HELMHOLTZ
ASSOCIATION
Alliance on Systems Biology

HelmholtzZentrum münchen
German Research Center for Environmental Health

TECHNISCHE
UNIVERSITÄT
MÜNCHEN

Novel network-based methods for multi-omics data analysis and interpretation

Daniel Parviz Gomari

February 2021

TECHNISCHE UNIVERSITÄT MÜNCHEN

Novel network-based methods for multi-omics data analysis and interpretation

Daniel Parviz Gomari

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Prof. Dr. Dmitrij Frischmann

Prüfer der Dissertation:

1. TUM Junior Fellow Jan Krumsiek
2. Prof. Karsten Suhre, Ph.D.

Die Dissertation wurde am 15.02.2021 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 16.04.2021 angenommen

Acknowledgements

I want to take this opportunity to thank the people whose support have made this PhD possible.

First off, thanks to Jan Krumsiek for all the fruitful discussions, invaluable feedback, and endless support throughout the four years of my PhD, despite the long distance from Munich to New York City, and creating a collaborative and fun work atmosphere in the group.

I am thankful to my collaboration partners, in particular Iman Achkar and Anna Halama, for the great cooperation on projects.

To my colleagues Elisa, Mustafa, and Annalise and former colleague Alida, I am grateful for all the great discussions, the breaks, and the times we spent in and outside of the office. My PhD time would not have been nearly as cool and fun without you all! I want to especially thank Elisa for always being up for discussions, her helpfulness, and honest advice whenever I needed it.

To all my friends, in particular Umberto and Pedro, thank you for all the emotional support in the past years, especially during challenging times.

I want to thank my brothers Syrus and Peiman, Julie, and my father for all the moral support throughout these years. I am also deeply grateful to my brother-in-law Stefan and my sister Laleh for helping me reach where I am today.

Lastly, I want to thank my late mother who inspired me to pursue my PhD and who taught me the value of always keeping an open mind moving forward.

Abstract

The latest advancements in 'omics' technologies have enabled the high-throughput generation of various types of molecular data. For instance, these advancements have made it possible to sequence whole genomes, analyze global transcript levels, and quantify proteomes and metabolomes, all of which are now regularly incorporated into biological study designs. Despite this progress, the understanding of the complex statistical findings produced by multi-omics data analysis poses significant challenges.

A widely used resource in the interpretation of multi-omics data are pathway and molecular interaction network databases, which are developed from comprehensive literature curation and *in silico* approaches. These databases are used in pathway and network analysis to summarize large lists of molecules from experimental data into smaller lists of predefined biological pathways and interactions. Another common data summarization approach are data-driven methods where molecular interactions are statistically inferred. Among these approaches are linear dimensionality reduction methods, such as principal component analysis (PCA). These untangle high-dimensional datasets into underlying core biological processes by finding smaller groups of related molecules. However, despite the popularity of the aforementioned techniques, none of them fully utilize the information contained in molecular interactions, which are essential in understanding complex systems-level characteristics such as diseases. For instance, knowledge database-driven methods usually do not use detailed information of single interactions in the analysis of high-throughput data. Similarly, linear dimensionality reduction methods, by definition, disregard nonlinear interactions that could be present in datasets.

Addressing these issues in high-throughput biological data analysis in this doctoral thesis, I focused on developing and evaluating new network-based methods for the

analysis and interpretation of multi-omics data, with a special focus on metabolomics data. I developed a tool that uses individual molecular interactions, which enables the generation of detailed molecular insights from multi-omics data. Furthermore, I implemented and evaluated variational autoencoders, a nonlinear dimensionality reduction method that utilizes nonlinear interactions in generating a better data-driven understanding of the metabolome.

First, I developed “piTracer”, an R Shiny application that enables the rapid and automatic reconstruction of molecular cascades at the multi-omics level. For instance, with inputs glucose as a substrate and pyruvate as a product, piTracer can automatically and accurately reconstruct the glycolysis pathway. I started by constructing a novel directed multi-omics network consisting of gene-gene, gene-metabolite, and metabolite-metabolite interactions. I developed a novel algorithm based on atom-tracing to construct a biochemically valid, directed metabolic network from a database of metabolic reactions. I developed this algorithm, since unprocessed metabolite interaction networks contain cofactors that act as shortcuts that connect all metabolites, are difficult to remove, and make it difficult to reconstruct metabolic cascades. After constructing my metabolic network, I then created a gene interaction network based on gene interactions found in public databases. I subsequently combined the gene interaction and metabolic networks to assemble the multi-omics network.

Second, with my directed multi-omics network as the backend, I used a k-shortest path algorithm to enable the reconstruction of multiple paths between pairs of molecules, i.e. genes and/or metabolites, in the network. It is pivotal to find several paths between molecule pairs, since in biological systems, molecules are connected through multiple molecular cascades of varying lengths. To enable the grouping of these paths into biologically similar mechanisms for easier interpretation, I implemented a pathway clustering algorithm. I then compared piTracer to state-of-the-art tools and showed that it significantly outperforms these tools in all biological pathway reconstructions.

Third, I applied piTracer to prioritize druggable genes to enable the streamlining of drug screening. Using breast cancer cell line metabolomics data as an input, I reconstructed tumor escape mechanisms with piTracer. With an additional drug target algorithm that I developed, I then assigned a druggability score to over 7,000 genes, prioritized 30 genes based on these scores, and selected 4 genes to experimentally validate. Through a successful validation experiment of our predictions, I showed that piTracer correctly identified and prioritized genes essential for tumor survival. This has big implications for metabolomics-based drug screening and repurposing efforts. For instance, the first crucial step in drug screening and repurposing studies is finding a viable target for a specific indication of interest, which is time intensive, expensive, and requires extensive domain expertise. piTracer can significantly expedite this process by automatically and rapidly selecting candidate genes to target with drugs.

Lastly, for the first time, I demonstrated the utility of Variational Autoencoders (VAEs) for metabolomics data. I trained a VAE on a large-scale metabolomics population cohort of human blood samples consisting of over 4,500 individuals. I analyzed the pathway composition of the latent space using a global feature importance score, which showed that latent dimensions represent distinct cellular mechanisms. In a validation step, I found that latent representations significantly correlated with patient groups in unseen metabolomics datasets on type 2 diabetes, schizophrenia, and acute myeloid leukemia patient groups, significantly outperforming our PCA baseline. This implies that, leveraging nonlinearities in metabolomics data, these representations capture disease-associated biological mechanisms. My findings suggest that VAEs are a powerful method that learns biologically meaningful, nonlinear and universal latent representations of metabolomics data.

Taken together, I created piTracer, which outperforms current methods in automatically and accurately reconstructing biological mechanisms. Interestingly, using only metabolomics data, piTracer is able to predict essential genes in cancer cell lines, which could massively expedite drug screening and repurposing efforts. In addition, I demonstrated that VAEs can learn universal representations of biological processes from a large-scale population cohort. This application is especially important, given the arrival of big datasets consisting of more than 500,000 individuals in the near future. In this doctoral work, I developed and applied novel network-based methods, which could substantially impact the way high-throughput, and especially metabolomics data will be utilized in future studies.

List of contributed articles

The following list shows manuscripts in submission or revision relevant for each chapter of this thesis.

Chapters 2, 3, and 4

- Gomari, D. P., Tabeling, J., Achkar, I., Halama, A., Krumsiek, J. piTracer. *Manuscript in preparation.*

Chapter 5

- Gomari, D. P., Schweickart, A., Cerchietti, L., Paietta, E., Fernandez, H., Al-Amin, H., Suhre, K., Krumsiek, J. Variational autoencoders learn universal latent representations of metabolomics data. *In review.*

Table of Contents

Chapter 1: Introduction

1.1 High-throughput “omics”

Recent developments in high-throughput “omics” technologies have paved the way to generate various levels of molecular data in a high-throughput manner [1]. For example, these technologies have made it possible to sequence whole genomes, to examine global transcript levels, and to measure the proteome and metabolome, all of which are now routinely incorporated into everyday biological study designs [1]. This has given rise to the field of “systems genetics”, where intermediate molecular phenotypes, such as transcript, protein, and metabolite abundances, that bridge the gap between DNA variation and phenotypic traits, are examined at a systematic level [2, 3]. Moreover, high-throughput omics measurements have allowed for the development of “systems medicine” approaches, in which multidisciplinary investigator teams integrate and validate multi-omics data for a better understanding of human disease for the benefit of patients [4]. For example, these approaches have found that escape mechanisms are an attractive drug target in cancer, because tumor cells are dependent on them for survival. Many escape pathways have metabolic adaptations as their foundation [5]. Thus, measuring the metabolome is crucial in elucidating these compensatory mechanisms and have a big potential in finding druggable genes involved in cancer survival. Given the proximity of the metabolome to such disease phenotypes, in this thesis, I will primarily focus on metabolomics data.

Systems-level computational approaches applied to high-throughput multi-omics data help us in profiling the molecular phenotype of disease and identifying the underlying pathological mechanisms of action [6–9]. Extracting such systemic effects from high-dimensional datasets requires data summarization and dimensionality reduction approaches to disentangle the high number of molecules into the processes in which they participate.

1.2 Common methods for high-throughput multi-omics data analysis

1.2.1 Pathway analysis and network-based methods

Pathway databases, constructed from extensive literature curation and *in silico* approaches, are commonly used to identify biologically interpretable mechanisms from the plethora of information contained in multi-omics datasets. Prominent databases widely used in the field include, among others, KEGG [10], the Gene Ontology (GO) [11], SMPDB [12], Recon 3.0 [13], STRING [14], OmniPath [15], and Reactome [16]. These databases cover metabolic reactions, protein-protein interactions, gene-regulatory interactions and other molecular relations between molecules.

With these databases, pathway analysis helps in summarizing large molecule lists from experimental data into smaller lists of predefined biological pathways [17]. Pathways are statistically tested for whether they accumulate significantly altered molecules relative to what is expected by chance. Common pathway analysis tools are Enrichr [18], SAFE [19], GAGE [20], and GSEA [21]. Furthermore, topology-based methods aim to enrich multi-omics datasets with mechanistic network information, enabling a better understanding of the underlying biological processes

(Figure 1.1a). These methods use graph algorithms on interaction networks, where nodes represent biomolecules and edges signify the interactions between these molecules. Similar to pathway enrichment-based methods, these approaches summarize high-throughput data. For instance, they identify network modules that are concomitantly affected in a biological system under study. Examples of these methods are module identification tools such as Hetionet [22] and Hierarchical HotNet [23] and specialized methods such as flux balance analysis (FBA) [24]. The aforementioned methods focus mainly on the usage of pathway and network databases on system-wide analyses of omics data.

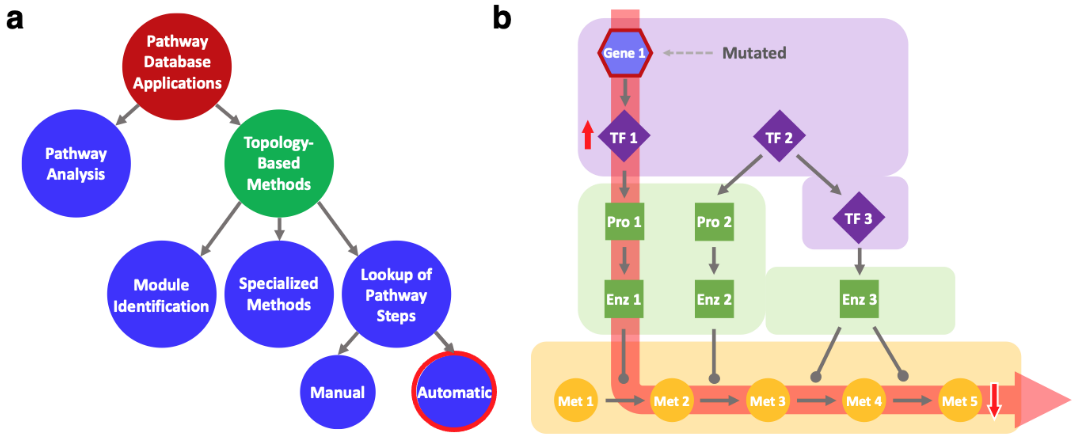


Figure 1.1. (**a**) Selected applications of pathway databases. Applications can be classified into either pathway analysis or topology-based methods, which include module identification, specialized methods (such as flux balance analysis), and manual or automatic lookup of pathway steps. (**b**) Example of a "trace" between a transcription factor gene, Gene 1, and its downstream effect on a metabolite, Met 5, in a metabolic pathway. TF: transcription factor, Pro: protein, Enz: enzyme, Met: metabolite.

Currently, detail-oriented analyses of experimental data by manually constructing molecular cascades between statistically significant molecules is an arduous task when molecules are far apart in an interaction network. For example, suppose that a mutated gene causes an increase in expression of a transcription factor and statistically associates with a decrease in a certain metabolite (Figure 1.1b). Given this observation, we may ask which molecular cascade confers the effect of the mutated gene onto the affected metabolite. However, finding such a biological path is challenging due to the colossal number of unique paths that connect molecules in an interaction network that can be chosen from. The problem with this type of analysis is further exacerbated if there are misleading shortcuts, such as metabolic cofactors or hub genes, in the interaction network. These shortcuts make it difficult to decide whether the generated "traces", i.e. reconstructed paths, are biologically meaningful. If it were possible to accurately reconstruct biological cascades, it would enable the detailed interpretation of results derived from statistical analysis. Moreover, it would be possible to rapidly generate hypotheses that can be integrated into experimental designs. For instance, an attractive application would be finding diseases-relevant molecular cascades in drug screening studies. Therefore, to create a tool that automatically finds biologically relevant molecular cascades, it is of paramount importance to exclude misleading shortcuts during the constructions of a molecular interaction network. Currently, existing pathway and network-based tools often overlook this important step due to the difficulty in removing shortcuts [25, 26], limiting their ability to fully utilize molecular interactions in multi-omics data analysis.

1.2.2 Dimensionality reduction methods and Autoencoders

Another common approach for generating insights from high-dimensional datasets are dimensionality reduction methods, which are widely used in the field of genomics, proteomics, metabolomics, and others. For example, these enable the disentanglement of the high number of metabolites into processes in which they participate. Prominently, linear dimensionality reduction methods, such as principal component analysis (PCA) [27] and independent component analysis [28], have been extensively applied to high-dimensional biological data, for instance to discover inherent metabolic processes in metabolomics data [29–33] or to deconvolute gene expression data into cell-type specific processes, drug response modules, and oncogenic regulatory pathways. [34–37]. However, metabolic systems, like most complex biological processes, contain nonlinear effects which arise due to high-order enzyme kinetics and upstream gene regulatory processes [38, 39]. The usage of metabolite ratios is a successful and intuitive example of the exploitation of such nonlinear effects in metabolomics data, approximating the steady state between products and educts of metabolic reactions [40, 41]. Extending this concept, systematic methods that take nonlinearities into account are required to correctly recover the functional interactions between metabolites in an unbiased fashion.

Autoencoders (AEs), which belong to the field of deep learning, were developed as a method of dimensionality reduction that can capture nonlinear effects [42]. AEs reduce high-dimensional data into latent variables through an encoding/decoding process which recreates the input data after passing through a lower dimensional space. Once the model is fitted, the latent variables represent a compact, often easier-to-interpret version of the original data. While AEs have been successful for prediction tasks on biological datasets [43, 44], they tend to learn a latent space specifically fitted to the input dataset, and are therefore not generalizable to unseen data. The Variational Autoencoder (VAE) was introduced as an extension to the AE architecture that uses variational inference to generate a probabilistic posterior distribution of latent embeddings [45, 46]. With this extension, the VAE not only reconstructs the input data, but infers the generative process behind the data, leading to high generalizability across datasets. The VAE architecture has, for example, proven effective for predicting single cell-level response to infection in transcriptomic data not available during training, and predicting drug response from gene expression data where drug response information is sparse [47, 48].

The application of deep learning architectures, including VAEs, to metabolomics datasets has significantly lagged behind all other omics [49]. As such, current state-of-the-art dimensionality reduction methods used for metabolomics data rely overwhelmingly on linear assumptions, and are therefore not able to pick up on possible nonlinearities [38] that are a result of functional interactions between metabolites.

1.3 Research goals

Despite fast-paced developments in multi-omics data analysis, none of the aforementioned approaches incorporate comprehensive information contained at the level of single interactions in the summarization of high-throughput data. Moreover, during analysis, these methods do not include nonlinear interactions that could be present in datasets.

In this thesis I will create and implement network-based methods that use individual molecular interactions and utilize nonlinear interactions present in data, with the goal to:

(I) Construct a multi-omics interaction network devoid of misleading shortcuts from various and heterogeneous pathway and network databases to enable the automatic assembly of true biologically cascades.

(II) Create an easy-to-use and open-source tool that enables the rapid reconstruction of biological pathways, using the interaction network from (I). By reconstructing molecular cascades between statistically relevant molecules, such a tool would allow for a detail-oriented analysis of multi-omics experimental results, which is currently a difficult undertaking.

(III) Apply the tool developed in (II) in predicting tumor-relevant druggable genes from reconstructed cancer escape mechanisms and validate these predictions in a drug experiment targeting these genes. This would demonstrate the hypothesis generation capability of the tool. In addition, a positive validation result would indicate the ability of the tool to streamline drug screening and repurposing efforts.

(IV) Learn a universal representation of metabolomics data through VAEs, which excel at utilizing nonlinearities present in data, that are a result of the interactions that exist between metabolites. By incorporating data nonlinearities, VAEs are expected to outperform linear-based methods, such as PCA, in learning lower-dimensional representations of metabolomics data.

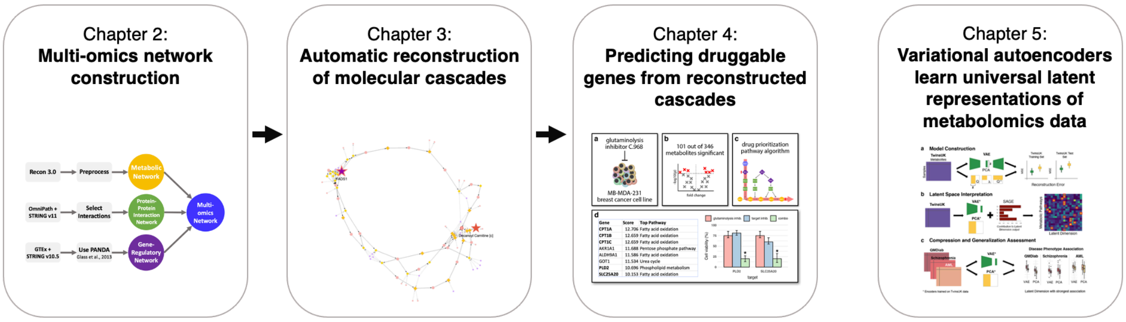

Chapter 2:
Multi-omics network construction
Recon 3.0
Preprocess
Metabolic Network
OmniPath + STRING v11
Select Interactions
Protein-Protein Interaction Network
GTEx + STRING v10.5
Use PANDA
Glass et al., 2013
Gene-Regulatory Network
Multi-omics Network
Chapter 3:
Automatic reconstruction of molecular cascades
Chapter 4:
Predicting druggable genes from reconstructed cascades
a
glutaminolysis inhibitor C.968
MB-MDA-231 breast cancer cell line
b
101 out of 346 metabolites significant
c
drug prioritization pathway algorithm
d
Chapter 5:
Variational autoencoders learn universal latent representations of metabolomics data
a Model Construction
b Latent Space Interpretation
c Compression and Generalization Assessment

Figure 1.2. Overview of the thesis. I introduce piTracer in Chapters 2 to 4. In Chapter 2, I describe how I constructed a multi-omics network used as the backend of piTracer. Then, in Chapter 3, I demonstrate that piTracer can accurately and rapidly reconstruct true biological cascades. I then apply piTracer to predict druggable genes and validate these predictions experimentally in Chapter 4. Finally, in Chapter 5, I demonstrate that variational autoencoders (VAEs) are able to learn universal latent representations of metabolomics data.

1.4 Overview of this thesis

The following is a brief outline of this thesis. A graphical overview is given in Figure 1.2.

In **Chapter 2**, I outline the steps in the construction of our directed multi-omics network, which forms the basis of our piTracer app (described in **Chapter 3**). To create my metabolic network, I develop an algorithm based on atom-tracing and apply it to metabolic reactions from the Recon 3.0 database. This step was necessary in order to remove cofactors, which connect all metabolites in unprocessed interaction networks that lead to false cascade reconstructions. Subsequently, I assemble a gene interaction network using publicly available databases, such as STRING, OmniPath, and Genotype-Tissue Expression (GTEx). I then combined the metabolic and gene interaction networks to create a directed multi-omics network.

Chapter 3 presents “piTracer”, an R Shiny application that enables the automatic and accurate reconstruction of biological cascades. For instance, given glucose as a starting molecule (substrate) and pyruvate as an end molecule (product), piTracer can reconstruct the glycolysis pathway. I implement Yen’s k-shortest path algorithm to construct paths between molecule pairs, i.e. genes and/or metabolites, in my multi-omics network. Subsequently, to enable the clustering of k-shortest paths between two molecules into biologically similar or redundant mechanisms, a path clustering algorithm is added to piTracer. I then compare the performance of piTracer with state-of-the-art pathway reconstruction tools, such as the QIAGEN’s Ingenuity Pathway Analysis (IPA) and ConsensusPathDB. The comparison demonstrates that piTracer significantly outperforms both tools in the accurate reconstruction of biological pathways.

