based on different properties of datasets, we have developed variant analytical methods(See Figure 1).

In detail, ZIKV dataset is with only 6 samples in transcriptomics with 20 thousand features, 16 samples in proteomics with 5 thousand features and 16 samples in metabolomics with 190 features. Rich in number of features but lack of samples render us not to apply AI algorithms but do just traditional statistics analysis on KEGG and adds restraints by using biological knowledge(See Figure 1 Strategy I). The strategy is suitable for all kinds of datasets.

Another problem is that it is hard to collect dataset filled with **matched** samples [2, 5]. The "matched" sample means a single sample with corresponding multi-omics data. Almost all available datasets have a great proportion of unmatched part, and in our work, we designed a feed-forward network for partially unmatched samples, successfully utilizing the unmatched samples(See Figure 1 Strategy II). We also implemented the codes that include the adjacency matrix combined with a residual widget, but due to the inexistence of many features in KEGG, the adjacency matrix is too scarce to be effective while we have got some discoveries from the residual part.

However, it is hard to recognise the important modules (or pathways), and then we applied SHAP [9] analysis onto XGBoost [10] model, and by taking the curved average SHAP importance, we also analyzed on module level(See Figure 1 Strategy III). It is suitable for medium-size datasets with hundreds of samples.

Getting idea from [7], we also implemented a hierarchical feed-forward network encoding interactions between features, modules and pathways(See Figure 1 Strategy IV). The model is

estimated to be suitable for thousands of matched samples, but all three datasets are with less than 600 matched samples, resulting an inevitable overfitting circumstance.

2 Related Work

The notion "omics" means a comprehensive, or global, assessment of a set of molecules [2]. Nowadays, it is easy to get large quantities of data of one omics or multiple omics from separate sets of samples. With traditional methods of analyzing on statistics including log2FoldChange, pvalueFDR and variance, the field of omic analysis has witnessed many breakthroughs: [4] used univariate linear regression on log2FoldChange information to analyze COVID-19 samples; [11] used pvalueFDR(q-value) and log2FoldChange to figure out the metabolism of diabetes patients.

And there are also lots of successful works applying AI methods on datasets with one omics information. For example, [5] uses XGBoost and binary logistic regression classifiers to assist classification on plasma protein abundance of COVID-19 samples; [7] uses a feed-forward network to analysis on three levels of genomics; [6] uses evolutionary enhanced Markov clustering to apply on proteomics data of yeast. And in our project, two or more omics data are integrated to unearth interactions among different kinds of entities.

3 Methods

3.1 Non-AI Biological Knowledge-informed Integration

Based on the fact that there are only tens of samples in ZIKV dataset, we have implemented an interface for querying the KEGG database and further processing(See Figure 2). This method is universal for almost all types of datasets.

In detail, we first gathered data of all three omics and updated the log2FoldChange and pvalueFDR values on KEGG database. log2FoldChange is the logarithm of the ratio of the average concentration in experimental group(ZIKV samples) and in control group(CTRL samples), denoting the concentration difference. Since the concentration of many compounds in human bodies may change exponentially, we just take the logarithm for computational convenience. PvalueFDR is the adjusted pvalue($H_1 : \theta \neq \theta_0$) in hypothesis testing theory of statistics for controlling positive false discovery rate[12]. The pvalueFDR is defined by:

$$\text{pvalueFDR}_i = \text{pvalue}_i \times \frac{N}{k},$$

where N is the number of features and k is the rank of pvalue of feature i in increasing order. In our experiments, the concentration of some compound changes credibly significantly(increasing or decreasing) if pvalueFDR < 0.05 and |log2FoldChange| is greater than a threshold.

After getting the total graph by Cypher language and transforming it to iGraph[2] via networkX[3], we implemented with some elaborated restraints. We need to find the largest "accessible" network, which means a connected network with "accessible" edges except for some undetected intermediary nodes. "Accessible" edges includes 2 circumstances:

1)In reactions, only edges between triples in Table 1 are accessible. And in practice, if in some reaction the majority of triples are accessible, we include the accessible ones.