In **Chapter 4**, the application of piTracer in accurately predicting essential disease genes is presented. With piTracer, I first reconstruct escape mechanisms of a breast cancer cell line using metabolomics data acquired before and after treatment of a glutaminase inhibitor. Genes involved in survival mechanisms are then ranked with a novel scoring method based on piTracer reconstructions. I then selected and experimentally validated our predictions in a drug combination treatment experiment, which showed that targeting our selected genes substantially decreased breast cancer cell viability. The results demonstrate that piTracer can significantly expedite drug screening and repurposing efforts by accurately prioritizing disease-essential genes.

For the first time in the metabolomics field, **Chapter 5** presents the application of Variational Autoencoders (VAEs) on metabolomics data. I train a VAE on a dataset of 217 measured metabolites in 4,644 twins of European-ancestery (4,256 females, 388 males) plasma samples from the TwinsUK population cohort. Furthermore, I interpret the VAE learned representation by calculating Shapley Additive Global importancE (SAGE) values, which represent global feature importance scores, at the level of metabolites and metabolic pathways. SAGE values demonstrate that the learned representations capture distinct cellular mechanisms. Moreover, compared to our PCA baseline, I show that VAE representations associate significantly with patient groups from unseen and very different datasets. That is, data from cohorts of mixed-gender and multi-ethnic from the US and Qatar, previously unseen by our VAE model which was trained on a mono-ethnic and predominantly-female population cohort. Our results demonstrate the ability of VAEs to learn universal representations of biological processes and could potentially replace linear dimensionality reduction methods in the metabolomics field.

Finally, in **Chapter 6**, I enumerate the scientific contributions of this thesis to the high-throughput data analysis field and discuss possible extensions and potential future projects.

Chapter 2: Multi-omics network construction

An essential prerequisite for reconstructing biological cascades is having a multi-omics interaction network composed of directed interactions and devoid of shortcut paths that arise from metabolic cofactors and hub genes. To this end, we constructed our multi-omics network for tracing (Figure 2.1), by first importing the human version of several public databases, focusing on those that are highly curated and that we considered to be the most advanced in the field. We then processed these interaction networks to exclude as many false positive molecular interactions, by removing cofactors and selecting gene interactions with high confidence scores. Lastly, we combined these heterogeneous networks by connecting them through overlapping molecules, such as enzymes, to assemble our final multi-omics network (Figure 2.1).

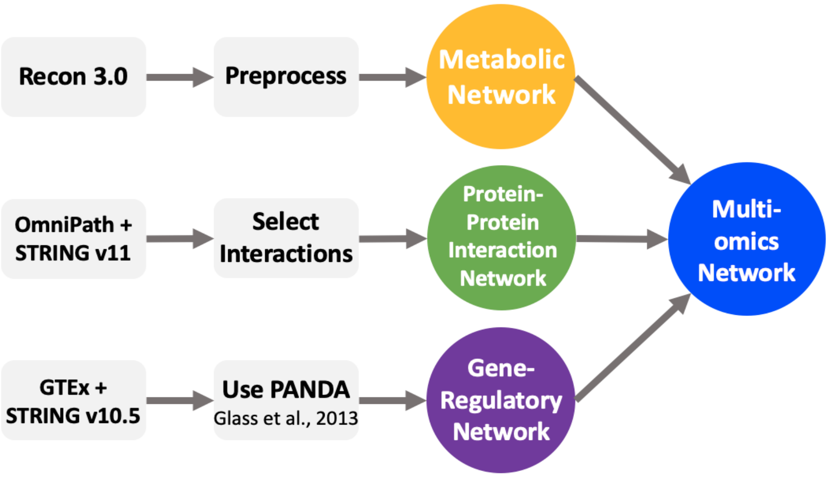


Figure 2.1. The databases and steps used in the construction of our multi-omics network.

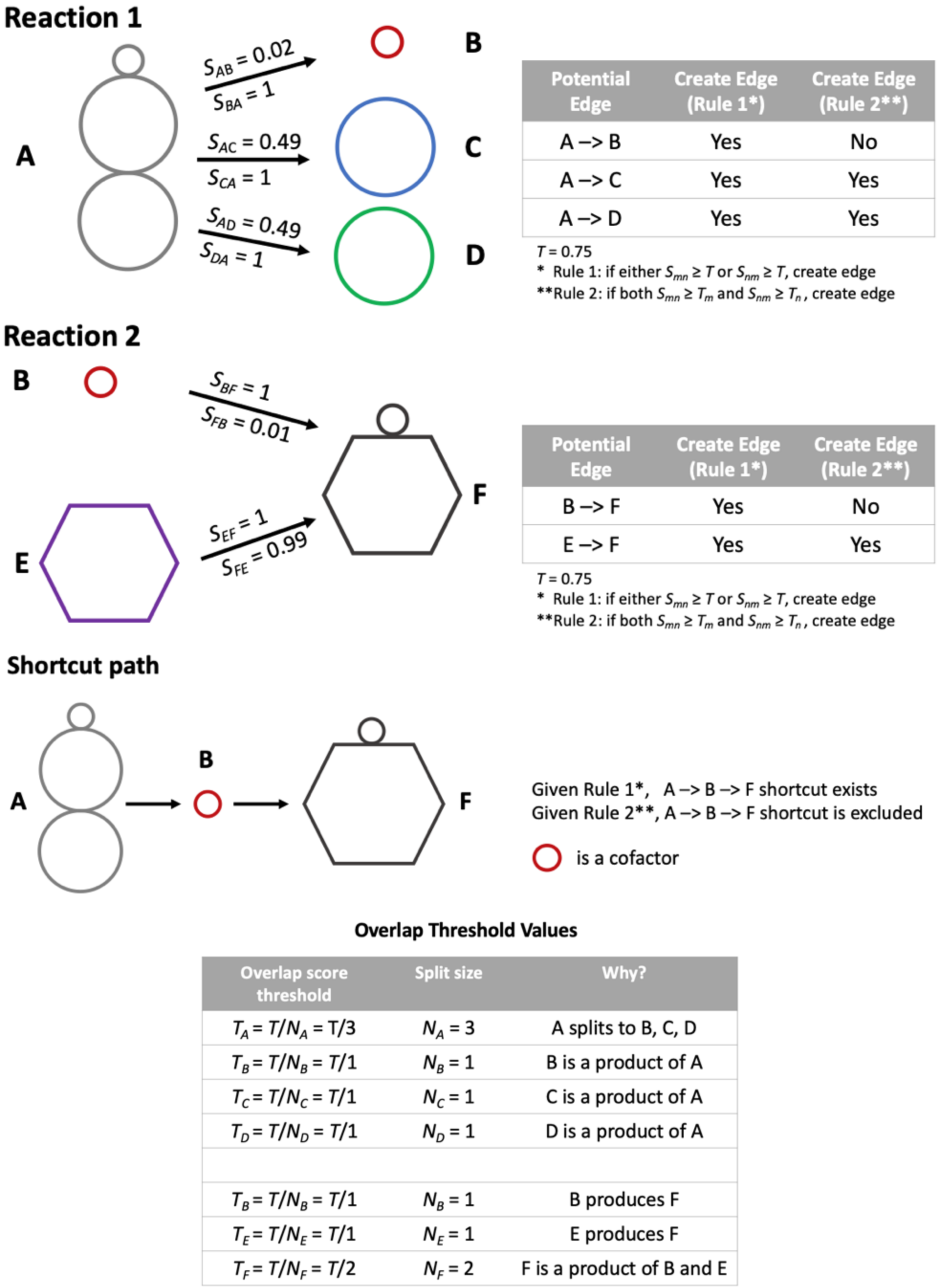
2.1 Metabolic network construction

Shortcut nodes, such as cofactors in metabolic networks and hub genes in gene networks, are a common problem in pathfinding in biological networks and are not readily excluded by simple filtering steps [25, 26]. During biological cascade reconstruction, these shortcuts can create misleading links between two molecules, leading to false molecular or uninterpretable paths. Therefore, we created a novel atom-tracing-based method to generate our metabolic network.

Recon 3.0 [13] was imported as the basis for our metabolic network. We performed the following preprocessing steps to ensure that metabolites are correctly linked and that shortcut nodes, such as cofactors, are pruned from the network (Figure 2.1).

We downloaded detailed metabolic reactions in the form of RXN files from the Virtual Metabolic Human database (VMH) [50]. RXN files contain structural data for the reactants and products of each reaction [51], including information on atoms that are transferred from reactants to products. When a metabolic reaction did not have a RXN file, we generated them from MOL files downloaded from VMH using the Reaction Decoder Tool v1.5.1 [52].

To connect metabolites in our metabolic network, we created a specialized scoring system to avoid shortcuts and ensure that for each step in the network, there is a certain amount of atomic overlap. For a given molecule pair *M* and *N*, defined by their respective sets of mapped atoms, we calculate the atomic overlap as $S_{MN} = |M \cup N|/|N|$. This overlap score $S_{MN}$is computed in a directional manner for all reactant/product and product/reactant pairs of each reaction.



Potential Edge	Create Edge (Rule 1*)	Create Edge (Rule 2**)
A –> B	Yes	No
A –> C	Yes	Yes
A –> D	Yes	Yes

T = 0.75

* Rule 1: if either $S_{mn} \geq T$ or $S_{nm} \geq T$, create edge

**Rule 2: if both $S_{mn} \geq T_m$ and $S_{nm} \geq T_n$, create edge

Potential Edge	Create Edge (Rule 1*)	Create Edge (Rule 2**)
B –> F	Yes	No
E –> F	Yes	Yes

T = 0.75

* Rule 1: if either $S_{mn} \geq T$ or $S_{nm} \geq T$, create edge

**Rule 2: if both $S_{mn} \geq T_m$ and $S_{nm} \geq T_n$, create edge

Given Rule 1*, A –> B –> F shortcut exists
Given Rule 2**, A –> B –> F shortcut is excluded

○ is a cofactor

Overlap Threshold Values

Overlap score threshold	Split size	Why?
$T_A = T/N_A = T/3$	$N_A = 3$	A splits to B, C, D
$T_B = T/N_B = T/1$	$N_B = 1$	B is a product of A
$T_C = T/N_C = T/1$	$N_C = 1$	C is a product of A
$T_D = T/N_D = T/1$	$N_D = 1$	D is a product of A
$T_B = T/N_B = T/1$	$N_B = 1$	B produces F
$T_E = T/N_E = T/1$	$N_E = 1$	E produces F
$T_F = T/N_F = T/2$	$N_F = 2$	F is a product of B and E

Figure 2.2. Example of how the overlap score threshold is used to assign edges between reactants and products, how the overlap score thresholds are calculated, and how shortcut paths are excluded.

However, simply applying an overlap threshold T to S_{nm} in order to determine which molecules should be connected in the metabolic network is not sufficient, because this would introduce shortcut paths and might miss important edges in the network, as shown in our example below. Therefore, we define a normalized overlap threshold as $T_m = T/N_m$for each reactant/product m, where T is a base overlap score threshold and N_m is the number of molecules this reactant/product is split to/synthesized from in the respective reaction. Using T_m we create an edge m-n in our network if and only if $S_{mn} \geq T_m$ and $S_{nm} \geq T_n$.

To illustrate the motivation behind this scoring method and how it is applied, we perform the following example. Suppose that in a given Reaction 1 (Figure 2.2) we have reactant A and products B, C, D. Since A is split into three products in this reaction, $N_A = 3$. Now assume that $S_{AB} = 0.02$, $S_{AC} = 0.49$ and $S_{AD} = 0.49$, $S_{BA} = 1$, $S_{CA} = 1$, and $S_{DA} = 1$, i.e. 2% of atoms in A are transferred to B, 49% of atoms in A are transferred to C, and so on. Assume we set a base overlap score threshold T of 0.75. For this reaction, $S_{AB} < T$ but $S_{BA} \geq T$. A choice must be made on how to create an edge between A and B in the network. If we require both the overlap scores to be greater than T, then there will be no edges between A and any of its products B, C, and D; hence, an "OR" rule would have to be used. However, if we only choose either of the overlap scores to be greater than T, then we will have an edge A-B (Figure 2.2 Rule 1). Let us further assume that we have another Reaction 2 (Figure 2.2) with reactants B, E and product F and $S_{BF} = 1$, $S_{EF} = 1$, $S_{FB} = 0.01$, $S_{FE} = 0.99$. Since F is synthesized from B and E, F has a split size $N_F = 2$. Here, $S_{FB} < T$ and $S_{BF} \geq T$ and since one of them is greater than T, we create an edge B-F (Figure 2.2 Rule 1). Given both reaction 1 and reaction 2, we now have a shortcut path A-B-F in our network (Figure 2.2 Shortcut Path). Note that B acts as a cofactor in both reactions, since B is much smaller than A and F. If this procedure is applied to all reactions in order to construct a metabolic network, we would get many shortcut paths passing through cofactors such as B.

To alleviate this problem, for each reactant-product pair, we calculate a reactant threshold value T_m and a product threshold value T_n (Figure 2.2, Overlap Threshold Values) as defined by $T_m = T/N_m$ and check whether $S_{MN} \geq T_M$ and $S_{NM} \geq T_N$. If both are greater than their corresponding thresholds, then an edge m-n is created in the network (Figure 2.2 Rule 2). For our example, this means that we will exclude shortcut edge A-B. Note that A is split into products B, C, and D, which means that $N_A = 3$ for A. For instance for reaction 1 and the pair A and B, $T_A = T/N_A = 0.75/3 = 0.25$ and $T_B = T/N_B = 0.75/1 = 0.75$. Since $S_{AB} < T_A$ and $S_{BA} \geq T_B$. Thus, we will not create an edge A-B. However, if we calculate this for the other pairs in reaction 1, we find that we will create edges A-C and A-D. For reaction 2 and the pair B and F, we have $T_B = T/N_B = 0.75/1 = 0.75$ and $T_F = T/N_F = 0.75/2 = 0.375$. Since $S_{BF} \geq T_B$ and $S_{FB} < T_F$, we do not create an edge B-F. In contrast, we do create an edge E-F following a similar calculation. Now the shortcut path A-B-F is effectively pruned from the network, leaving reaction 1 and reaction 2 disconnected. We used a $T = 0.75$ for the processing of our metabolic network.

We observed that reactions involving the binding of coenzyme A (CoA) to a metabolite were filtered out using our scoring method. For instance, the succinyl-CoA to succinate reaction, an important step in the TCA cycle, was omitted from our metabolic network. Thus, we manually curated a list of these reactions and added them back into our metabolic network.

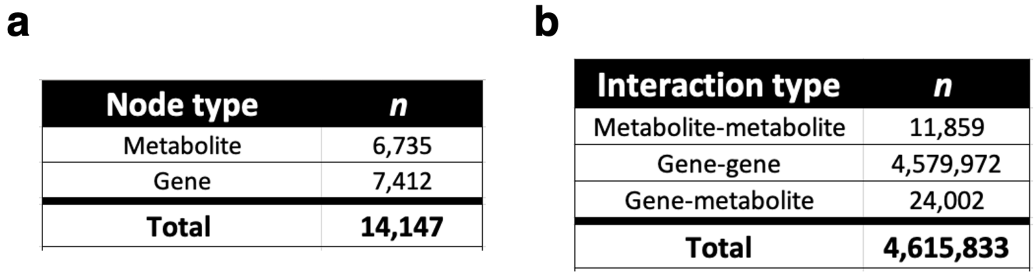
To further exclude cofactors from our metabolic network, we used a curated blacklist that includes NAD and derivatives, FAD and derivatives, nucleotides, ions, CoA, acetyl CoA, and others. The blacklist can be found together with our scripts on our github repository.

2.2 Gene Interaction network construction

We constructed a gene interaction network (Figure 2) using the following databases: (1) STRING v11 [14] was used for the protein-protein interaction (PPI) network. Only PPIs containing directional information, i.e. activation, inhibition, and catalysis and with scores above 400 for the STRING "experiments" and "experiments transferred" scores were selected. (2) Signaling cascades were imported from the OmniPath database [15] and interactions with OmniPath "consensus directionality" criteria were chosen. (3) We further integrated gene regulatory information into our multi-omics network by including a network created by Sonawane et al. [53] using PANDA [54] on data from the Genotype-Tissue Expression (GTEx) project [55] and STRING v10 [56] without any further preprocessing.

2.3 Final network

Given that both of our metabolic and gene interaction networks have common genes, i.e. enzymes, we overlapped the two networks to assemble our directed multi-omics network. Our final network consisted of 6,735 metabolites and 7,412 genes and 11,859 metabolite-metabolite, 4,579,972 gene-gene, and 24,002 gene-metabolite interactions (Figure 2.3). This high quality multi-omics network enabled our tool, described in **Chapter 3**, to retain biologically meaningful molecular relationships, to avoid false links during biological cascade reconstructions, and to reconstruct accurate molecular pathways.



a

Node type	*n*
Metabolite	6,735
Gene	7,412
Total	**14,147**

b

Interaction type	*n*
Metabolite-metabolite	11,859
Gene-gene	4,579,972
Gene-metabolite	24,002
Total	**4,615,833**

Figure 2.3. Our multi-omics network in numbers. a, node-level and **b**, interaction-level numbers. There are a comparable number of genes and metabolites. However, gene-gene interactions dominate the multi-omics network.

Chapter 3: Automatic reconstruction of molecular cascades

In this chapter, we present piTracer, an easy-to-use tool that can automatically and rapidly generate candidates for biologically accurate molecular cascades between molecules across different omics. To the best of our knowledge, only two tools, ConsensusPathDB (CPDB) [57] and the commercial Ingenuity Pathway Analysis (IPA) software [58], have attempted to address this problem. However, these tools have two core problems: (1) Misleading shortcuts exist in their multi-omics network. For piTracer, we have addressed this problem in **Chapter 2**, which is the basis for the ability of piTracer to construct true molecular cascades. (2) multiple molecular paths of different lengths connect two molecules in biological networks. Since CPDB and IPA only use shortest paths with a constant length to connect two molecular entities, many biologically informative paths can be missed in a query. In contrast, piTracer allows for the simultaneous reconstruction of multiple molecular cascades of various lengths.

Terminology	Definition
k-shortest paths	Paths in a network including the shortest path and k-1 other shortest paths, which may be longer than the shortest path
Path	Refers to one of the k shortest paths
Trace	As a noun: Defined as a collection of paths between a start and end node pair. As a verb: The act of generating multiple paths between a start and end node(s)

Table 1. Terminology used in reconstructing molecular cascades.

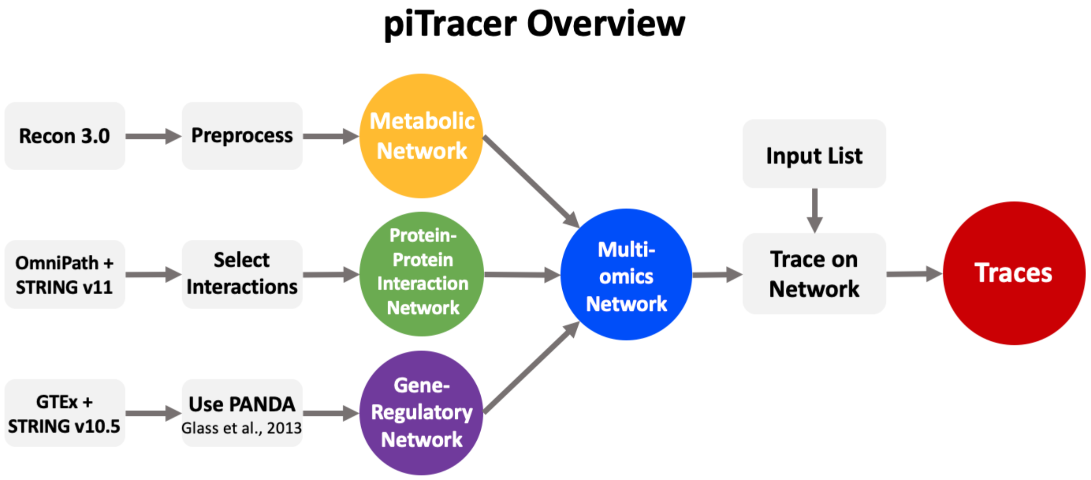


Figure 3.1. The piTracer app backend, showing the databases used in the construction of the multi-omics network, as shown in Figure 2.1, and a scheme of the tracing approach. The input to the method is a list of genes and/or metabolites, e.g. derived from statistical analysis of an experiment.

piTracer is a Shiny-based web application that enables the querying of distant functional relationships between molecules and powerful interactive visualizations of the molecular traces (Figure 3.1). We combined multiple databases across different omics after extensively curating and applying new processing methods to these databases to remove shortcut nodes such as cofactors. We also implemented an algorithm that allows for the creation of traces containing differing molecular path lengths and a clustering algorithm to aid visualization of the traces. piTracer is written in R, open-source and freely available. We performed metabolite-to-metabolite, gene-to-metabolite, and gene-to-gene tracing to validate the ability of piTracer to automatically construct traces between molecules.