2)Between protein and protein/gene: if the relationship is activation/(gene) expression, the corresponding edge is accessible if both are increasing or decreasing; if the relationship is inhibition/repression, the corresponding edge is accessible if one increases and the other decreases.

[2]See `https://igraph.org/` for details.
[3]See `https://github.com/networkx/networkx` for details.

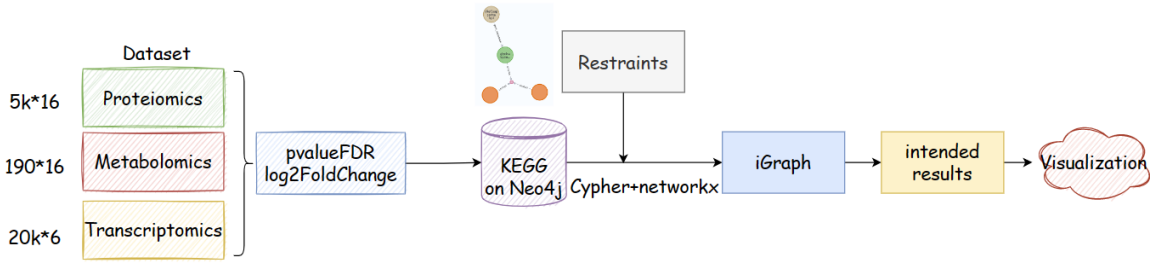

Figure 2: The diagram of Non-AI method. We used the pvalueFDR and log2FoldChange information of each feature and update them onto KEGG database. By using Cypher language and networkX, we can get the intended results, and we can also use iGraph for visualization.

Substrate	Enzyme	Product
No obvious change/decreasing	Decreasing	Decreasing
No obvious change/increasing	Increasing	Increasing
Increasing	No obvious change/None	Increasing
Decreasing	No obvious change/None	Decreasing

Table 1: The accessible triples in reactions.

For undetected intermediary nodes, we may predict the changing trend of it and only consider the second type of accessible edges. If over 2/3 of the detected neighbouring nodes agree with one trend(increasing/decreasing), we include edges connecting them with it. We found all accessible networks by DFS, and the designed algorithm shown in Algorithm 1.

Since some pairs of compounds may be related to too many reactions, we also design programs to remove these duplicate reaction nodes.

3.2 Further Analysis using Non-AI method

By previous method, we processed the ZIKV dataset and obtained a network that consists of relations between elements such as compounds, genes. The relation network can be visualize using iGraph on KEGG database.

We wanted to directly analyze the relationship network and tried to find some important information using non-AI method. The following are some proper methods we have tried.

The first method is to do pruning on the relation network. By adjusting the threshold of log2FoldChange, we obtained different connected graphs from the original relation network. Then we did farther pruning on these connected graphs. Some important parts of the relationship network may be found by comparing these graphs.

The second method we used is to study the relation network at module level. Module is a concept in multi-omics. KEGG database provides the information of module. We partitioned the relation network in the sense of module. Then we could study the importance of each module. For instance, we studied the importance of each module by the number of compounds in the module.

We then focused on the compounds in the important modules. There are reactions taken part between these compounds. We selected the reactions that involve compounds belonging to different modules. These reactions could be viewed as important information of the relationship network.

3.3 Feed-forward Network Analysis

COVID-19 dataset has 806 samples of proteomics data and 671 samples of metabolomics data from the healthy people as well as T1, T2 and T3 periods of COVID-19 patients[5]. And we designed the method of analyzing interactions through weights of feed-forward layer(See Figure 3, a).

Algorithm 1 Finding Accessible network

Get the transformed graph G.
Find all the accessible edges of detected nodes and undetected intermediate nodes.
Clear off all the nodes with no accessible edges.
Denotes the number of remaining nodes as $|V|$ and initialize $cnt = 0$, $networkList = []$ $avail = [True, ..., True]$ denoting for the availability of each node.
while $cnt < |V|$ **do**
  if not $avail[cnt]$ **then**
    $cnt \leftarrow cnt + 1$
  else
    Initialize $stack = [cnt]$, $network = []$ for a new accessible network
    while $len(stack) > 0$ **do**
      $v \leftarrow stack.pop()$ and append v to $network$.
      $avail[v] \leftarrow False$
      For edges $e = (v, u)$, if $avail[u] == True$ and u is not in stack, push u to the stack.
    end while
    Append $network$ to $networkList$.
  end if
end while
return $networkList$

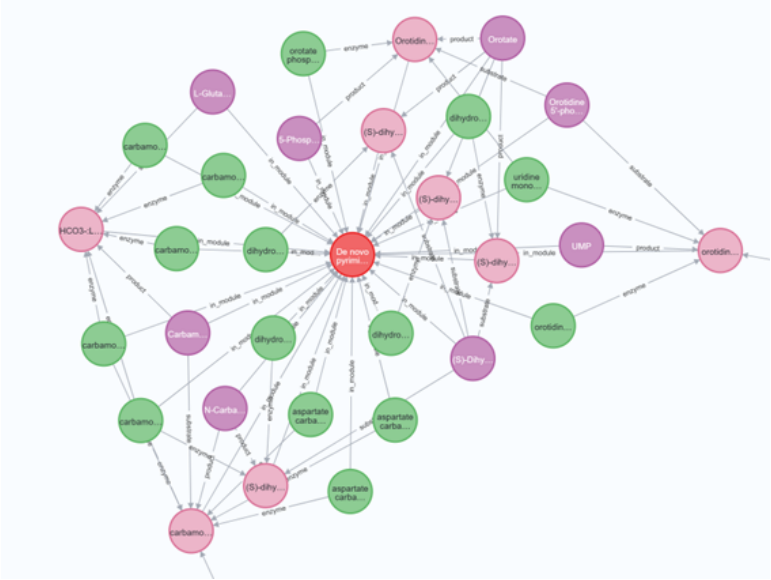

Figure 3: An example of module

However, the two sample groups are not completely coincident. About a half of samples are with only one omics information. For efficient utilization of the datasets, we creatively introduced input of "valid", indicating the availability of corresponding part of data. And similar to P-NET[7], we constructed two such networks for both omics. While for the multi-omics, to include the interaction between two omics, we creatively combined two feed-forward networks with a third one with a residual connection in the first layer, which can reflect this kind of relationship to some degree. The final result is a weighted average of two separate networks, and the third network calculated the weights by softmax, where we introduced an additional input of the "valid" parameter. After training the classification model, we analyzed the heat map of the residual layer for salient interactions.

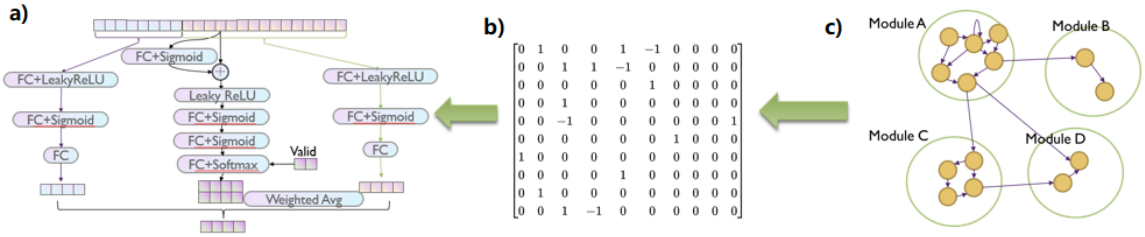

Figure 4: The structure of the feed-forward network analysis. a) the structure of feed-forward network; c) KEGG graph; b) adjacency matrix taken from KEGG graph as a "mask" of weights in the residual part.

We have also implemented an interface with the embedding of the adjacency matrix of KEGG in the feed-forward network as the prior knowledge. For positive interactions(activation/(gene) expression) or accessible reactions, we just set the elements 1 and for negative interactions(inhibition/ repression), we just set -1. And we used it to "mask" the weight matrix of the residual part, denoting the connections in prior knowledge. However, since the number of detected features in the dataset is much fewer than the number of nodes in KEGG, the resulting adjacency matrix is too scarce for training, and we did not train with the adjacency matrix.