3.1 Methods

In the following, we describe the different algorithmic steps to extract paths and traces between a start and an end node from the multi-omics network.

3.1.1 Yen's *k*-shortest path algorithm

Yen's k-shortest path algorithm calculates the *k*-shortest paths between a pair of nodes in a network. The algorithm initially finds the shortest path between a pair of nodes, and subsequently enumerates the *k*-th shortest path based on node deletion and recalculation of a shortest path based on previously calculated *k*-1 shortest paths. The assumption this algorithm makes is that the *k*-th shortest path shares edges and sub-paths with the (*k*-1)-th shortest paths and these can thus be used for calculating the *k*-th shortest path [59].

3.1.2 Path clustering

To improve the visualization of the *k*-shortest paths in a trace, especially for large values of *k*, we developed a clustering algorithm to group paths. We first compute the distances between the *k*-shortest paths between two nodes using the CoMapPa2 algorithm [60]. Given these distances, a path dendrogram is then created using the hclust function in R [61]. Path clusters are obtained through dynamic tree cutting using cutreeHybrid function from the Dynamic Tree Cut R package [62]. Clusters of paths are expected to represent shared biological characteristics and can be useful in selecting groups of paths for further investigation.

3.1.3 Heuristic for speed-up

To speed up trace calculations, we dynamically reduce the search space of the *k*-shortest paths algorithm per trace. We achieve this by confining the search to a subnetwork of our dense multi-omics network for each pair of start and end nodes. This subnetwork will contain a subset of all the shortest paths that exist between a pair of nodes. That is, it might not contain all possible *k*-shortest paths for a trace and will cap the value of *k*, but this will significantly accelerate the calculation of traces. We implemented the heuristic as follows: (1) Initially, we extract the set N_{start} of all reachable nodes from the start node and another set N_{end} of all nodes that can reach the end node. (2) We define a hyperparameter v, such that we can calculate an overlap set $O \subseteq N_{start} \cap N_{end}$ for which $|O| \leq v$. For instance, $v = 5$ means that at most there are 5 molecules in O . Given that $|O| \leq v$, the hyperparameter v determines the number of maximum shortest path k that can be calculated for a trace. We currently set $v = 5$, which on average allows 300 shortest paths to be found on our network, depending on the start and end nodes chosen. Increasing v increases the number *k*-shortest paths that can be calculated per trace in exchange for an increase in computational time. (3) We calculate a set D_s of the shortest distances between all pairs of the start node and nodes in O. We then extract the v shortest distances from D_s and select the maximum value d_s. The same procedure is performed using the end node to calculate d_e. (4) Afterwards, a subnetwork of the multi-omics network is created by including all nodes and edges within distance $\mathrm{d_s}$ towards the O nodes and distance $\mathrm{d_e}$ from the O nodes. (5) Tracing is then performed on the final, restricted multi-omics subnetwork.

3.1.4 IPA and CPDB

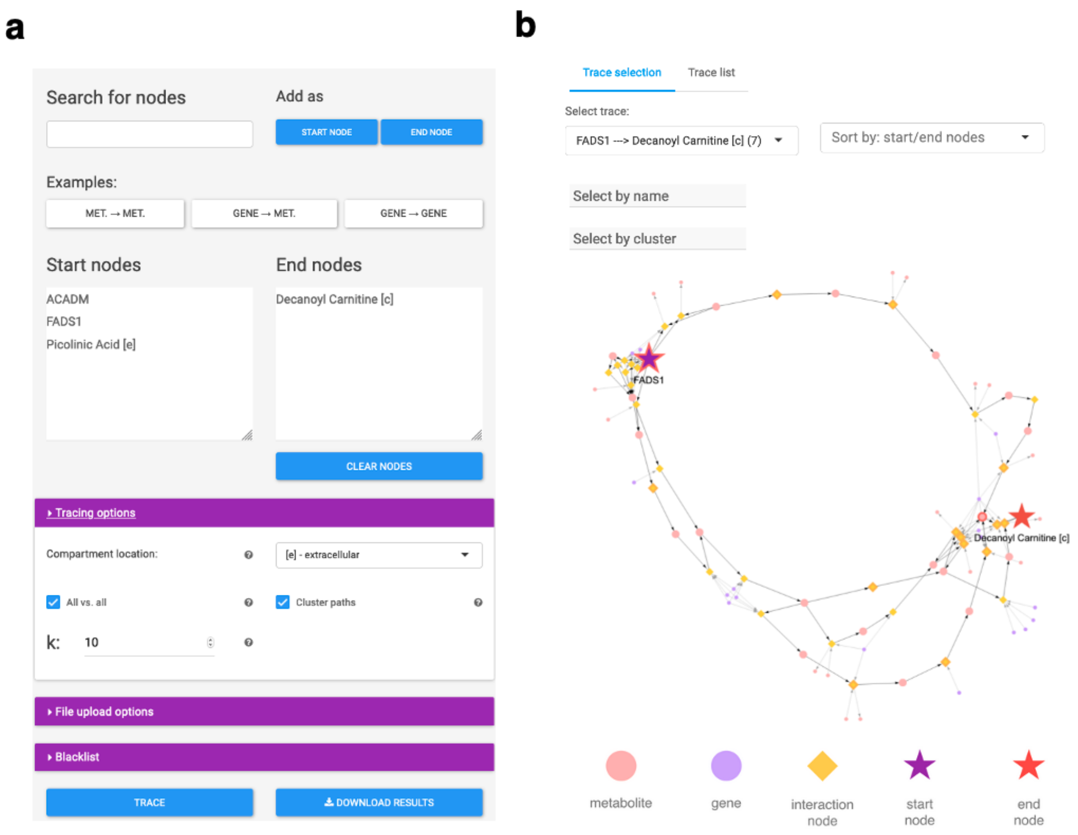
The traces produced by piTracer were compared to traces generated from two other state-of-the-art tools: (1) IPA, which is a commercial software by QIAGEN. To create traces in IPA version 01-16, the Path Explorer tool was used. Additionally, we used either “direct interaction” or both “direct and indirect interactions” options for tracing, depending on which option resulted in a more meaningful trace. Moreover, we selected the “relaxed filters” option for the “Species” and “Tissues and Cell Lines” options for all traces. (2) CPDB, which is a web-based tool that integrates human molecular interaction data and provides computational methods and visualization tools to explore these data [57]. For CPDB, we used the “shortest interaction path” function to create and the “Visualize path” function to visualize traces. We used the “exclude” function to blacklist nodes in CPDB, in order to produce traces with minimal shortcut nodes.

3.1.5 piTracer code, docker image, and web page

piTracer has been developed in R [61] version 3.6.1 and Shiny 1.4.0.2 [63]. Free access to a hosted version of piTracer is provided at http://cbsunemo.biohpc.cornell.edu/pitracer/. The codes used in piTracer are separately available at the GitHub repository https://github.com/krumsieklab/piTracer.

3.2 Results

3.2.1 The piTracer Shiny app user interface



Start nodes: ACADM, FADS1, Picolinic Acid [e]

End nodes: Decanoyl Carnitine [c]

Compartment location: [e] - extracellular

☑ All vs. all

☑ Cluster paths

k: 10

Select trace: FADS1 —> Decanoyl Carnitine [c] (7)

Sort by: start/end nodes

Figure 3.2. Screenshot of the piTracer app in a web browser. (**a**) Input panel. Users have the option to search for molecules in the piTracer database, trace between start and end node pairs or trace between all start nodes and all end nodes, upload lists directly into the start and end nodes text boxes, blacklist nodes in the traces, set the number of shortest paths *k*, and cluster traces based on similarity. Users can also download the visualizations as HTML files. (**b**) Results tab including interactive network visualizations and a detailed trace list.

The easy-to-use user interface of piTracer starts with an input area (Figure 3.2a) that enables users to enter a list of metabolites and/or genes. Alternatively, a comma separated value (CSV) or Excel spreadsheet file containing a list of metabolites and genes can be uploaded. piTracer accepts identifiers from HMDB, KEGG, PubChem, and Recon for metabolites and HGNC gene symbols for genes. Specific genes/metabolites can be dynamically searched for in the app database using an autofill textbox. Users can specify whether to trace between pairs of start and end nodes as they are arranged in the text area or between all pairs of start and end nodes. Molecules can also be blacklisted by using a user-defined list prior to tracing. Moreover, the number of shortest paths between start and end nodes to be calculated for traces, and whether to cluster paths for easier visualization can be specified. Once traces have been generated, both visualizations and a list containing all the generated paths in the traces can be downloaded.

After submitting the start and end nodes list for tracing, users are presented with the traces (Figure 3.2b) where they can dynamically select and sort which traces to view. Moreover, the visualizations are interactive, allowing for network zooming and moving of nodes. Users additionally have access to a trace list, where the list of paths for each start and end node trace is provided.

3.2.2 Gene-to-metabolite traces

Interactions between genes and metabolites are not always clear-cut as metabolic reaction cascades. For instance, a statistical association found between a single nucleotide polymorphism (SNP) at a specific locus and a certain metabolite in a genome-wide association study (GWAS) could be caused by a variety of different mechanisms, from simple enzyme-to-reactant relationships to complex cascades of genetic regulation. Given this nontrivial task, we here show that piTracer automatically and rapidly generates biologically meaningful hypotheses that provide explanations for gene-to-metabolite associations. As a test case we used several associations found in the currently highest powered GWAS study involving metabolites [64] to illustrate the hypothesis-generation functionality of piTracer.

Uncovering the correct regulatory path for the complex genetic association between rs2403254 and 2-hydroxyisovalerate

In their initial GWAS paper, Shin et al. [64] explained the association between SNP rs2403254 and 2-hydroxyisovalerate with the HPS5 gene, since that SNP is located in in an intronic region of HPS5. Other candidate genes for the SNP found by a different study were GTF2H1, SAA1, and LDHA [65]. This later study demonstrated that in fact LDHA was the correct causal gene for the association by showing that LDHA converts 3-methyl-2-oxobutanoate into 2-hydroxyisovalerate in an experiment [65]. In contrast, they did not find any biochemical explanations for the HPS5 to 2-hydroxyisovalerate association [65]. We assessed all four gene candidates using the piTracer app.

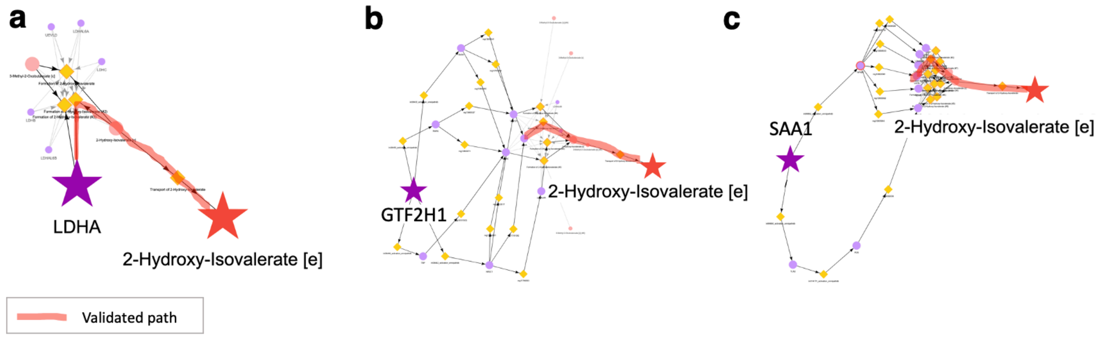

a
LDHA
2-Hydroxy-Isovalerate [e]
Validated path
b
GTF2H1
2-Hydroxy-Isovalerate [e]
c
SAA1
2-Hydroxy-Isovalerate [e]

Figure 3.3. Visualization of traces between (**a**) LDHA, (**b**) GTF2H1, and (**c**) SAA1 and 2-hydroxyisovalerate using piTracer. The highlighted paths in **b** and **c** are the LDHA shortest path in A.

First, there was no trace from HPS5 to 2-hydroxyisovalerate, since the gene does not have any functional annotations in our database. This is in line with the findings of Heemskerk et al. [65]. While we acknowledge that research bias might play a role here, the lack of functional information in extensive databases provides evidence against HPS5 being the correct gene. We then generated traces (k = 10) from LDHA, GTF2H1, and SAA1 to 2-hydroxyisovalerate (Figures 3.3a, 3.3b, and 3.3c, respectively) and subsequently ranked gene-to-metabolite associations by plausibility. We found that the LDHA trace contains the shortest path to 2-hydroxyisovalerate (Figure 3.3a, highlighted path) and that LDHA directly catalyzes a reaction involving 2-hydroxyisovalerate. Interestingly, the GTF2H1 and SAA1 traces, which consisted of longer paths, are extensions of the LDHA shortest path (Figure 3.3b and Figure 3.3c, highlighted paths). Using these traces, it would have been immediately clear that LDHA had the highest likelihood to be the correct gene associated with 2-hydroxyisovalerate.

We similarly traced between LDHA, GTF2H1, and SAA1 and 2-hydroxyisovalerate using IPA (Figures 3.4a to 3.4c). However, all traces had a constant path length, none contained the correct enzyme-metabolite relationship, and were substantially longer and less specific than the shortest path found by piTracer. Additionally, we also found a trace between the wrong gene HPS5 and 2-hydroxyisovalerate with the same path lengths, suggesting that traces by IPA are uninformative and do not allow for the prioritization of the gene to metabolite association hypotheses in this case. We similarly attempted to generate traces between the 4 genes and 2-hydroxyisovalerate using CPDB. However, the metabolite did not exist in the CPDB database.

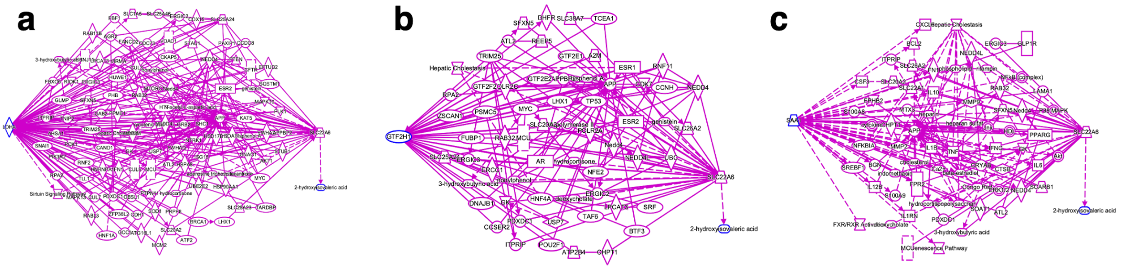


Figure 3.4. Visualization of traces between (**a**) LDHA, (**b**) GTF2H1, and (**c**) SAA1 and 2-hydroxyisovalerate using IPA.

Beta-oxidation and carnitine shuttle pathway between ACADM and hexanoylcarnitine

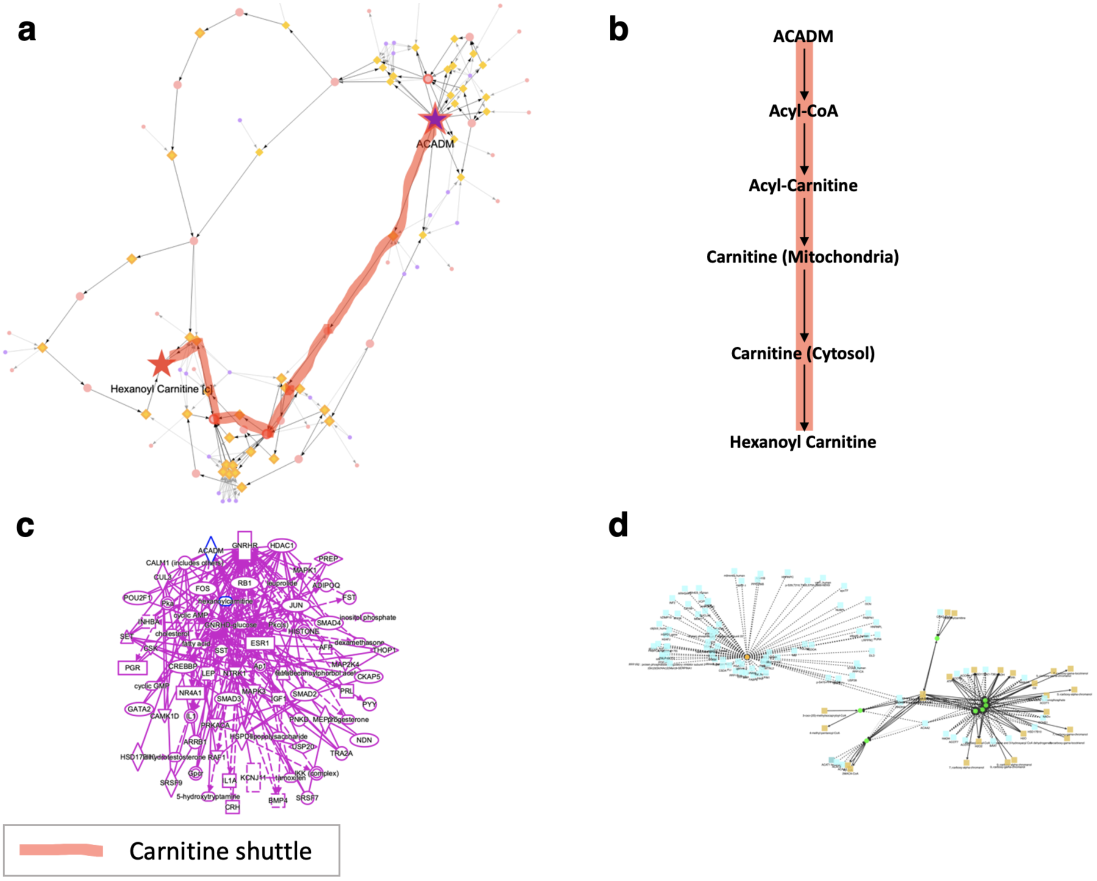


Figure 3.5. ACADM-hexanoylcarnitine trace. (**a**) Trace produced by piTracer. (**b**) The carnitine shuttle. Traces produced by (**c**) IPA and (**d**) CPDB. ACADM and carnitine were used for the CPDB trace, since hexanoylcarnitine does not exist in its database.

Another highly significant association found by Shin et al. [64] and various other previous studies [40, 66] was between a SNP in ACADM and the metabolite hexanoylcarnitine. The biochemical cascade behind this association is well understood. Briefly, ACADM is an enzyme involved in mitochondrial beta-oxidation and hexanoylcarnitine is a transport variant of the substrate and product of this enzyme [67]. Tracing between ACADM and hexanoylcarnitine with $k = 10$ indeed recovers all steps of this mechanism (Figures 3.5a and 3.5b).

We also attempted to construct traces using IPA and CPDB (Figures 3.5c and 3.5d). We traced from ACADM to carnitine using CPDB, since hexanoylcarnitine did not exist in its database. Neither could recover a known mechanism connecting ACADM with hexanoylcarnitine. IPA recovers paths that all pass through gonadotropin-releasing hormone (GNRH), its receptor GNRHR, and a synthetic hormone leuprolide, while CPDB finds paths between ACADM and carnitine through cofactors.

Further validation examples

We analyzed further SNP/metabolite pairs from the Shin paper (cite again) to provide additional validation of piTracer. These include traces between ALDH18A1 and citrulline, MCCC1 and 3-hydroxyisovalerylcarnitine, and DBH and vanillylmandelate (Figures 3.6-3.8, respectively). In all cases, we found the correct biochemical paths using piTracer, in contrast to IPA and CPDB.