3.4 XGBoost with SHAP analysis

XGBoost is a machine learning algorithm with forest-shape structure used for prediction (classification)[10], and SHAP tool can compute the SHAP values of a model to determine the important features for classification. We used the SHAP to analyze at entity, module and pathway levels.

The algorithm of XGBoost is in Algorithm 2.

Although there is also a function calculating feature importance in XGBoost, we used another analyzing tool called SHAP(SHapley Additive exPlanation)[9]. In the SHAP model, suppose the original model is f and the explanation model is g. We initialize a random set of $z' \in [0,1]^N$, where N is the number of features, and it tries to ensure $g(z') = \phi_0 + \sum_{j=1}^{N} \phi_j z'_j \approx f(h_x(z'))$ whenever $z' \approx x'$ and $x = h_x(x')$ be a mapping function. Then we get

$$\pi_x(z') = \frac{(N-1)}{\binom{N}{|z'|}|z'|(N-|z'|)}$$

, and we need to make linear regression to

$$g(x') = \phi_0 + \sum_{j=1}^{N} \phi_j$$

Algorithm 2 XGBoost[10]

Require: training dataset $\{(x_i, y_i)\}_{i=1}^N$, loss function $L(y, F(x))$, number of weak learners M, learning rate r.

1: Initialize model with a constant value: $\hat{f}_{(0)}(x) = \arg\min_\theta \sum_{i=1}^N L(y_i, \theta)$.
2: **for all** $m = 1, ..., M$: **do**
3: Compute the gradient matrix and Hessian matrix: $\hat{g}_m(x_i) = [\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}]_{f(x)=\hat{f}_{(m-1)}(x)}$, $\hat{h}_m(x_i) = [\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}]_{f(x)=\hat{f}_{(m-1)}(x)}$.
4: Fit a learner using the training set $\{x_i, -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\}_{i=1}^N$ by solving the optimization problem: $\hat{\phi}_m = \arg\min_\phi \sum_{i=1}^N \frac{\hat{h}_m(x_i)}{2}[-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \phi(x_i)]^2$, $\hat{f}_m(x) = r\hat{\phi}_m(x)$
5: Update the model by $\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \hat{f}_m(x)$.
6: **end for**
7: **return** $\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^M \hat{f}_m(x)$.

by decreasing the loss of $L(f, g, \pi_x) = \sum_{z' \in Z}[f(h_x(z')) - g(z')]^2 \pi_x(z')$. Then ϕ_0 is the mean of $f(x)$ and ϕ_i is the SHAP value for feature i. And finally we can get

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{N} \phi_i x_i'. \tag{1}$$

In Equation (1), the SHAP value ϕ_i is just like the contribution of feature i, hence we can use SHAP value to analyze the importance of features. Moreover, we also conducted researches at module and pathway levels according to average SHAP value of detected entities. We ignored the pathways and modules with less than 3 detected entities, and got the ones with top 20 average SHAP values for further analysis.

3.5 Improvement on P-NET

Based on P-NET[7], we add additional layers to reflect the interaction between units of different levels(See Figure 5). In the model, between layers of neurons is a tanh activation after a full connected layer masked by actual connections in KEGG. And there is also a sigmoid output branch from each full connected layer representing the information from all levels of the hierarchy. And the outcome is the average of the branches in each layer. However, it is very likely to overfit with matched samples fewer than 1,000, so we expect a better dataset.

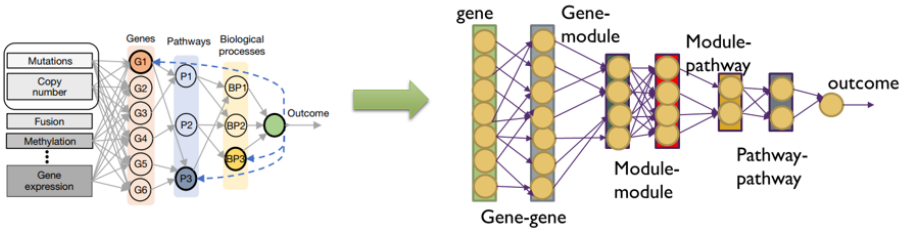

Figure 5: The structure of P-NET(left) and our improved model(right).