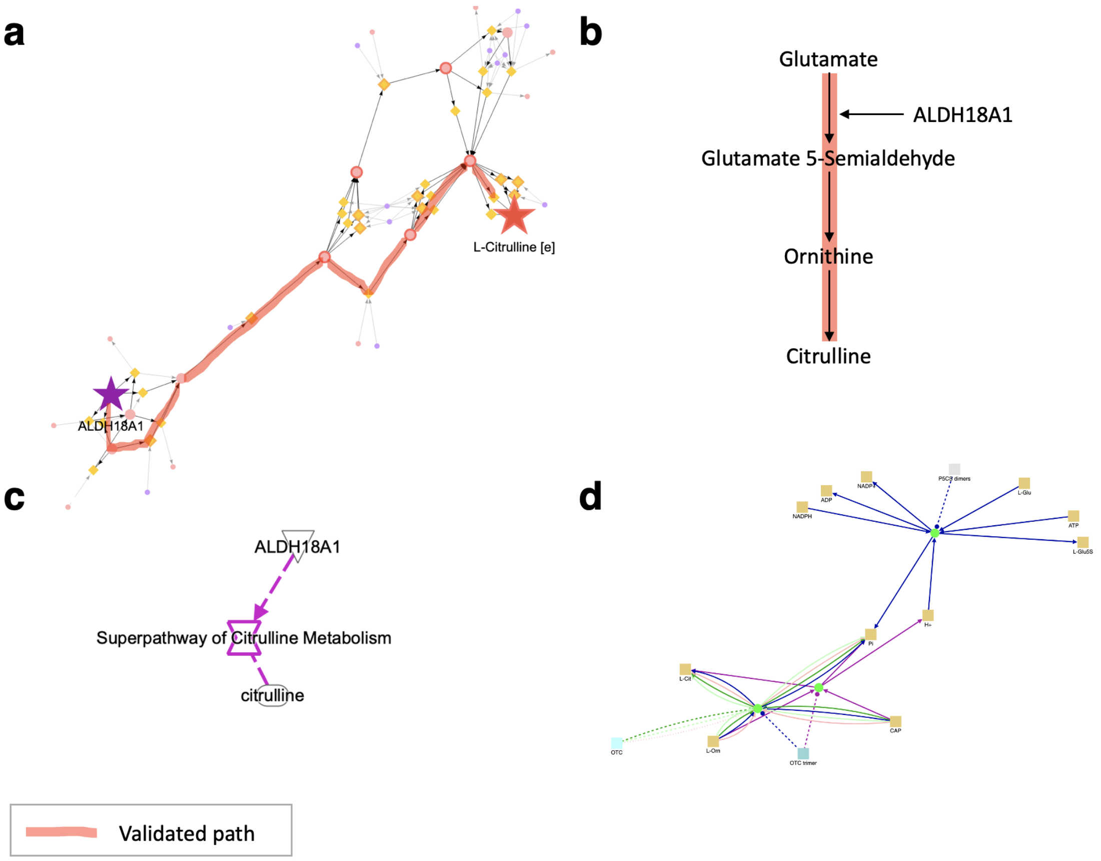


Figure 3.6. Visualization of trace between ALDH18A1 and citrulline. (**a**) Trace produced by piTracer. (**b**) Textbook version, (**c**) M IPA trace, and (**d**) CPDB trace of the pathway.

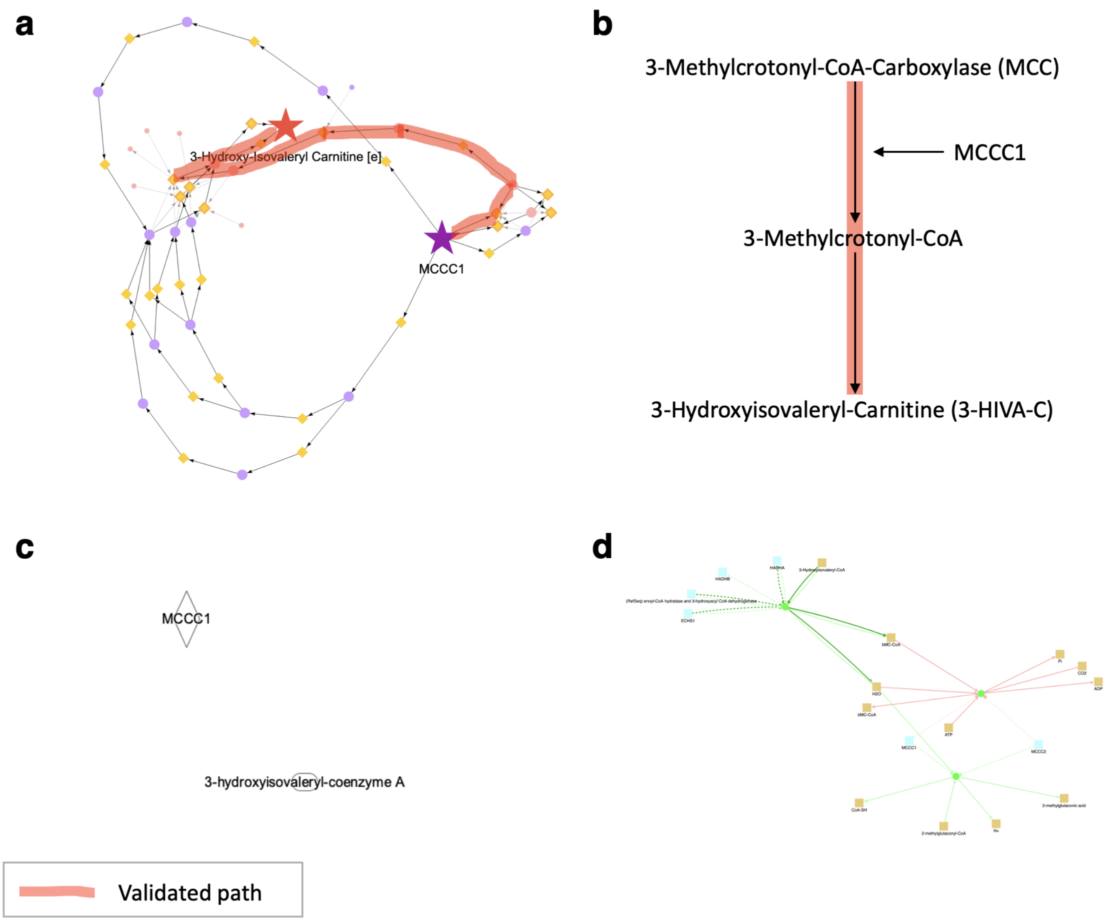

a
3-Hydroxy-Isovaleryl Carnitine [e]
MCCC1
b
3-Methylcrotonyl-CoA-Carboxylase (MCC)
MCCC1
3-Methylcrotonyl-CoA
3-Hydroxyisovaleryl-Carnitine (3-HIVA-C)
c
MCCC1
3-hydroxyisovaleryl-coenzyme A
d
HADHA
3-Hydroxyisovaleryl-CoA
HADHB
(RefSeq) enoyl-CoA hydratase and 3-hydroxyacyl CoA dehydrogenase
ECHS1
bMC-CoA
Pi
CO2
ADP
H2O
bMC-CoA
ATP
MCCC1
MCCC2
CoA-SH
2-methylglutaconic acid
2-methylglutaconyl-CoA
H+
Validated path

Figure 3.7. Visualization of trace between MCCC1 and 3-hydroxyisovaleryl carnitine. (**a**) Trace produced by piTracer. (**b**) Textbook version, (**c**) IPA trace, (**d**) CPDB trace of the pathway. Note that for IPA and CPDB, MCCC1 was traced to 3-hydroxyisovaleryl-coenzyme A.

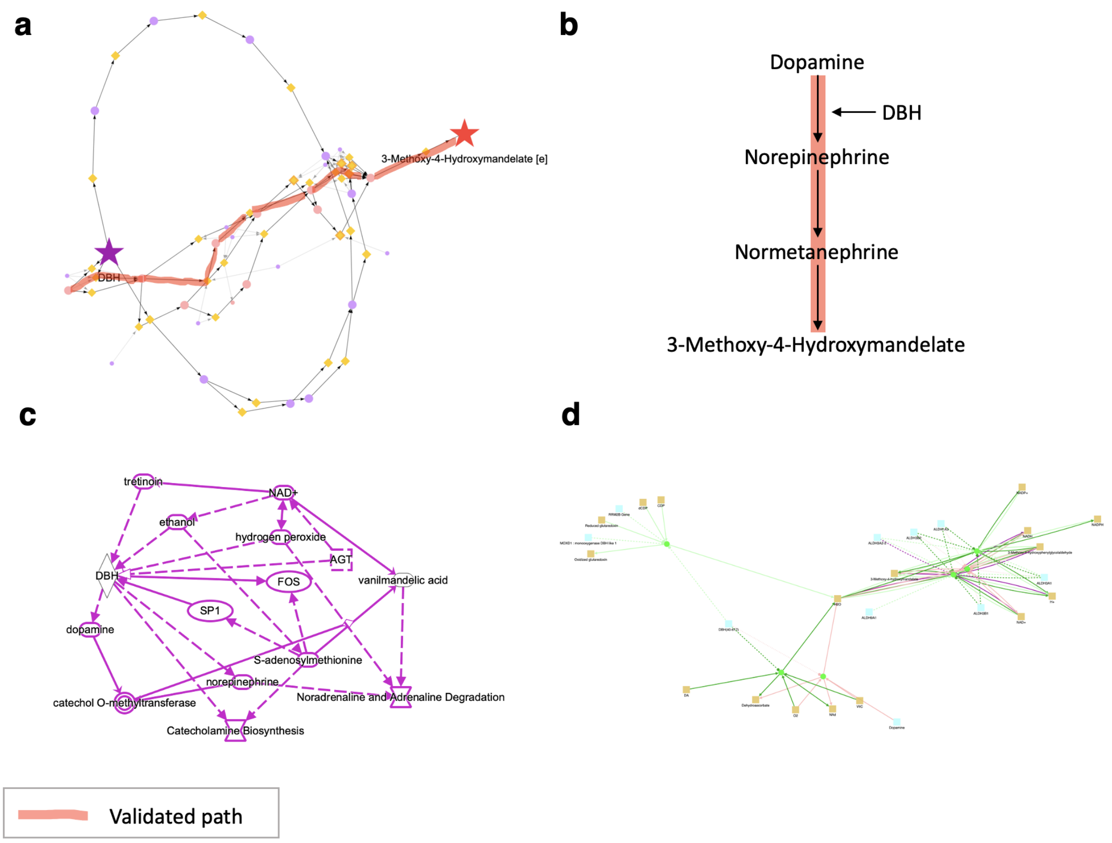


Figure 3.8. Visualization of trace between DBH and vanillylmandelate (3-methoxy-4-hydroxymandelate). (**a**) Trace produced by piTracer. (**b**) Textbook version, (**c**) IPA trace, (**d**) CPDB trace of the pathway.

3.2.3 Metabolite-to-metabolite traces

Finding multiple reaction steps connecting metabolites is often a nontrivial task. For instance, a statistical correlation between two metabolites can result from an experiment, but it may not be apparent how these are metabolically connected, unless they are the direct reactant and product of the same reaction. Moreover, there may be multiple paths that explain the connection between two metabolites and enumerating these can be challenging. Here we show that piTracer reconstructs well-established metabolic pathways as a positive control

Finding paths in central carbon metabolism

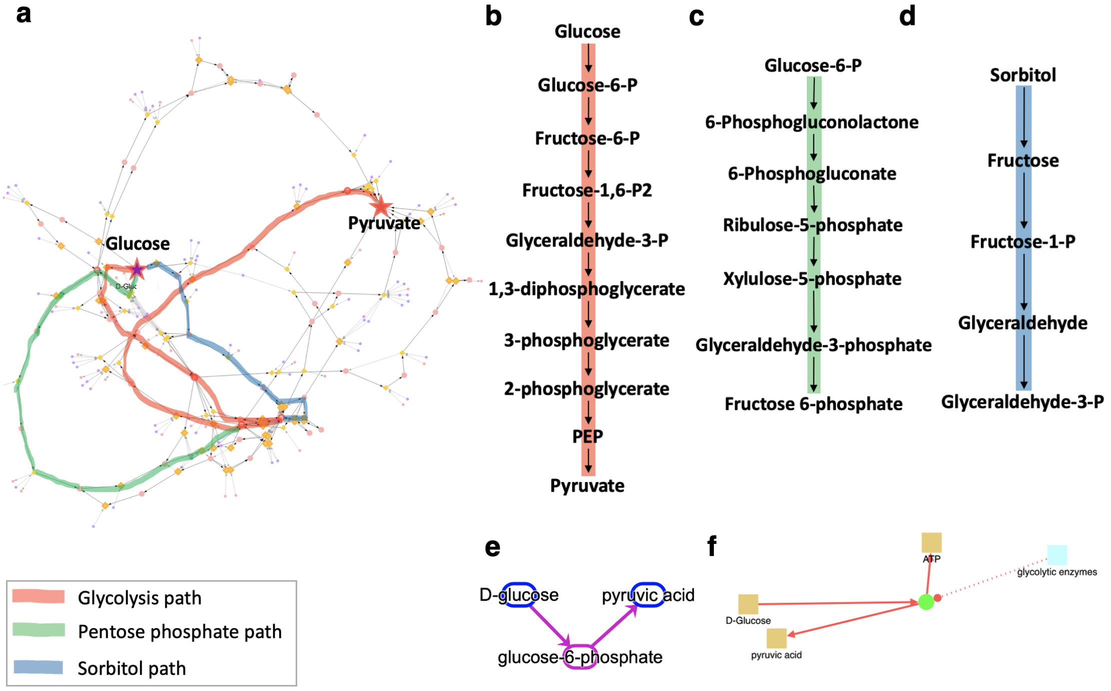


Figure 3.9. The glycolysis, pentose phosphate, and sorbitol pathways. Traces were produced by using glucose as the starting node and pyruvate as the end node (**a**) Trace constructed using piTracer with manually added highlighting of three well-known carbohydrate metabolism pathways. (**b-d**) Textbook versions of the glycolysis, pentose phosphate, and sorbitol pathways. (**e**) IPA and (**f**) CPDB reconstructions, resulting in single-step paths between glucose and pyruvate.

We started by tracing between glucose and pyruvate, two major metabolites of central carbon metabolism. To this end, the top 70 shortest paths connecting the two metabolites were queried. We chose the top 70 shortest paths since this was the minimum number of paths that could reconstruct the glycolysis pathway. The tracing (Figure 3.9a) resulted in three major axes: The glycolysis pathway (Figure 3.9b), the pentose phosphate pathway (Figure 3.9c), and the sorbitol pathway (Figure 3.9d). These three pathways are major constituents of carbohydrate metabolism and are essential for the survival of all cells. Notably, the other paths contained in the trace may also represent valid molecular steps, however only partial stretches of the alternative paths have published studies associated with them. For instance, in the trace, there is a path that connects glucose to pyruvate via glucose-6-phosphate (G6P), myo-inositol-1-phosphate, myo-inositol, phosphatidate, phosphatidyl-serine, and serine. It has been established that glucose affects myo-inositol [68], phosphatidate is involved in the production of phosphatidyl-serine [69], and serine yields pyruvate [70]. However, there are no studies showing the synthesis of pyruvate from glucose via this path.

We also queried the connection between glucose and pyruvate in IPA and CPDB. The IPA path connected glucose to pyruvate via glucose-6-phosphate (Figure 3.9e) and CPDB returned a single step reaction between glucose and pyruvate (Figure 3.9f). Despite using the blacklist functionality in CPDB and choosing the best interaction filters for IPA (see Methods 3.1), neither could construct any of the known carbohydrate metabolism pathways.

Reconstructing the citric acid cycle

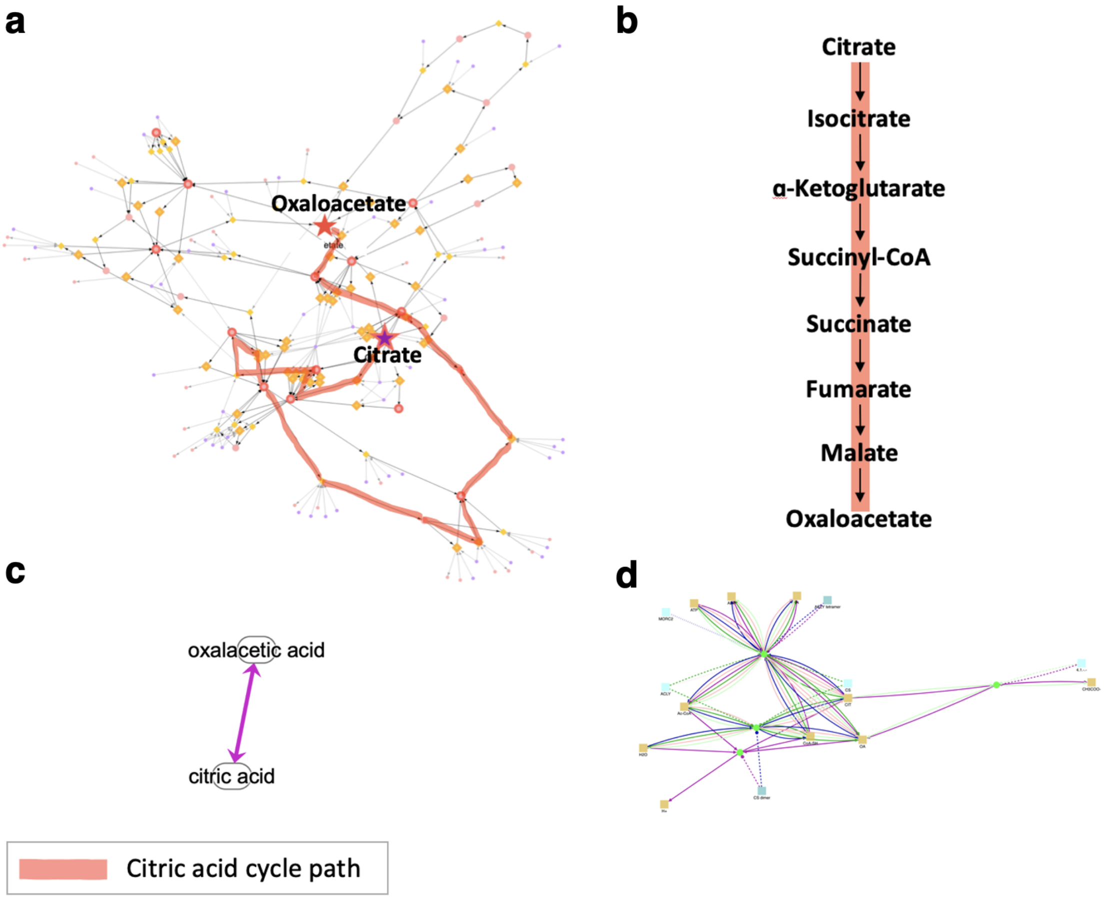


Figure 3.10. The citric acid cycle. Traces were produced by using citrate as the starting node and oxaloacetate as the end node. (**a**) Trace produced by piTracer. (**b**) Textbook version, (**c**) IPA trace, (**d**) CPDB trace of the pathway. Only piTracer is able to recover the citric acid cycle, although CPDB and IPA produced biochemically valid paths other than the pathway.

We further validated piTracer by tracing from citrate to oxaloacetate, the start and end point of the citric acid cycle. By querying the 10 shortest paths between the two metabolites, we were able to find a path that reconstructs the citric acid cycle (Figure 3.10a and 3.10b). IPA and CPDB were also used to generate citric acid cycle traces. Both tools produced the same biochemically valid single reaction step (Figure 3.10c and 3.10d), a connection from the end (oxaloacetate) to the start (citrate) of the citric acid cycle, rather than a path around the cycle.

3.2.4 Gene-to-gene traces

Functional interactions between genes are at the core of all cellular processes. Many of these interactions have been curated and stored in network databases. However, similar to metabolic interactions, querying these databases for gene-to-gene relationships can lead to dense networks that are difficult to interpret. Part of the reason for the density of these networks is that different interaction types exist between genes; for example, genes can bind, activate or inhibit other genes. Network databases collect these relations into large sets of gene-to-gene interactions. Thus, it becomes challenging to choose paths that best explain gene relationships. Here we show that our stringent network filtering steps and our path-finding algorithm enables piTracer to generate biologically valid gene-to-gene traces that are sparse and easy to interpret by using several well-known molecular cascades as examples.

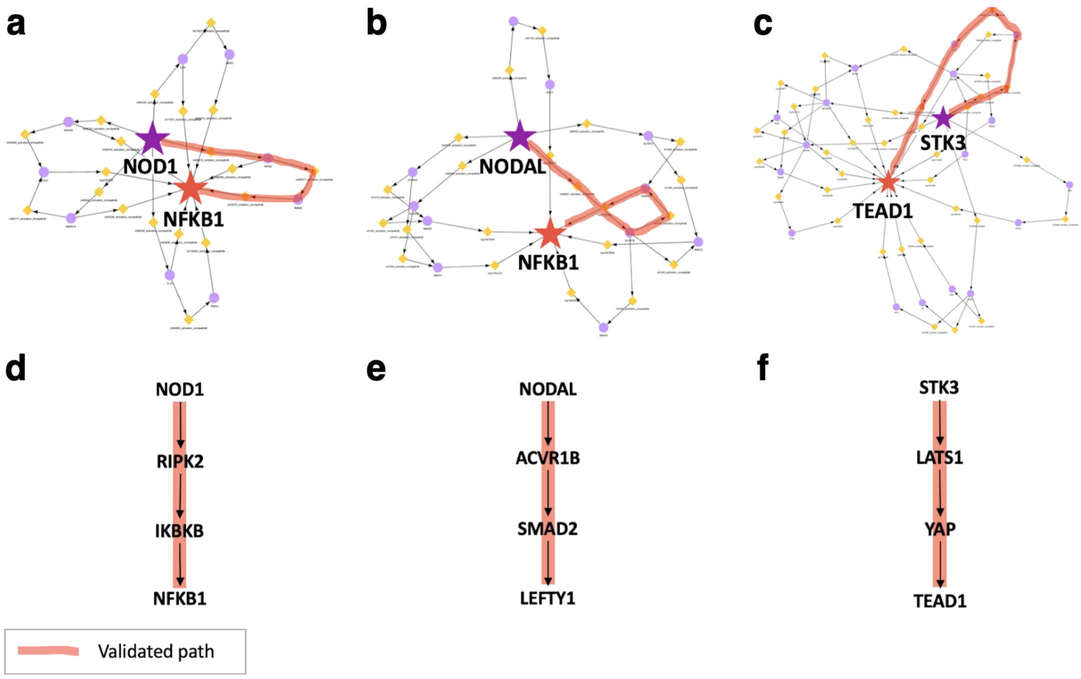


Figure 3.11. Traces between (**a**) NOD1 and NFKB1, (**b**) NODAL and LEFTY1, and (**c**) STK3 and TEAD1 produced by piTracer. Adapted textbook versions of (**d**) NOD-like receptor, (**e**) NODAL, and (**f**) Hippo signalling pathways.

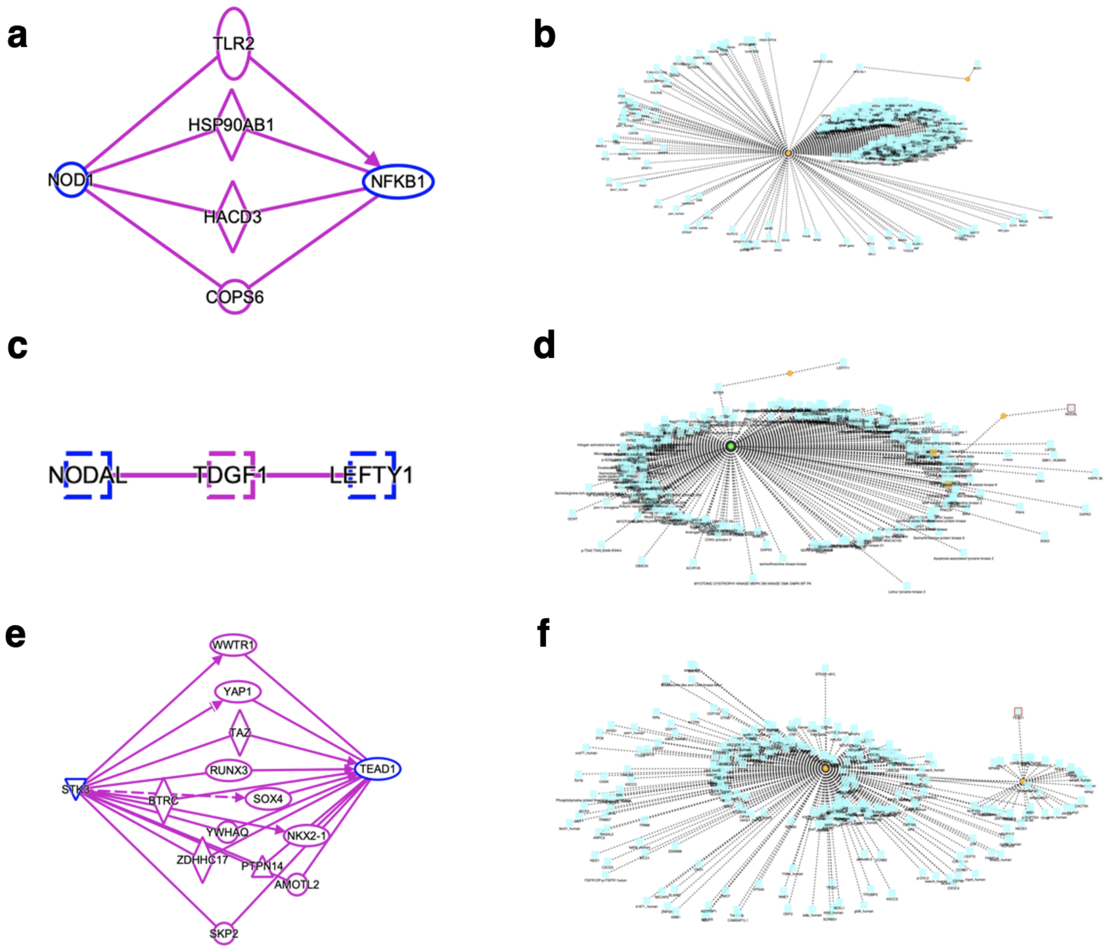


Figure 3.12. IPA and CPDB traces between (**a**) NOD1 and NFKB1, (**b**) NODAL and LEFTY1, and (**c**) STK3 and TEAD1.

The NOD-like receptor signalling pathway

We attempted to reconstruct the NOD-like receptor signaling pathway by tracing the top 10 shortest paths between NOD1 and NFKB1 (Figure 3.11a). NOD1, together with NOD2, senses conserved motifs in bacterial peptidoglycan and induces anti-microbial responses and pro-inflammatory cytokines through NF-κB activation [71, 72]. piTracer successfully found the major steps in the signaling pathway, namely the NOD1, RIPK2, IKBKB and NFKB1 path (Figure 3.11d).

We also traced between NOD1 and NFKB1 using IPA and CPDB (Figure 3.12a and 3.12b, respectively). Both IPA and CPDB indicated the existence of a connection between the two genes, however they did not capture any of the specific steps (Figure 3.11d) constituting the pathway.

The NODAL signalling pathway

As another example, we reconstructed the NODAL signalling pathway, by tracing between NODAL and its inhibitor LEFTY1 ($k = 10$) (Figure 3.11b). This pathway plays a central role in the maintenance of embryonic stem cell pluripotency, the patterning of the early embryo during mesoendoderm induction, and the dorsal-ventral axis specification in embryos [73]. Not only did our trace recover this signaling pathway (Figure 3.11b highlighted path), but also included other equally valid paths that pass through different subcomponent combinations of Activin type 1 (ACVR1) and 2 receptors (ACVR2) and different SMADs involved in the pathway (Figure 3.11e).

Using IPA and CPDB, we also traced between NODAL and LEFTY1. While IPA connected the two genes in two steps using different gens (Figure 3.12c), CPDB found a path between NODAL and LEFTY1 via MTOR, which is part of a different pathway, namely the mTOR signalling pathway (Figure 3.12d).

The Hippo signalling pathway

As a third example, we reconstructed the Hippo signalling pathway using piTracer by tracing between STK3, a core kinase of the signalling pathway, and TEAD1 (Figure 3.11c). STK3 is a core kinase of the Hippo signalling pathway [74] and activates the transcription factor TEAD1 which leads to cell proliferation and survival [75], oncogenesis and chemotherapeutic resistance [76]. Our app successfully reconstituted the pathway (Figure 3.11f) and IPA included all the major steps of the pathway but one, namely the LATS1/2 step (Figure 3.12e). However, CPDB generated a trace that was not biologically meaningful (Figure 3.12f).

3.3 Summary & Discussion

piTracer is an app that integrates various human molecular interaction databases across omics to enable the rapid and automatic generation of biologically meaningful pathway cascades between molecules. To the best of our knowledge, it is the only tool that enables the automatic construction of valid molecular interaction paths across omics, even for distant interactions, an otherwise onerous task to perform manually. Being able to generate these cascades is essential, since statistical results from high-throughput datasets are often not readily interpretable. For instance, a biological relationship is obvious when a gene-to-metabolite association is found between a gene encoding an enzyme and a metabolite of the corresponding catalyzed reaction. However, when two molecules do not share a direct relationship, which is often the case, it can be challenging to find biological cascades connecting the two. We showed the capability of piTracer to address this problem by generating metabolite-to-metabolite, metabolite-to-gene, and gene-to-gene traces, which were validated by previously published studies. Furthermore, the validation examples could not be reproduced using most widely used tools IPA and CPDB. Taken together, these results show that piTracer is the only existing tool that can accurately and automatically construct true biological cascades between molecules, which can be used in applications such as interpretation of complex genetic associations.

We traced between glucose and pyruvate and found several well-known glucose metabolism pathways embedded in our trace, including glycolysis, pentose phosphate, and sorbitol pathways. The other paths contained in the trace may also be valid molecular steps, however only parts of these paths have been previously reported in previous studies. We also generated a trace between citrate and oxaloacetate where we recovered the citric acid cycle. Only piTracer found these essential carbohydrate/nucleotide metabolism pathways, as opposed to IPA and CPDB. This clearly demonstrates that our metabolic network processing retains valid metabolic interactions which can be used with our pathfinding algorithm to construct true metabolic cascades.

Statistical associations between SNPs and metabolites, like those found in Shin et al. [64], are often difficult to interpret, mainly due to the challenging task of finding the correct causal gene and biochemical pathway explaining a SNP-to-metabolite association. Originally, Shin et al. [64] incorrectly reported that HPS5 was the causal gene explaining the rs2403254 and 2-hydroxyisovalerate association. We used piTracer to correctly identify LDHA as the causal gene for this association, which was the same gene identified by an independent study [65]. This demonstrates the ability of piTracer to prioritize which gene explains a SNP-to-metabolite association.

Given that gene networks are dense, it is challenging to automatically extract sparse and biologically meaningful molecular cascades between pairs of genes. Furthermore, a sparse trace does not guarantee that the reconstructed paths are accurate. For instance, when we traced between the endpoints of known signalling cascades, namely the NOD-like receptor, NODAL, and Hippo pathways using IPA and CPDB, we were not able to reconstruct the entirety of the pathways. IPA returned sparse traces, but frequently missed major steps of the pathways, while CPDB returned dense networks that were difficult to untangle. This can be partially explained by the gene networks used by these tools, which have not been processed and were used as is. In contrast, we specifically selected gene regulatory, activating, and inhibitory interactions that are experimentally validated from various databases in order to create our multi-omics network and make it less dense.

The number of shortest paths k is a central parameter of piTracer that substantially affects the results produced by the algorithm. We used a value of $k = 10$ for the majority of our validation traces. However, we generated our final traces for the STK3 to TEAD1 pair (Hippo pathway) with $k = 20$, for the glucose to pyruvate pair, $k = 70$, and for the citrate to oxaloacetate pari, k = 25. Although we could generate valid paths with $k = 10$ for these pairs, we observed that for the glucose to pyruvate trace, a $k = 70$ was the minimum number required to include the glycolysis pathway in our final result. This is due to the fact that many biochemically valid paths connect glucose and pyruvate, as shown in our trace, and a higher value of k is needed in order to recover the majority of them. We made the same observations for the STK3 to TEAD1 and citrate to oxaloacetate traces. With these three traces being the only exceptions, and that almost all of our validation traces were generated with a $k = 10$, we suggest that a $k = 10$ is an adequate initial value to find paths between molecules. Furthermore, when it is hypothesized that many paths exist between a molecule pair, higher values up to $k = 100$ could be required to reconstruct all major paths between them.

piTracer is, to the best of our knowledge, the first freely available and open-source app that can rapidly generate correct biological paths of varying lengths between molecules for the interpretation of complex statistical results.

Chapter 4: Predicting druggable genes from reconstructed cascades

This chapter demonstrates the hypothesis generation capability of piTracer, by focusing on the ability of piTracer in predicting druggable genes in tumor cells.

Drug-resistance remains a monumental challenge in enabling effective cancer treatment, causing up to 90% of cancer-related deaths [77, 78]. A diverse range of mechanisms of drug resistance has been described that arise from anti-cancer treatments, such as the presence of compensatory metabolic pathways [79]. It is known that metabolomics provides a functional readout beyond the information covered by genetics technologies [80, 81]. For instance, while transcriptomics represents a noisy state of regulation, metabolomics data condenses complex and highly combinatorial genetic states into a discrete series of cascades the cell is able to operate in. Our hypothesis is that the metabolome enables us to discover cancer escape mechanisms, which when targeted, causes cancer cell death. Although significant efforts have been undertaken in this direction to better understand cancer-specific molecular mechanisms of resistance [77, 78, 82], reconstructing compensatory metabolic pathways have remained an onerous task. With its ability to automatically reconstruct biological pathways, we investigated whether we can apply piTracer on metabolomics data to find tumor survival pathways that could be targeted to induce lethality.

4.1 Methods

4.1.1 Target gene prioritization

We prioritized gene targets based on metabolic readouts of escape mechanisms after treatment with a primary drug using the following steps: (1) Given a metabolomics dataset before and after treatment with the original drug of interest, perform differential abundance analysis on the data. (2) Using piTracer, calculate the 10-shortest paths between all metabolite pairs in the dataset to create a "context network" ***C***, which is composed of only metabolites. (3) Assign weights ***w*** to metabolites in ***C***, with 1 for significant metabolites and 0 for the rest, which includes both non-significant metabolites and additional metabolites introduced in ***C*** via the 10-shortest paths calculations. (4) Remove directionality in ***C***. This step is required since metabolite fluxes can be affected either by upstream or downstream factors. (5) Remove PPI edges from piTracer's gene interaction network, i.e. STRING, to get a new network ***G***. We removed STRING edges due to the presence of many hub genes, which could obfuscate relevant gene-to-metabolite connections by connecting virtually all genes to all metabolites in the piTracer network. (6) Combine ***G*** with ***C***. (7) Calculate an inverse distance matrix ***l*** between all genes in ***G*** and all metabolites in ***C*** in the ***G*** + ***C*** network. The inverse distance is calculated as *1/D*, where *D* is the number of steps between a gene and a metabolite pair. ***l*** should have as rows genes in ***G*** and as columns metabolites in ***C***. (8) Create a list ***E*** that contains genes directly interacting with metabolites, i.e. enzymes in metabolic reactions, in the ***G*** + ***C*** network. (9) Assign gene weights, ***g*** = 1 - (# of ***E***s reached)/|***E***| for each gene, to all genes in ***G***. The assumption here is that perturbing genes that regulate or interact with many enzymes could have detrimental effects on metabolic paths not relevant to the biological system. ***g*** downweights these genes and will enable specific metabolic traces relevant to the system to be targeted. (10)

Calculate gene scores $\boldsymbol{S}$ = diag($\boldsymbol{g}$)$\boldsymbol{\iota}\boldsymbol{w}$. For each gene, $\boldsymbol{S}$ contains the sum of the inverse distances *1/D* between the gene and all significant metabolites it reaches in the $\boldsymbol{G} + \boldsymbol{C}$ network, weighted by the number of enzymes the gene reaches. A gene receives a high score if it is connected to many significant metabolites via metabolic fluxes rather than through the regulation of or interaction with many enzymes.

For our proof of concept, we picked the top 30 genes based on our calculated scores and further added pathway level annotations to each gene. To this end, we calculated a score for each pathway based on diag($\boldsymbol{g}$)$\boldsymbol{\iota}$. Each row of diag($\boldsymbol{g}$)$\boldsymbol{\iota}$ is a gene and each column is a weighted inverse distance score of the gene with a metabolite. For each pathway and each gene, we summed up this weighted inverse distance score between the gene and all metabolites in the pathway. Each sum represents the pathway score for each gene.

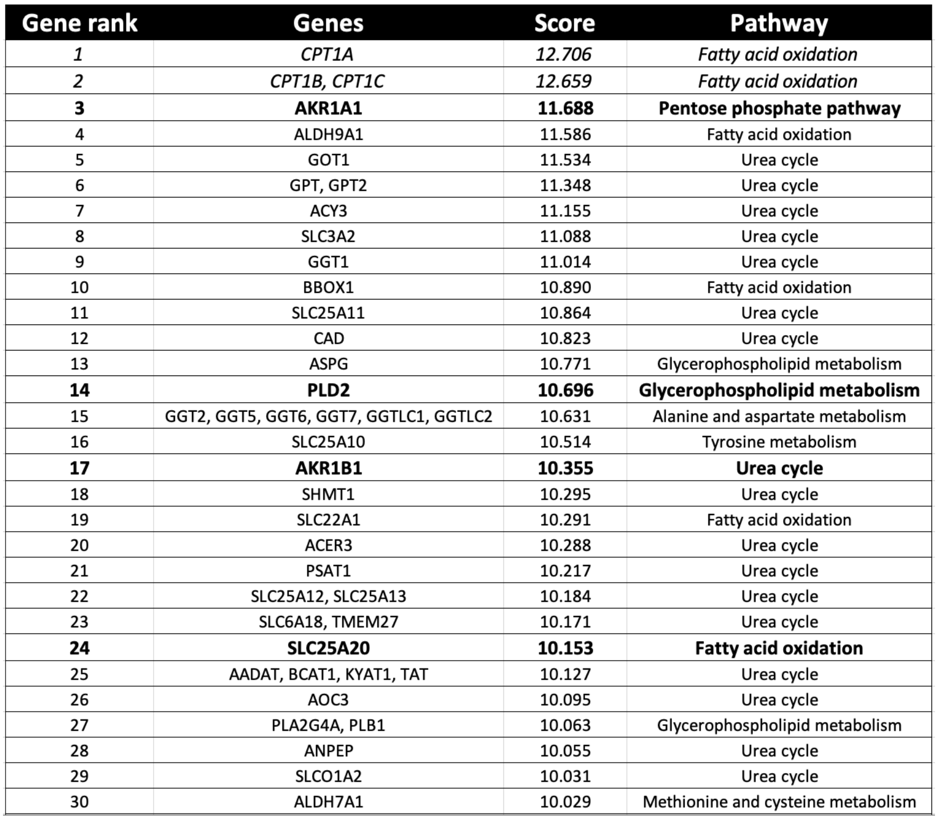
4.1.2 Experimental validation

Our collaboration partners, Dr. Anna Halama and Iman Achkar at Weill Cornell Medicine - Qatar, performed the validation experiments. All experiments described as follows were done in triplicates. They first measured the metabolic profile of a triple negative breast cancer cell line (MB-MDA-231) before and after treatments with glutaminase inhibitor C.968. Using Western blot, they verified the expression of our selected proteins, i.e. SLC25A20, PLD2, AKR1A1, and AKR1B1 in MB-MDA-231 before and after C.968 treatment. They then assessed cell viability after the treatment of MB-MDA-231 with C.968 and drugs targeting each of the selected proteins by MTT assay and microscopic pictures at 24h, 48h, and 72h post treatment. PLD2 was targeted by CAY10594, SLC25A20 was targeted by Ingenol Mebutate, AKR1A1 was targeted by Imirestat, and AKR1B1 was targeted by Ranirestat. The concentration used for C.968 was 10 µM, for CAY10594 was 0.001, 0.01, 0.1, 1, 10 µM while for Ingenol Mebutate, Imirestat, and Ranirestat was 0.01, 0.1, 1, 10, 100 µM. In a similar setting, they assessed cell viability after combination treatment with C.968 and drugs targeting the selected proteins. They used a concentration of 10 µM for C.968, 1 and 10µM for CAY10594, and 10 and 100µM for Ingenol Mebutate, Imirestat, and Ranirestat. They subsequently quantified MB-MDA-231 cell viability at 24h, 48h, and 72h post treatment.

4.2 Results

Based on a previous study [83] and as a proof of principle, we sought a drug to use in combination with glutaminase inhibitor C968 on the triple negative breast cancer cell line (MB-MDA-23) in order to induce cell death. We leveraged piTracer and metabolomics data from MB-MDA-23 before and after C968 treatment to calculate the 10-shortest paths between all pairs of metabolites in the dataset. The combination of all of these paths is a “context” network that contains potential escape mechanisms of MB-MDA-23. We then assigned a druggability score to each gene in the piTracer network. Briefly, we calculate gene scores based on how many significant metabolites a gene reaches in the context network, downweighted by the number of other enzymes it interacts with to reach significant metabolites (for details, see Methods 4.1.1). Our collaboration partners performed all subsequent validation experiments.

4.2.1 piTracer-based potential drug target candidates



Gene rank	Genes	Score	Pathway
1	*CPT1A*	*12.706*	*Fatty acid oxidation*
2	*CPT1B, CPT1C*	*12.659*	*Fatty acid oxidation*
3	**AKR1A1**	**11.688**	**Pentose phosphate pathway**
4	ALDH9A1	11.586	Fatty acid oxidation
5	GOT1	11.534	Urea cycle
6	GPT, GPT2	11.348	Urea cycle
7	ACY3	11.155	Urea cycle
8	SLC3A2	11.088	Urea cycle
9	GGT1	11.014	Urea cycle
10	BBOX1	10.890	Fatty acid oxidation
11	SLC25A11	10.864	Urea cycle
12	CAD	10.823	Urea cycle
13	ASPG	10.771	Glycerophospholipid metabolism
14	**PLD2**	**10.696**	**Glycerophospholipid metabolism**
15	GGT2, GGT5, GGT6, GGT7, GGTLC1, GGTLC2	10.631	Alanine and aspartate metabolism
16	SLC25A10	10.514	Tyrosine metabolism
17	**AKR1B1**	**10.355**	**Urea cycle**
18	SHMT1	10.295	Urea cycle
19	SLC22A1	10.291	Fatty acid oxidation
20	ACER3	10.288	Urea cycle
21	PSAT1	10.217	Urea cycle
22	SLC25A12, SLC25A13	10.184	Urea cycle
23	SLC6A18, TMEM27	10.171	Urea cycle
24	**SLC25A20**	**10.153**	**Fatty acid oxidation**
25	AADAT, BCAT1, KYAT1, TAT	10.127	Urea cycle
26	AOC3	10.095	Urea cycle
27	PLA2G4A, PLB1	10.063	Glycerophospholipid metabolism
28	ANPEP	10.055	Urea cycle
29	SLCO1A2	10.031	Urea cycle
30	ALDH7A1	10.029	Methionine and cysteine metabolism

Table 2. Top 30 piTracer-predicted target gene candidates. Rows in bold indicate genes that were selected for experimental validation. Italicized rows, i.e. gene ranks 1 and 2, indicate genes that were successfully validated independently [84]. Each pathway represents the top scoring pathway for each target gene.