4 Experiments

4.1 Non-AI method

We have tried different thresholds of log2FoldChange from 0.0 to 1.5, and added different kinds of restraints to get different results shown in Figure 6. It is shown that there is a sharp decline

between threshold = 0.5 and 0.6, indicating an adequate value of threshold. Then we analyzed the results when threshold = 0.6 by strict restraints and the largest accessible metabolic network is shown in Figure 7. It is clear enough and ready to be further biological analysis.

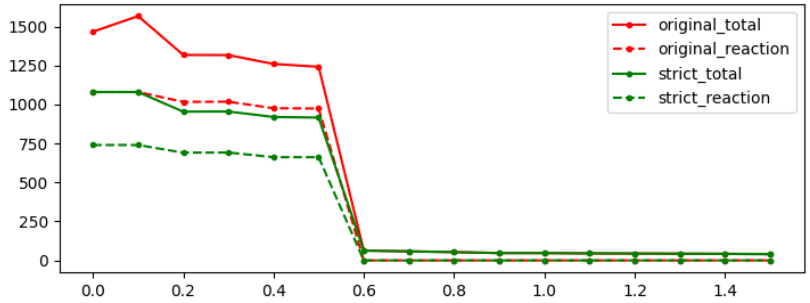

Figure 6: The numbers of total and reaction nodes in the largest accessible networks of different thresholds of log2FoldChange and different restraints. The strict restraints are to remove reaction nodes with only one detected neightbours.

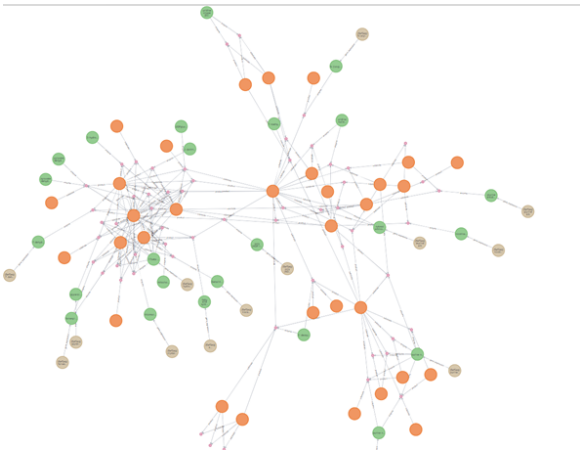

Figure 7: The largest accessible metabolic network for strict restraints when log2FoldChange threshold=0.6. It is clear and ready for further biological analysis.

4.2 Feed-forward Network Method for COVID-19 Dataset

We have trained the classification model on COVID-19 dataset by using information of only proteomics, only metabolomics and both omics. And the validation and test accuracies are shown in Table 2.

Split	Proteomics	Metabolomics	Both
Validation	78.4%	62.8%	77.3%
test	79.0%	71.8%	74.2%

Table 2: The validation and test accuracies of the three experiments.

With the accuracies near 75%, we then got the residual weight matrix and drew a heat map(Figure 8). It is found that there are obvious connections between IFNG and some other compounds, showing a potential research direction.

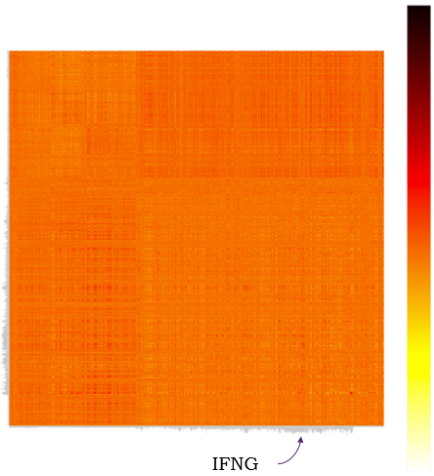

Figure 8: The heat map of the residual weight matrix. The weights of IFNG to some other compounds are the most salient, showing a potential research direction.