We screened for potential drug targets out of the 7,253 genes in our modified piTracer network. Based on our gene scores, we shortlisted the top 30 genes (Table 2). Together with our collaboration partners, we then selected 4 genes based on their involvement in different metabolic pathways inferred from the data, druggability, and drug availability. The genes were: (1) PLD2, a phospholipase which catalyzes the hydrolysis of phosphatidylcholine to produce phosphatidic acid and choline and also involved in lipid pathways. (2) SLC25A20 which encodes a protein important for fatty acid oxidation. (3) AKR1A1 and (4) AKR1B1 which are both part of the aldo-keto reductase family 1. AKR1A1 is involved in the reduction of biogenic and xenobiotic aldehydes, while AKR1B1 catalyzes the reduction of glucose to sorbitol. We used these 4 genes to select potential drugs that could have a potential synergistic effect with C968. Note that the 3 top ranking genes, CPT1A, CPT1B, and CPT1C were validated in an independent study [84].

4.2.2 Verifying the presence of target gene products

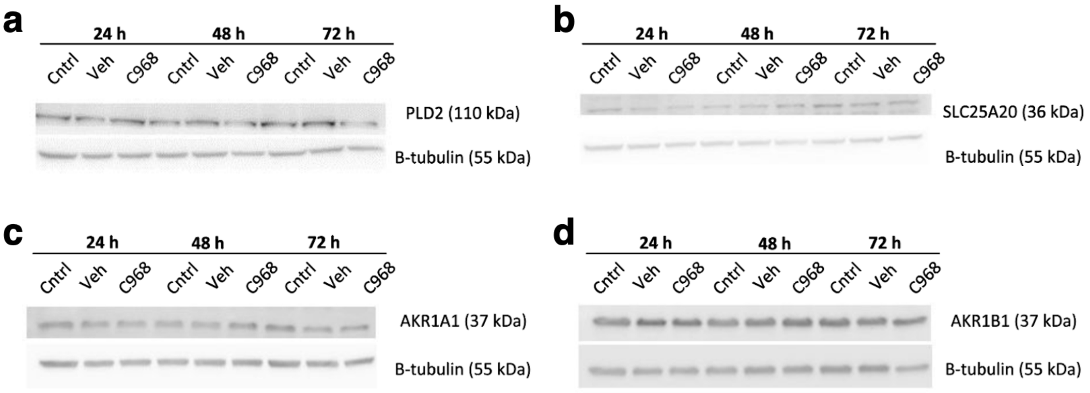


Figure 4.1. Western blot of the 4 predicted target genes. Blots for (**a**) PLD2, (**b**) SLC25A20, (**c**) AKR1A1, and (**d**) AKR1B1 vs. β-tubulin at different timepoints. All genes are expressed in MDA-MB-231. Cntrl stands for control and Veh for vehicle.

Prior to drug validation experiments, our collaboration partners checked whether the proteins of the 4 selected genes are expressed in MDA-MB-231 cells under normal conditions and after treatment with C968 (Figure 4.1). All proteins were present and stable in the corresponding Western blots.

4.2.3 Single drug treatment of the MDA-MB-231 cell line

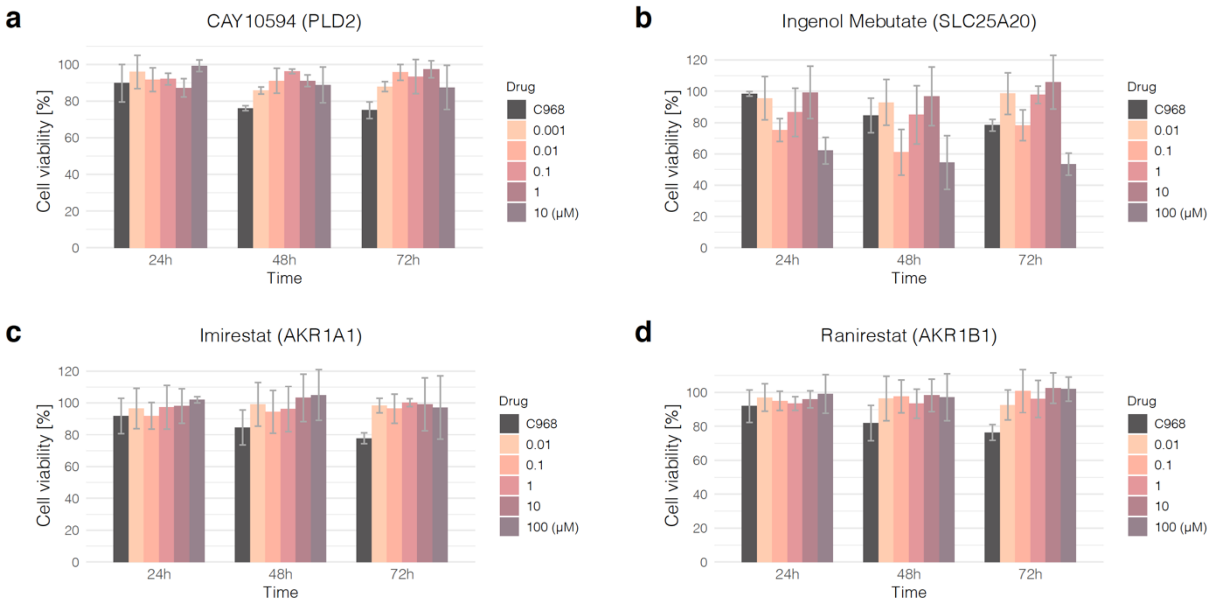


Figure 4.2. Cell viability assay results for single drug treatment of the MDA-MB-231. (**a**) PLD2 was targeted by CAY10594, (**b**) SLC25A20 by Ingenol Mebutate, (**c**) AKR1A1 by Imirestat, and (**d**) AKR1B1 by Ranirestat. Treatment with single drugs had minimal impact on MDA-MB-231. The blue line indicates C968 treatment of the cell line and the different shades of black represent different concentrations of drugs for a target gene.

Our collaboration partners then tested whether treatment of MDA-MB-231 with drugs targeting only PLD2, SLC25A20, AKR1A1, and AKR1B1, and without the effects of glutaminolysis inhibition, affected cell viability. The impact of each drug on the viability of MDA-MB-231 was tested with 5 different concentrations (Figure 4.2). They inhibited PLD2 using CAY10594 (Figure 4.2a), a PLD2 inhibitor typically used in ameliorating acetaminophen-induced acute liver injury. They used Ingenol Mebutate, an inducer of cell death, to target SLC25A20 (Figure 4.2b). Lastly, they used aldose reductase inhibitors Imirestat and Ranirestat to target AKR1A1 and AKR1B1, respectively (Figures 4.2c and 4.2d). These drugs are commonly used in the treatment of diabetes. The selected components as a single agent had minimal impact on MDA-MB-231.

4.2.4 Combination drug treatment of MDA-MB-231

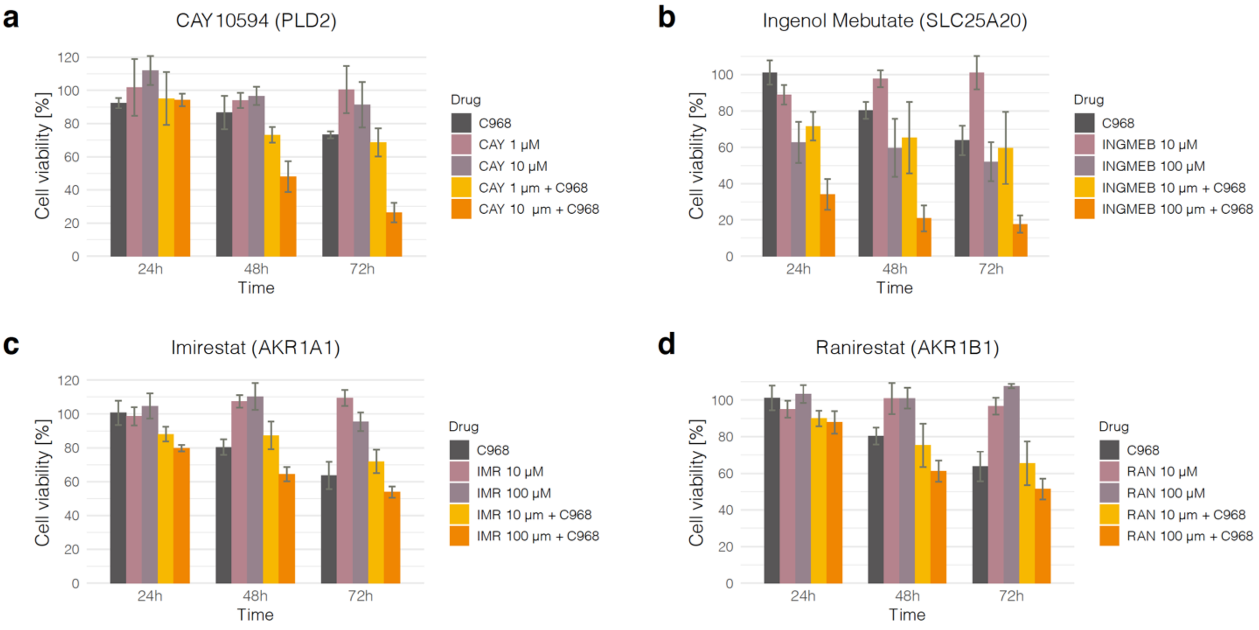


Figure 4.3. Cell viability assay results for combination drug treatment of MDA-MB-231. Treatment of the cell line with C968 in combination with (**a**) CAY10594 to target PLD2, (**b**) Ingenol Mebutate to target SLC25A20, (**c**) Imirestat to target AKR1A1, and (**d**) Ranirestat to target AKR1B1. A combination of C968 and CAY10594 or C968 and Ingenol Mebutate exert the strongest decrease in cell viability. A concentration of 10 μM C968 was selected for all experiments based on a previous study [83].

Lastly, it was tested whether C968 treatment of MDA-MB-231 in combination with drugs targeting PLD2, SLC25A20, AKR1A1, and AKR1B1 impacted cell viability (Figure 4.3). The impact of each of the drug combinations on the viability of MDA-MB-231 was tested with two different concentrations of the gene-specific drugs and one concentration of C968, i.e. 10 μM [83]. Combination treatment with all selected drugs and C968 impacted the viability of MDA-MB-231. The most significant cell viability decrease was observed for C968 and 10 µM CAY10594 (PLD2) from 100% to 20% in 72h (Figure 4.3a) and C968 and 100 µM Ingenol Mebutate (SLC25A20) from 100% to 20% in 24h (Figure 4.3b). Taken together, these results show that by predicting relevant gene targets, piTracer successfully aided in screening for a drug to use in combination with C968 in order to induce cell death in MB-MDA-23.

4.3 Summary & Discussion

The ability of piTracer to predict drug targets, and by extension streamline drug screening, is remarkable given that it solely relies on metabolomics data to generate predictions. The first crucial step for drug screening or drug repurposing which is time intensive, expensive, and requires extensive domain expertise, is finding a viable target for a specific indication of interest [85–87]. In cancer, escape mechanisms are a desirable drug target since cancer cells rely on these for survival. Many escape mechanisms have metabolic adaptations at their core [5]. Thus, measuring the metabolome is crucial in elucidating these compensatory pathways. We used piTracer and metabolomics data to reconstruct escape mechanisms in the triple negative breast cancer cell line MDA-MB-231. Out of 7,253 genes, we prioritized 30 genes using piTracer compensatory pathway reconstructions, selected 4 genes, and validated them experimentally. Targeting either of the 4 genes, i.e. PLD2, SLC25A20, AKR1A1, or AKR1B1 with a drug in combination with glutaminase inhibitor C968 resulted in a decrease in cell viability of MDA-MB-231. In particular, targeting either PLD2 or SLC25A20 significantly reduced the viability of the cell line. In an identical setting, our top 3 predicted target genes, CPT1A, CPT1B, and CPT1C, which we excluded from our experiments, were previously validated by others [84]. Remarkably, out of the many potential drug targets, i.e. 7,253 genes, piTracer enabled us to reduce this to a few genes that were true drug targets. Note that this approach is not limited to cancer and may be applied in other settings where the metabolome is measured between two conditions, e.g. healthy vs. unhealthy cells. These results indicate that piTracer can rapidly and reliably generate viable drug target candidates relevant for a disease, which is invaluable in the efficient screening and repurposing of drugs.

Chapter 5: Variational autoencoders learn universal latent representations of metabolomics data

Current well-known dimensionality reduction techniques used for metabolomics data rely primarily on linear assumptions, and are therefore unable to detect possible nonlinearities [38] that arise from functional interactions between metabolites. Moreover, the application of deep learning models, such as Variational Autoencoders (VAEs), that are well suited in capturing these nonlinearities by implicitly modeling feature interactions, has significantly lagged behind in the metabolomics field [49].

To this end, in this chapter, we trained a VAE model on 217 metabolite measurements in 4,644 blood samples from the TwinsUK study [88] and evaluated our model performance in comparison to a linear PCA model (Figure 5.1a). To investigate the biological relevance of the learned VAE and PCA latent dimensions, we employed the Shapley Additive Global Importance (SAGE) method [89], which determines the contribution of each input to each latent dimension. We calculated SAGE values at different granularities, i.e. metabolites, *sub-pathways*, and *super-pathways* (Figure 5.1b). We then applied the model on three additional blood metabolomics datasets to test its ability to recover disease phenotypes in unseen datasets: Type 2 Diabetes diagnosis in The Qatar Metabolomics Study on Diabetes (QMDiab, n = 358), therapy response in an acute myeloid leukemia dataset (AML, n = 85), and schizophrenia diagnosis in a third validation dataset (n = 207) (Figure

5.1c). We provide pre-trained VAE and PCA models as well as scripts to reproduce our analysis at https://github.com/krumsieklab/mtVAE.

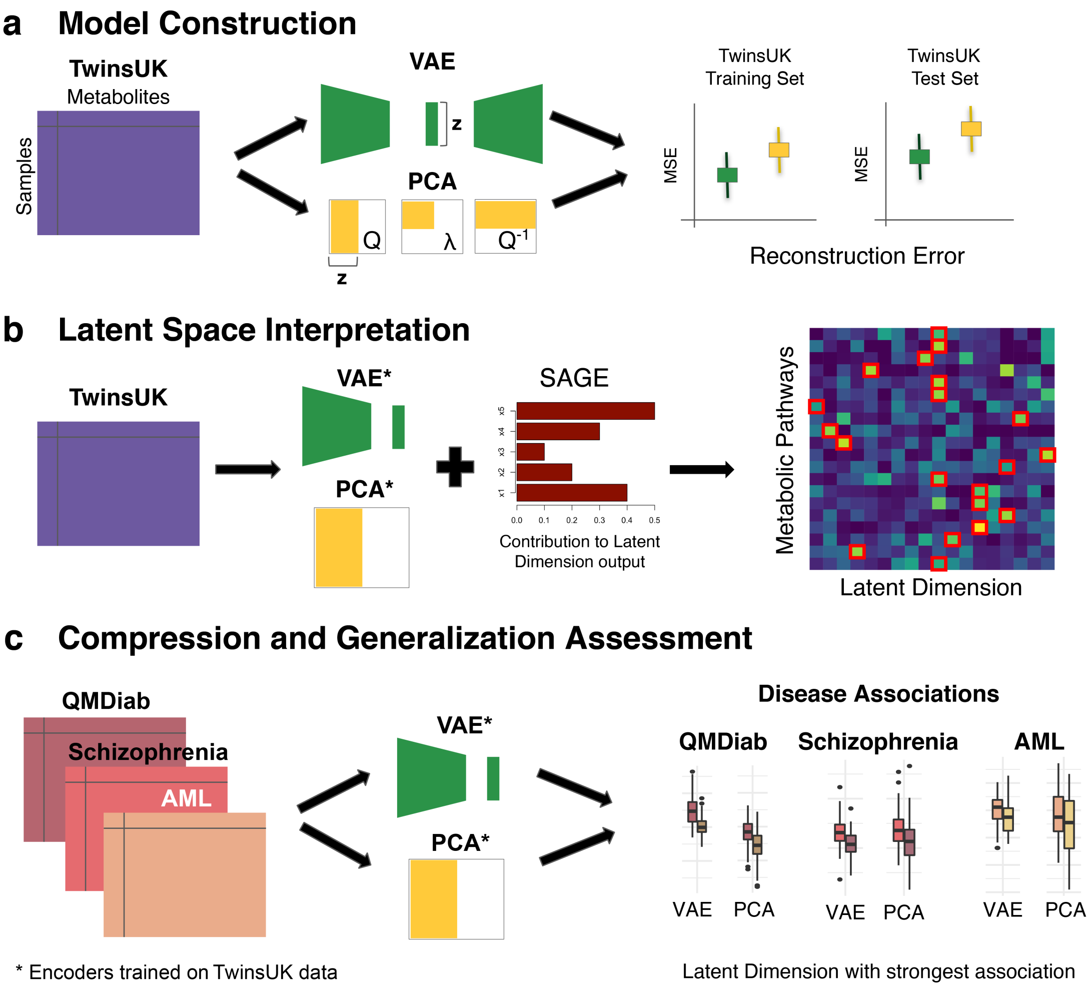


Figure 5.1. Overview of our approach. (**a**) VAE and PCA models were trained using training and test partitions in the TwinsUK dataset (*n*=4,644 samples, *p*=217 metabolites). Model performance was then evaluated using Mean Squared Error (MSE) of correlation matrix reconstruction. (**b**) The SAGE method was applied to the models to calculate the contribution of individual metabolites, sub-pathways and super-pathways to each latent dimension. (**c**) QMDiab (*n* = 358), AML (*n* = 85), and Schizophrenia (*n* = 207) datasets were encoded using VAE and PCA models trained on the TwinsUK data. Latent dimensions of each model were then associated with disease phenotypes.

5.1 Methods

5.1.1 Datasets

The TwinsUK registry is a population-based study of around 12,000 volunteer twins from all over the United Kingdom. The participants have been recruited since 1992 and are predominantly female, ranging in age from 18 to 103 years old. Study design, sampling methods, and data collection have been described elsewhere [88]. For this thesis, we included data from 4,644 twins (4,256 females, 388 males), the subset of TwinsUK for which plasma metabolomics measurements were available. Ethical approval was granted by the St Thomas' Hospital ethics committee and all participants provided informed written consent.

The QMDiab study was conducted between February and June of 2012 at the Dermatology Department of Hamad Medical Corporation (HMC) in Doha, Qatar. The study population was between the age of 23 and 71, predominantly of Arab, South Asian, and Filipino descent. Data collection and sampling methods have been previously described elsewhere [90]. For this thesis, we included plasma data of 358 subjects (176 females, 182 males; 188 diabetic, 177 non-diabetic). The study was approved by the Institutional Review Boards of HMC and Weill Cornell Medicine-Qatar (WCM-Q). Written informed consent was obtained from all participants.

For the schizophrenia analysis, metabolomics samples were taken from an antipsychotics study conducted in Qatar between December 2012 and June 2014 [91]. A total of 226 participants between the ages of 18 and 65 years of age were recruited, predominantly of Qatari and Arab descent. For this thesis, we included plasma metabolomics measurements from 207 subjects (84 females, 142 males; 102

schizophrenic, 105 non-schizophrenic). Approval for the study was obtained from the HMC and WCM-Q Institutional Review Boards, and all participants provided written informed consent.

The cohort of patients with *de novo* acute myeloid leukemia (AML) comes from the ECOG (Eastern Cooperative Oncology Group) E1900 trial [92]. This study was conducted between December 2002 and March 2009, recruiting 657 patients with AML between the ages of 17 and 60. A subset of these patients had follow-up profiling to determine their response to therapy. For this thesis, we include the serum metabolomics measurements of 85 subjects of which 43 responded to therapy and 42 did not (34 females, 51 males). The study was approved by the institutional review board at the National Cancer Institute and each of the study centers, and written informed consent was provided by all patients.

5.1.2 Metabolomics measurements and metabolite annotations

Non-targeted liquid chromatography/mass spectrometry (LC/MS)-based metabolomic profiling for all four cohorts was performed on the Metabolon platform as previously described [93]. Notably, the AML dataset was based on serum samples, while TwinsUK, QMDiab, and schizophrenia metabolomics was run on plasma samples. However, previous research has shown that these two sample types are comparable, as shown by high correlations and good reproducibility between plasma and serum measurements in the same blood sample [94].

For each metabolite measured on the Metabolon platform, a super-pathway and sub-pathway annotation was provided. For super-pathways, we have nine annotations referring to broad biochemical classes, namely “Amino acid”, “Carbohydrate”, “Cofactors and vitamins”, “Energy”, “Lipid”, “Nucleotide”, “Peptide”, “Xenobiotics”, and “Unknown”. Note that “Unknown” is assigned to unidentified metabolites. Furthermore, we have 54 sub-pathway which represent more functional metabolic processes, such as “Carnitine metabolism”, “TCA Cycle”, and “Phenylalanine and Tyrosine Metabolism”.

5.1.3 Data processing and normalization across datasets

For each dataset, metabolite levels were scaled by their cohort medians, quotient normalized [95] and then log-transformed. Samples with more than 30% missing metabolites and metabolites with more than 10% missing samples were removed. Missing values were imputed using a k-nearest neighbors imputation method [96]. Datasets with BMI measurements (Schizophrenia, QMDiab, and Twins) were corrected for that confounder and then mean-scaled. 217 metabolites were overlapping between the 4 datasets and were kept for further analysis.

Semi-quantitative, non-targeted metabolomics measurements are inherently challenging to compare across datasets due to heterogeneity between studies. This prevents any machine learning model from being transferable from one study to the other. To ensure comparability, datasets were normalized using a uniform group of participants as a reference set. This group was selected as follows: Male, within a 20 year age range (30-50 for TwinsUK, QMdiab, and schizophrenia, 40-60 for AML due to low sample size of younger participants), BMI between 25 and 30 (not available for AML data, thus not filtered for that dataset), and in the respective control group. Each metabolite in each dataset was then scaled by the mean and standard deviation of their respective uniform sample groups. The assumption of this approach is that the uniform group of reference participants has the same distributions of metabolite concentrations.

5.1.4 Variational Autoencoders

We trained our VAE model as follows. We split TwinsUK data into 85% training and 15% test sets. We then fixed our VAE architecture to be composed of an input/output layer, an intermediate layer which contains the only nonlinear activation functions in the model, and latent layers. The latent layers consist of a mean vector $\boldsymbol{\mu}$ and a standard deviation vector $\boldsymbol{\sigma}$ which parametrize the latent space $\mathbf{z}$. $\mathbf{z}$ is constructed by the simultaneous learning of the $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ encoder through the use of a reparameterization trick that enables back propagation during training [45]. For a $\mathbf{z}$ with a latent dimensionality $\boldsymbol{\mu}$ is a of length d. The d x d covariance matrix $\boldsymbol{\Sigma}$ of the underlying Gaussian is assumed to be diagonal (i.e. no correlation across latent dimensions), allowing the covariance matrix to be represented by a single vector $\boldsymbol{\sigma}$ of length d.

For the parameter fitting procedures, all weights were initialized using Keras' default model weight initialization, Glorot uniform [97], and leaky rectified linear units (ReLUs) [98] were used for nonlinear activation functions. The VAE models were trained for 1000 epochs using MSE loss for sample reconstruction and a batch size of 32.

To select the latent dimensionality d of our VAE model, we initially fixed this value to d = 50. We then optimized the model hyperparameters using Keras Tuner [99] and TwinsUK training set and identified the following optimized values:

Intermediate layer dimensionality = 200, learning rate = 0.001, and Kullback-Leibler (KL) divergence weight = 0.01. Using these hyperparameters, we then optimized *d* by calculating the reconstruction MSE of the correlation matrix (CM-MSE) of metabolites for *d* = 5, 10, 18, 30, 40, 60, 80, 100, 120, 160, and 200 on the TwinsUK test set. Our final model consisted of a 217 dimensional input/output layer (the number of metabolites in our datasets), a 200 dimensional intermediate layer, and an 18 dimensional latent layer. For all sample embeddings in this thesis, we used their respective $\boldsymbol{\mu}$ values.

All models were computed on a deep learning-specific virtual machine running on Google Compute Engine with two NVIDIA Tesla K80 GPU dies and 10 virtual CPUs.

5.1.5 PCA embedding and reconstructions

We used PCA with $d = 18$ latent dimensions as a baseline model. On the mean-centered TwinsUK data matrix with $n = 3{,}947$ samples (rows) and $k = 217$ metabolites (columns), we calculated the rotation matrix $\boldsymbol{Q}$, a k x k matrix of eigenvectors ordered by decreasing magnitudes of eigenvalues. To embed a new m x k dataset $\boldsymbol{X}$ with m samples into the m x d PCA latent space $\boldsymbol{A}$, we first calculated $\boldsymbol{XQ} = \boldsymbol{A}$ and subsetted to the first d columns, denoted by $\boldsymbol{A}_{*,d}$. To simulate the process of encoding and decoding in PCA for dataset $\boldsymbol{X}$, we calculated the reconstructed dataset as $\hat{\boldsymbol{X}} = \boldsymbol{A}_{*,d}\boldsymbol{Q}^{-1}{}_{d,*}$.

5.1.6 Model assessments

We assessed our PCA and VAE models using sample reconstruction mean squared error (MSE) and metabolite-wise correlation matrix MSE (CM-MSE). We calculated CM-MSE by first computing the metabolite-wise correlation matrix of an input dataset and reconstructed input dataset. Afterwards, we calculated the MSE between the upper triangular matrix of the two symmetric correlation matrices.

To calculate a confidence interval for both MSE and CM-MSE between our input and reconstructed data, we randomly sampled the same samples with replacement from the two datasets and then calculated MSE and CM-MSE. We performed this for 1000 iterations.

5.1.7 Model interpretation

In order to interpret each latent dimension for our VAE and PCA models, we calculated Shapley Additive Global Importance (SAGE) values [89] for metabolites, sub-pathways, and super-pathways. Briefly, SAGE is a model-agnostic method that quantifies the predictive power of each feature in a model while accounting for interactions between features. This is achieved by quantifying the decrease in model performance when combinations of the other model variables are removed. Since there are exponentially many combinations of variables, the current approach is to sample that space of combinations sufficiently. For each of the tested combinations, a loss function, such as MSE, is used to quantify the decrease in performance compared to the model output (here the latent layer) computed using the full model. Then, the mean of all MSEs is calculated, which represents the contribution of the variable to the latent dimension. To calculate pathway-level SAGE values, metabolites were grouped into pathways and each pathway was treated as a single variable. For each of our VAE and PCA models, we ran SAGE using our TwinsUK test set with default parameters, e.g. marginal sampling size of 512, as suggested by Covert, et al. (2020) [89]. We used the SAGE code from https://github.com/iancovert/sage.

5.2 Results

5.2.1 VAE model construction and fitting

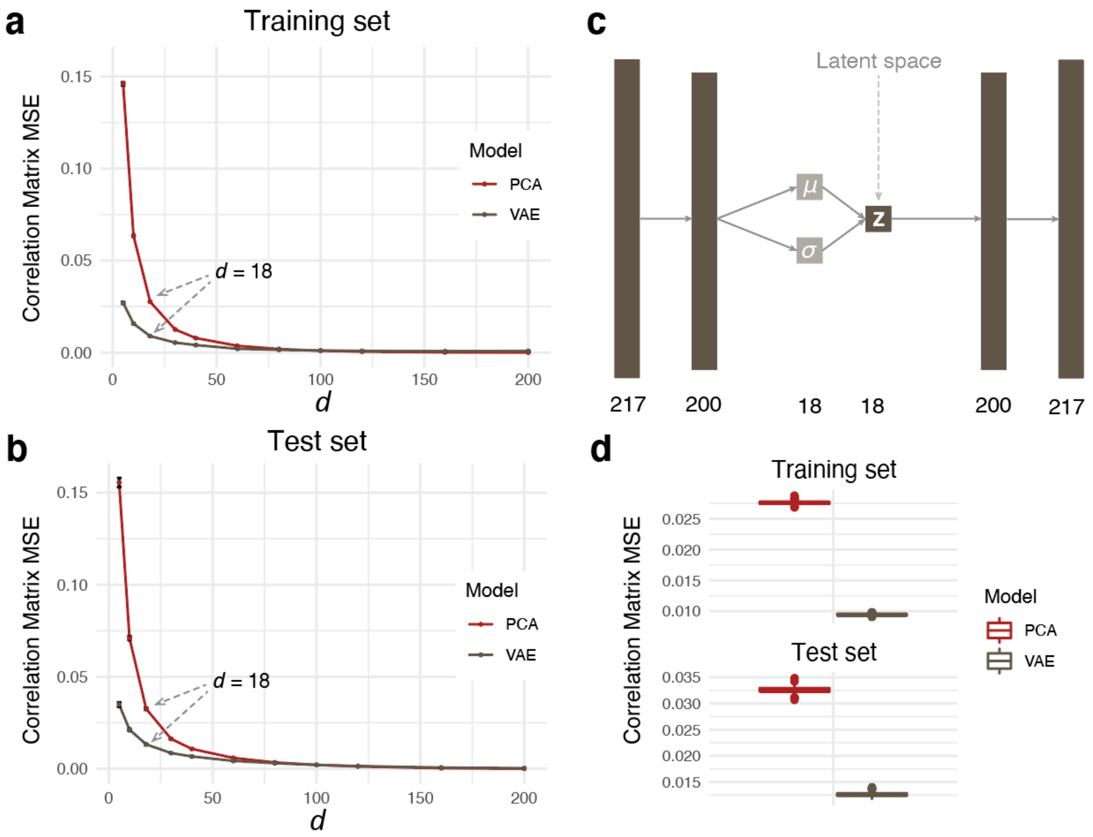


Figure 5.2. VAE and PCA model construction on the TwinsUK dataset. (**a**) Training and (**b**) test set metabolite correlation matrix reconstruction for a range of latent dimensionality values *d*. The slope of the VAE curve plateaus after *d* = 18. Error bars correspond to one standard deviation from bootstrapping. (**c**), Final VAE architecture, where μ is the mean vector and σ is the standard deviation vector that generates the latent space **z**. (**d**), Reconstruction MSE for latent dimensionality *d* = 18 on training (top) and test sets (bottom). The VAE preserved feature correlations substantially better than PCA.

Our VAE architecture consisted of an input/output layer, an intermediate layer and a latent layer. We split the TwinsUK cohort into an 85% training and a 15% test set, and the training set was used to optimize the hyperparameters in the VAE model. Keras Tuner [99] identified the following optimal hyperparameters: Intermediate layer dimensionality = 200, learning rate = 0.001, and Kullback-Leibler (KL) divergence weight = 0.01. With these parameters fixed, we optimized the dimensionality d of the latent layer $\mathbf{z}$ by calculating the reconstruction MSE of the correlation matrix (CM-MSE) of metabolites (Figures 5.2a+b). We observed that the CM-MSE curve plateaus after d = 18, indicating that increasing the latent dimensionality beyond this value only marginally improves the models. The final architecture of the model consisted of a 217 dimensional input/output layer (the number of metabolites in our datasets), a 200 dimensional intermediate layer, and an 18 dimensional latent layer (Figure 5.2c).

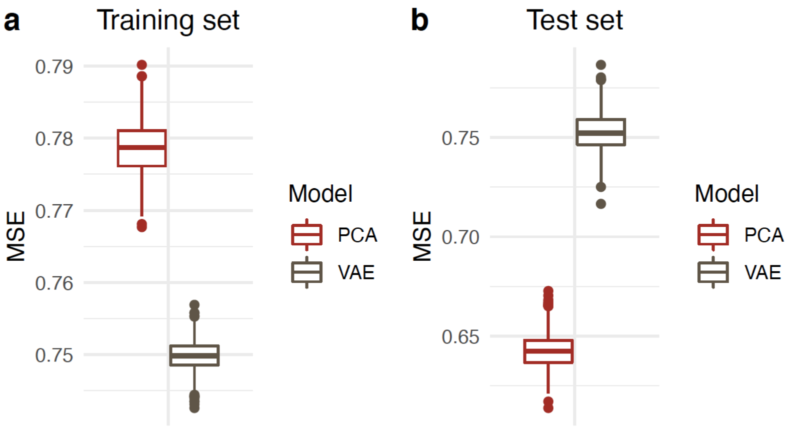


Figure 5.3. Sample reconstruction MSE. (**a**) TwinsUK training and (**b**) test set sample reconstruction MSE for latent dimensionality d = 18. VAE has a lower reconstruction error in the training set. However, PCA has a lower reconstruction MSE in the train set, implying that PCA performs better at sample reconstruction.

We used principal component analysis (PCA) as a baseline model to compare the VAE to a linear latent variable embedding method. To this end, we fitted a PCA on the TwinsUK train data and extracted the first d = 18 dimensions, i.e. principal components. While PCA reconstructs the data matrix better (Figure 5.3), the VAE outperforms PCA in terms of correlation matrix reconstruction via CM-MSE in both the TwinsUK train and test set (Figure 5.2d). These results suggest that while the VAE does not reconstruct the original data matrix precisely, it is superior at preserving metabolite correlations compared to PCA.

5.2.2 Interpretation of VAE latent space dimensions in the context of metabolites and pathways

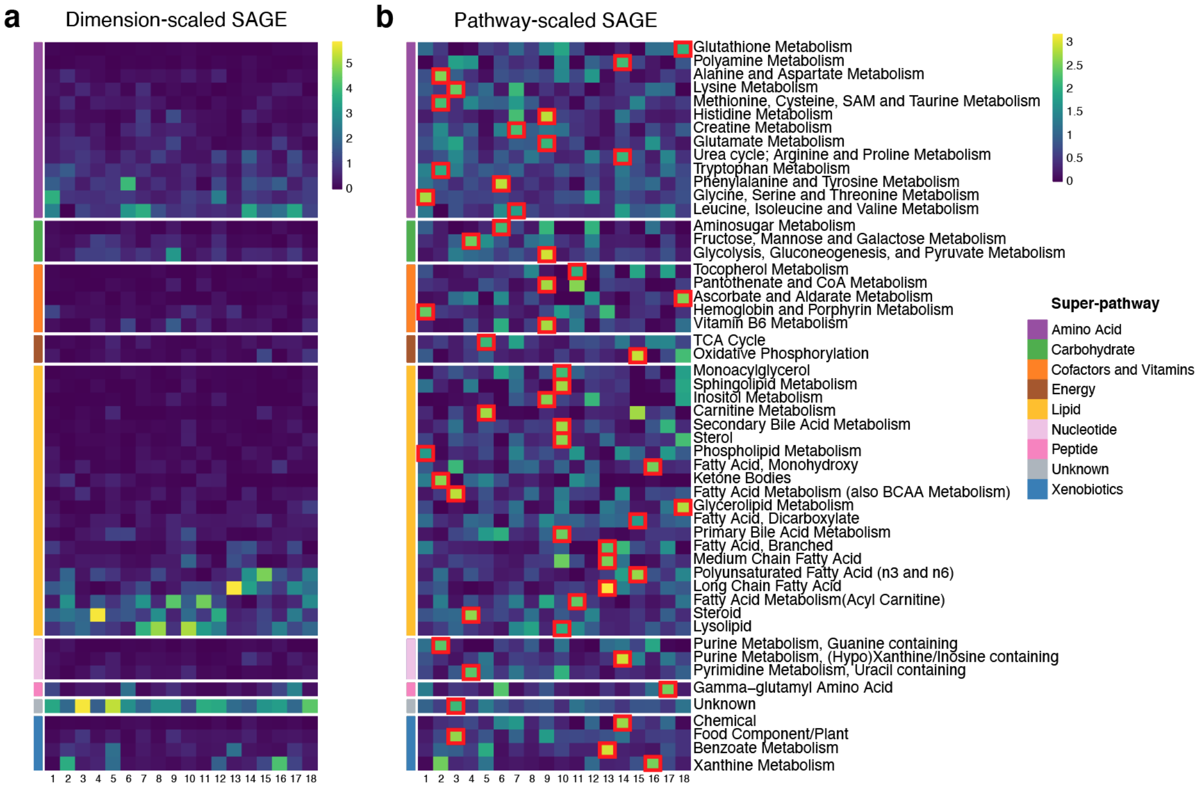


Figure 5.4. Sub-pathway-level SAGE values for the VAE latent dimensions. (**a**) SAGE values were scaled by dimension, i.e. set to standard deviation 1 for each column in the matrix. This highlights

pathways that contribute the most to each dimension. Lipid and amino acid super-pathways showed the highest values for to most dimensions, which can likely be attributed to the high number of metabolites in those pathways. (**b**) SAGE values were scaled by pathway, i.e. set to standard deviation 1 for each row in the matrix. This highlights dimensions that contribute to a pathway the most. Taking into consideration the largest scaled SAGE values per pathway (red square marks), almost all sub-pathways are represented by unique dimensions. The combination of these key subpathways of a dimension outlines the distinct cellular mechanisms a dimension encodes.

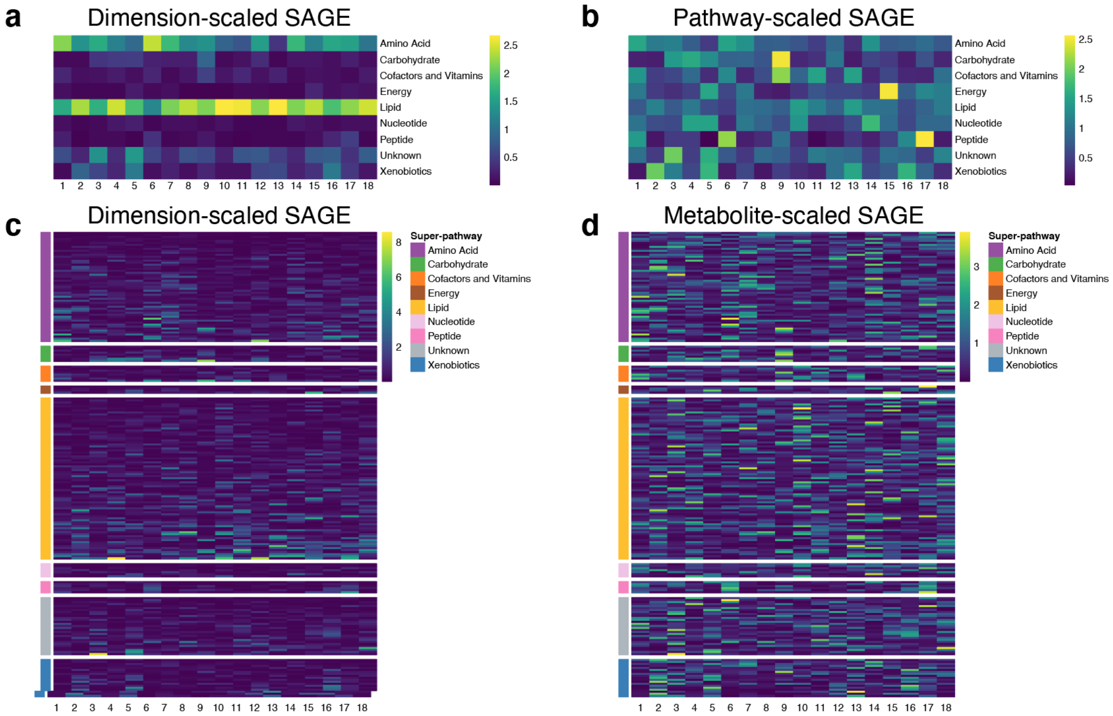


Figure 5.5. Scaled SAGE value heatmaps for VAE. (**a**) super-pathway SAGE values are scaled by dimension to highlight super-pathways that contribute the most to a dimension. The VAE heatmap indicates that in most dimensions, lipids contribute the most to each dimension due to their size. (**b**) super-pathway SAGE values are scaled by super-pathway to highlight dimensions that represent a super-pathway the most. Taking into consideration the highest scaled SAGE values, almost all super-pathways are represented by a unique dimension. (**c**) absolute metabolite SAGE values are scaled by dimension to highlight metabolites that contribute the most to a dimension. (**d**) absolute metabolite SAGE values are scaled by metabolite to highlight dimensions that represent a metabolite the most.

We evaluated the composition of all latent dimensions in the context of metabolic pathways. For each metabolite in our dataset, a “sub-pathway” and “super-pathway” annotation was available (see Methods 5.3.2). Sub-pathways refer to biochemical processes such as TCA cycle and sphingolipid metabolism, while super-pathways are broad groups such as lipid and amino acid. To provide insights into the processes represented by different VAE dimensions, we computed SAGE scores, a measure of model feature relevance, at the level of metabolites, sub-pathways and super-pathways (Figure 5.4 and Figure 5.5).

The VAE sub-pathway heatmap (Figure 5.4a) shows that nearly all dimensions have major contributions by lipid and amino acid super-pathways. The prevalence of the two super-pathways can be attributed to the fact that those groups contain the largest number of metabolites in the dataset. Note that we deliberately ignored the “Unknown” molecule group, which refers to unidentified metabolite that could originate from any pathway.

Inspecting the SAGE values in the other direction, almost all sub-pathways are predominantly represented by a single VAE dimension that captures the respective pathway the most (Figure 5.4b, red square marks). For instance, “glycolysis, gluconeogenesis and pyruvate metabolism” and other functionally related sub-pathways of central carbon metabolism are represented by VAE dimension 9. Another interesting example is VAE dimension 15, which captures functionally related sub-pathways that involve essential mitochondrial processes, such as oxidative phosphorylation, dicarboxylic fatty acids, and n3 and n6 polyunsaturated fatty acid metabolism. Taken together, these results show that VAE latent dimensions capture a complex mix of functionally related sub-pathways, thus capturing major metabolic processes in the dataset.