4.3 XGBoost with SHAP Analysis

We conducted SHAP analysis on the XGBoost model trained on COVID-19 and TNBC datasets. On COVID-19 dataset, the test accuracy is 78.5% and the SHAP analysis result is shown in Figure 9. We can see that the LIF protein and cysteine sulfate are the most significant indicators of the healthy status of samples, while the DDX58 is the most important indicator for infection of COVID-19, which can be potential for a further medical analysis.

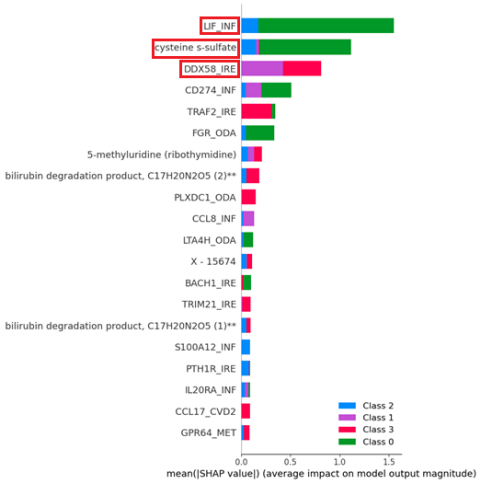

Figure 9: The bar chart for mean absolute SHAP importance. Classes 0, 1, 2 and 3 indicate healthy group, T1, T2 and T3 periods of COVID-19 patients, respectively.

We also conducted researches on TNBC dataset. The trained XGBoost model has an accuracy of 77.6% on test dataset. On the level of entity, we have discovered that ENTPD1 is the most salient. It encodes CD39, a popular research objective of immunotherapy, which can hydrolyze extracellular ATP into AMP. The expression level of CD39 has been increased among various tumors. Our discoveries are consistent with the experiments of [1]., it is mentioned that CD39 and CD73 are expressed in TNBC cell lines, and experiment results show that CD39 is the main enzyme responsible for ATP hydrolysis in TNBC. It is expected to do more biological experiments to validate other results of this method and reveal more mechanisms.

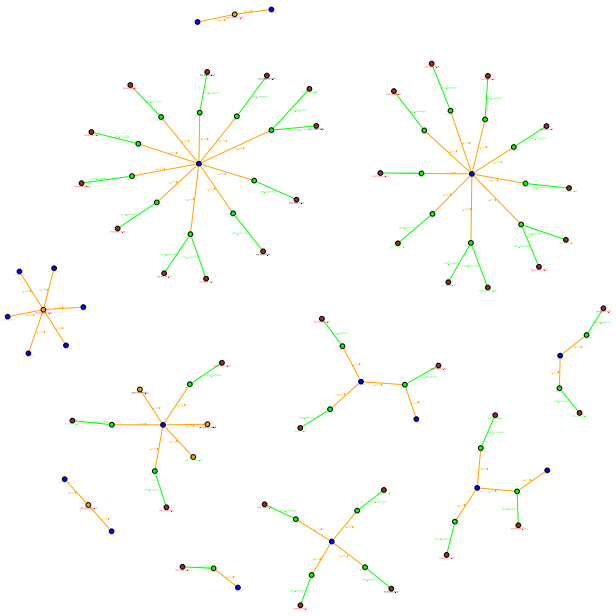

Figure 10: Visualization of the 20 most important modules with highest SHAP importances. Blue, green, red and orange nodes are modules, orthologs, genes and compounds, respectively.

On the level of pathway, we have visualized the 20 most important modules with highest SHAP importances, shown in Figure 10. We have found that metabolic pathways related to Tryptophan and C5 isoprene are most potential indicators of subtypes of TNBC.

5 Discussion & Conclusion

Some of our methods are just implemented with code but not applied because we are short of datasets with many matched samples. And in our applied models, due to the few number of samples, the accuracies are not so high and the validation and test losses are still high, showing a sign of overfitting. On one hand, we should improve the structure of our model for preventing overfitting; on the other hand, we should apply the model with datasets of more samples for a better result. Hopefully, currently Hu-Lab is trying to collect more comprehensive and consistent datasets; by using the methods we have developed, we can anticipate further potential medical discoveries in the future.

Moreover, we have discovered some potential research objectives including LIF, DDX58, IFNG, etc. However, these findings should be tested in biological experiments which would take a long period of time.

All in all, we have implemented 1 rule-based and 3 AI methods for 3 datasets, from the generally suitable strategy to strategy suitable for high-throughput high-quality datasets. We have found ENTPD1 for TNBC which is validated by biological experiments, proving the efficiency of our methods to some degree. And we have also discovered some potential research directions for further analysis.

References

[1] Laura Schäkel, Salahuddin Mirza, Riekje Winzer, Vittoria Lopez, Riham Idris, Haneen Al-Hroub, Julie Pelletier, Jean Sévigny, Eva Tolosa, and Christa E Müller. Protein kinase in-

hibitor ceritinib blocks ectonucleotidase cd39 – a promising target for cancer immunotherapy. *Journal for ImmunoTherapy of Cancer*, 10(8), 2022.

[2] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):1–15, 2017.

[3] Ali Ebrahim, Elizabeth Brunk, Justin Tan, Edward J O'brien, Donghyuk Kim, Richard Szubin, Joshua A Lerman, Anna Lechner, Anand Sastry, Aarash Bordbar, et al. Multi-omic data integration enables discovery of hidden biological regularities. *Nature communications*, 7(1):1–9, 2016.

[4] Katherine A Overmyer, Evgenia Shishkova, Ian J Miller, Joseph Balnis, Matthew N Bernstein, Trenton M Peters-Clarke, Jesse G Meyer, Qiuwen Quan, Laura K Muehlbauer, Edna A Trujillo, et al. Large-scale multi-omic analysis of covid-19 severity. *Cell systems*, 12(1):23–40, 2021.

[5] Yapeng Su, Dan Yuan, Daniel G Chen, Rachel H Ng, Kai Wang, Jongchan Choi, Sarah Li, Sunga Hong, Rongyu Zhang, Jingyi Xie, et al. Multiple early factors anticipate post-acute covid-19 sequelae. *Cell*, 185(5):881–895, 2022.

[6] Konstantinos Theofilatos, Niki Pavlopoulou, Christoforos Papasavvas, Spiros Likothanassis, Christos Dimitrakopoulos, Efstratios Georgopoulos, Charalampos Moschopoulos, and Seferina Mavroudi. Predicting protein complexes from weighted protein–protein interaction graphs with a novel unsupervised methodology: evolutionary enhanced markov clustering. *Artificial intelligence in medicine*, 63(3):181–189, 2015.

[7] Haitham A Elmarakeby, Justin Hwang, Rand Arafeh, Jett Crowdis, Sydney Gang, David Liu, Saud H AlDubayan, Keyan Salari, Steven Kregel, Camden Richter, et al. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880):348–352, 2021.

[8] Yi Xiao, Ding Ma, Yun-Song Yang, Fan Yang, Jia-Han Ding, Yue Gong, Lin Jiang, Li-Ping Ge, Song-Yang Wu, Qiang Yu, et al. Comprehensive metabolomics expands precision medicine for triple-negative breast cancer. *Cell research*, 32(5):477–490, 2022.

[9] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[10] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016.

[11] Karsten Suhre, Christa Meisinger, Angela Döring, Elisabeth Altmaier, Petra Belcredi, Christian Gieger, David Chang, Michael V Milburn, Walter E Gall, Klaus M Weinberger, et al. Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. *PloS one*, 5(11):e13953, 2010.

[12] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Author Contributions

The term "we" in the main part is just for formality. In fact, all the main tasks are done by Liu Yifeng except for Section 3.2 by Pan Yunzhou & Tian Yuchen. We also thank Prof. Yuan Yang and Prof. Hu Zeping for instruction, and thank Zhao Yizi and Li Biao for construction of KEGG database and provision of some ideas of these methods.