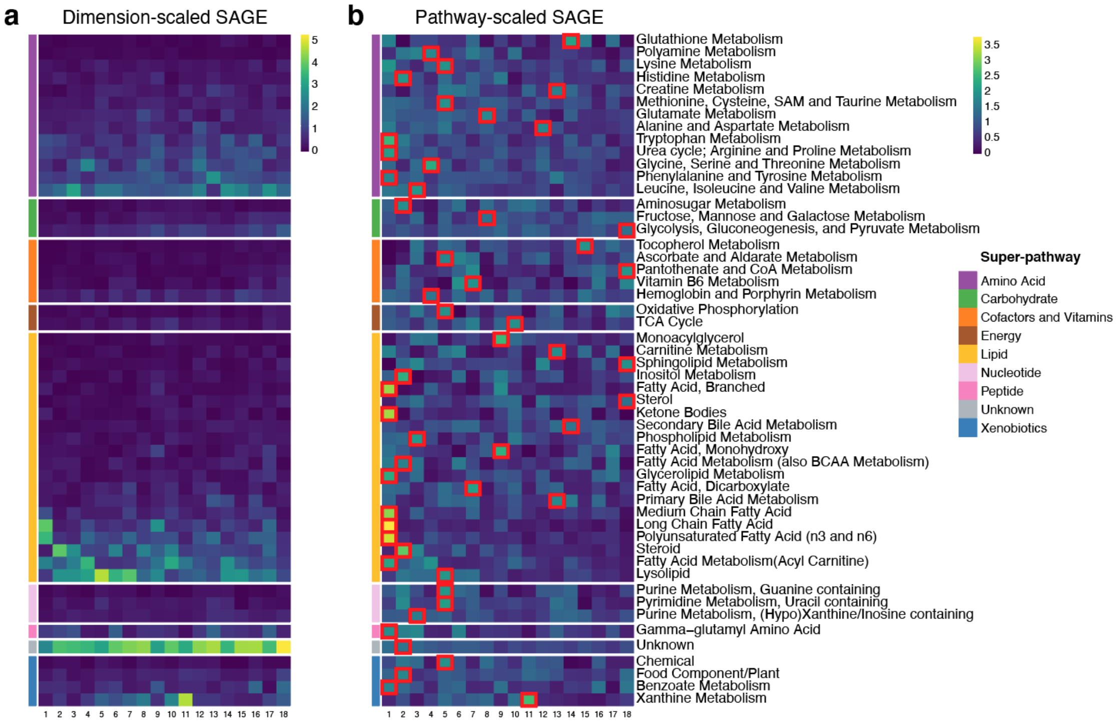


Figure 5.6. Sub-pathway-level SAGE value heatmaps for PCA. (**a**) SAGE values are scaled by dimension to highlight pathways that contribute the most to a dimension. The lipid and amino acid super-pathways contribute the most to the dimensions. (**b**) SAGE values are scaled by pathway to highlight dimensions that represent a pathway the most. Taking into consideration the largest scaled SAGE values, dimension 1, the first principal component, simultaneously represents many sub-pathways. Red squares indicate top 1 dimensions that represent a sub-pathway. Sub-pathways concentrate on the first few dimensions, especially on dimensions 1-3.

In contrast, PCA dimensions 1 to 3, which by construction represent the highest linear variations in the data, nonspecifically represents various sub-pathways (Figure 5.6). With the exception of PCA dimensions 4 and 5, the remaining dimensions contain primarily unrelated sub-pathways.

5.2.3 VAE latent space captures signals in unseen diabetes, cancer, and schizophrenia metabolomics datasets

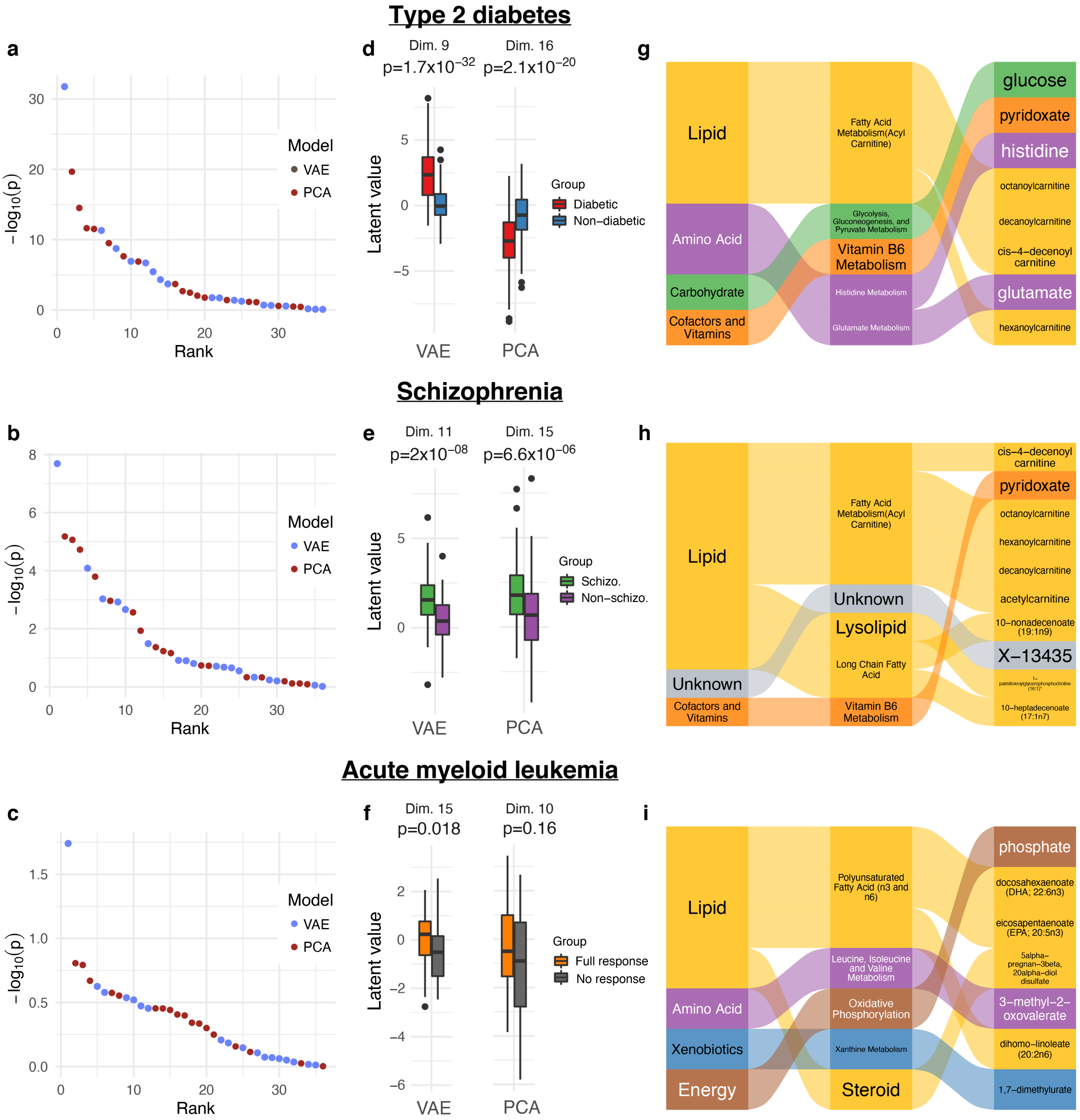


Figure 5.7. VAE latent space associations with clinical outcomes. (**a**), (**b**), (**c**), Sorted -$\log_{10}$(p-value) for all VAE and PCA dimensions for the type 2 diabetes, schizophrenia and AML datasets, respectively. The highest scoring VAE dimensions showed lower p-values than the highest scoring PCA dimensions for all datasets. (**d**), (**e**), (**f**), Latent space dimensions with the lowest p-values for the three datasets. (**g**),

(**h**), (**i**), Contributions of super-pathways, sub-pathways and metabolites to the highest scoring VAE latent dimensions, determined by SAGE values. All dimensions are driven by lipid metabolism and a mixture of other super-pathways, with differing sub-pathways contributing to the different dimensions. p = p-value. Schizo. = schizophrenic.

We investigated whether VAE latent dimensions learned on the TwinsUK data contained information that is generalizable to other datasets. To this end, we computed latent dimensions in three clinical datasets, type 2 diabetes, schizophrenia, and acute myeloid leukemia (AML) using the encoders from the TwinsUK data. For each VAE and PCA latent dimension, we performed a two-sided t-test between diabetic vs. non-diabetic individuals, schizophrenic vs. non-schizophrenic individuals, and full response vs. not in an AML clinical trial, respectively. Across all datasets, the respective best performing VAE dimensions associated substantially stronger with the patient groups than any of the PCA dimensions (Figures 5.7a-f). To better understand the driving factors of these associations in the VAE, we ranked pathways and metabolites by their calculated SAGE values (Figures 5.7g-i). The strength of associations between VAE dimensions and disease parameters were comparable to single metabolite associations. 1,5-anhydroglucitol (1,5-AG) associates with type 2 diabetes with a p $=1.14\text{x}10^{-35}$ (VAE dimension 9 $p=1.7\text{x}10^{-32}$), beta-hydroxyisovalerate associates with schizophrenia with a $p=1.48\text{x}10^{-9}$ (VAE dimension 11 $p=2.0\text{x}10^{-8}$), and trans-4-hydroxyproline associates with AML with a $p=5\text{x}10^{-3}$ (VAE dimension 15 p=0.018). However, unlike the VAE dimensions, these univariate associations do not represent system-level mechanisms related to the diseases.

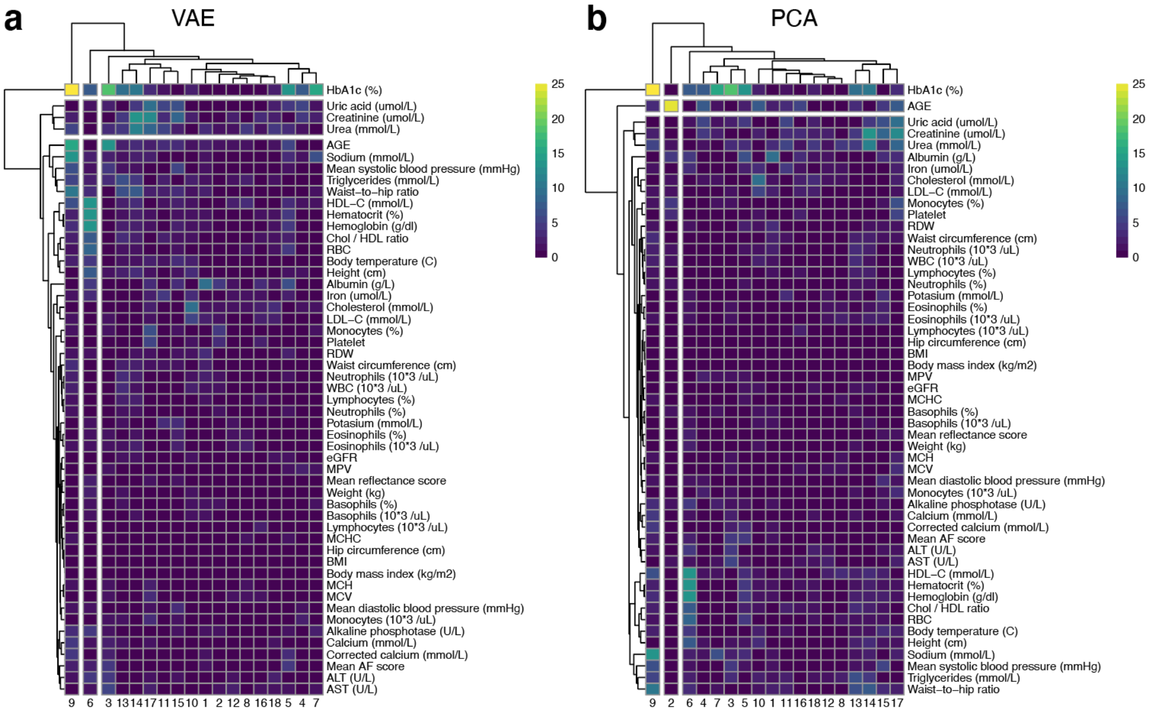


Figure 5.8. Association heatmap between (**a**) VAE and (**b**) PCA latent dimension values and clinical variables in QMDiab. Each latent dimension is associated with different combinations of clinical variables. Both VAE dimension 9 and PCA dimension 16, which associate with QMDiab diabetes groups, strongly associate with HbA1c (%). VAE dimension 9 HbA1c association p = 5.6x10^{-56}, PCA dimension 16 HbA1c association p = 1.1x10^{-30}.

Type 2 diabetes. VAE latent dimension 9 showed the highest association with type 2 diabetes, with a substantially stronger signal than the highest correlating PCA dimension 16 (p=1.7x10^{-32} vs. p=2.1x10^{-20}, respectively; Figure 5.7d). The top ranking metabolite in dimension 9 was glucose, which is directly affected by the disease and thus serves as a positive control. The top sub-pathways were “acyl carnitine fatty acid metabolism”, “glycolysis, gluconeogenesis, and pyruvate metabolism”, “vitamin B6 metabolism”, and “histidine metabolism”. Other high-ranking metabolites were pyridoxate, histidine, and medium chain acyl-carnitines (Figure 5.7g). Vitamin B6 metabolism, which includes pyridoxate, has been shown to associate with type 2 diabetes and associate with the predisposition of diabetic patients to other diseases [100, 101]. Additionally, circulating medium chain acyl-carnitines have been shown to be associated with early stages of type 2 diabetes [102, 103]. We furthermore correlated dimension 9 with clinical lab measurements from the QMDiab study, and found a strong association between this dimension and HbA1c ($p=5.6x10^{-56}$ compared to PCA $p=1.1x10^{-30}$, Figure 5.8), a widely used diabetes biomarker [104, 105]. This finding demonstrates how a quantitative disease biomarker can carry more information than a crude disease yes/no classification, and further highlights the higher information content in the VAE latent layer compared to PCA.

Schizophrenia. VAE dimension 11 had a stronger association with schizophrenia than PCA dimension 15 ($p=2.0x10^{-8}$ vs. $p=6.6x10^{-6}$, respectively; Figure 5.7e). The top scoring metabolites for this dimension (Figure 5.7h) were mainly acyl-carnitines, such as 4-decanoylcarninite, octanoylcarnitine, and hexanoylcarnitine, and a series of lysolipids. Acyl-carnitines, which are involved in energy metabolism and reflect an individual's mitochondrial beta-oxidation capacity, have been shown to associate with schizophrenia previously [106, 107]. Vitamin B6 metabolism, through pyridoxate, is also one of the highest ranking pathways for this dimension. Previous studies have demonstrated that low levels of vitamin B6 associates with a subgroup of schizophrenic patients [108, 109].

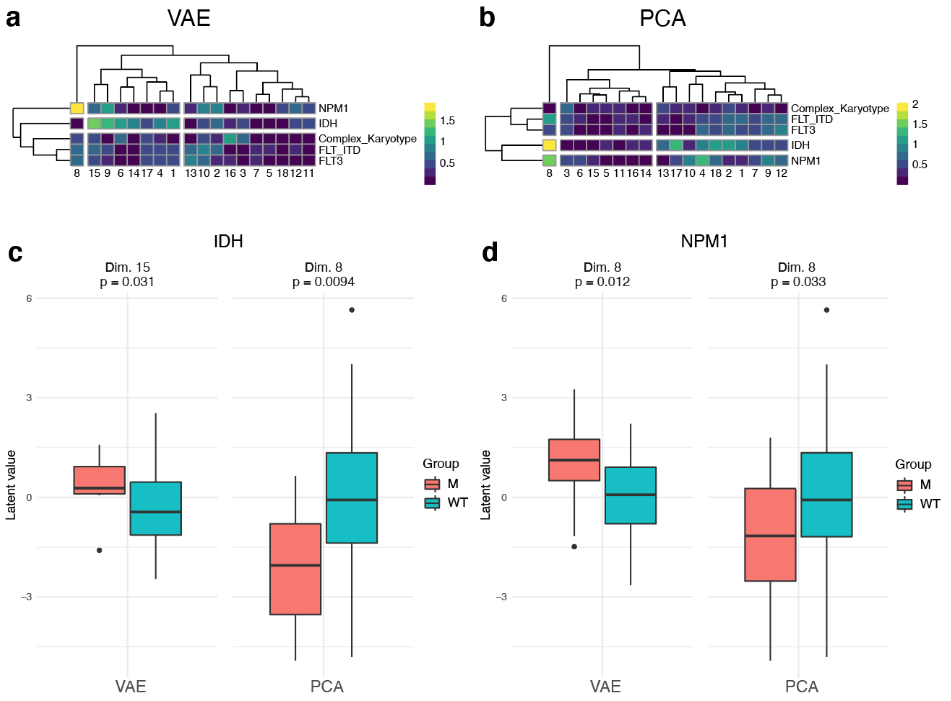


Figure 5.9. AML mutation profile and latent dimension associations. (**a**) VAE latent dimensions association heatmap. (**b**) PCA latent dimension association heatmap. IDH and NPM1 show the strongest associations to the latent dimensions. **c**, boxplot of VAE and PCA latent values for IDH. PCA dimension 8 associates stronger with IDH. **d**, boxplot of VAE and PCA latent values for NPM1. VAE dimension 8 associates more with NPM1. Color bars for **a** and **b** are -log$_{10}$(p-value). For **c** and **d** M = mutant, WT = wildtype.

Acute myeloid leukemia. AML response groups associated an order of magnitude stronger with VAE dimension 15 than PCA dimension 10 (p=0.018 vs. p=0.16, respectively; Figure 5.7f). Note that the p-value would not withstand multiple testing correction; the detected signal is thus merely suggestive and requires replication in future studies. Phosphate, which regulates the oxidative phosphorylation pathway and is involved in energy metabolism, is the most important metabolite for dimension 15. It has been previously demonstrated that oxidative phosphorylation plays a paramount role in AML survival and drug resistance [110–112] and could be an effective target for combination therapy in chemoresistant AML [111, 113]. Additionally, dimension 15 is driven by various metabolites from the n3 and n6 polyunsaturated fatty acid (PUFA) sub-pathway, such as docosahexaenoate (DHA) and eicosapentaenoate (EPA) (Figure 5.7i). It has been shown that treatment of AML cell lines with DHA and EPA has a deleterious effects on their mitochondrial metabolism which leads to cell death [114–117]. We furthermore investigated correlations of the latent dimensions with 17 major AML-related mutations; the analysis revealed no noteworthy results (Figure 5.9).

Taken together, these results suggest that our VAE has learned representations of metabolic processes that are essential for unseen clinical outcomes.

5.3 Discussion & Conclusion

In this chapter, we trained a VAE on metabolomics data from the TwinsUK population cohort and applied the learned latent representations on unseen data. VAE outperformed PCA in terms of correlation matrix reconstruction. Interpretation of VAE latent dimensions at the metabolite, sub-pathway, and super-pathway level revealed that these dimensions represent functionally related and distinct cellular processes. Moreover, VAE latent dimensions show stronger disease associations than PCA in unseen data, as shown in Type 2 Diabetes, schizophrenia and AML datasets. This implies that the VAE learned a latent representation of metabolomics data that is biologically informative and transferable across different cohorts.

Calculating the mean squared error (MSE) [45] between true and model-reconstructed samples is commonly used for the training of models on normally distributed data. However, this metric does not always correspond to model performance on biological data. For instance, we observed that PCA had a lower MSE compared to VAE (Figure 5.3) despite VAE significantly outperforming PCA in associating with disease groups. A similar discrepancy was demonstrated in another study [48, 118]. Interestingly, when we calculated the MSE between the original and reconstructed metabolite correlation matrices, we found that VAE outperformed PCA in this metric. Given our observation, we postulate that improved model performance arises by capturing intrinsic correlation structures in data, rather than better sample reconstruction.

The generalizability of the VAE across different datasets is especially remarkable given the vastly different underlying populations of the datasets we analyzed. The VAE was trained on the TwinsUK population cohort, a European-ancestery population cohort consisting predominantly of British women (~92%), while the validation datasets are mixed-gender and multi-ethnic cohorts from the US and Qatar. Despite the existence of these variations in our datasets, our VAE learned a generalized representation of metabolomics data which was able to identify disease-related differences.

To the best of our knowledge, this is the first time a universal latent representation of metabolomics data is constructed using VAEs. Our results show that VAEs are well-suited for metabolomics data analysis and can potentially replace dimensionality reduction approaches, such as PCA, in creating universal, systems-level understanding of metabolism.

Chapter 6: Conclusion & Outlook

Incorporating molecular interactions into the analysis of high-throughput data is essential for our understanding of biological mechanisms at the systems level. However, current commonly-used pathway and network-based approaches and linear dimensionality-reduction techniques have two main limitations: in analyzing high-throughput data, (1) the above methods do not include information contained at the level of single molecular interactions and (2) nonlinearities present in datasets are omitted.

In this thesis, we contributed in the development and implementation of novel network-based methods that extensively utilize molecular interactions in the analysis of high-throughput data, specifically metabolomics data. Furthermore, we showed that our methods significantly outperform state-of-the-art techniques that are widely used in the field.

6.1 Scientific accomplishments

The following is a list of new and significant scientific contributions and insights that my work has provided to the field of high-throughput biological data analysis and interpretation.

- Current network-based methods do not adequately address misleading shortcuts, such as cofactors and hub genes, that exist in their interaction networks, consequently limiting their use of molecular interactions. This includes state-of-the-art methods such as Ingenuity Pathway Analysis (IPA) and ConsensusPathDB. We constructed a multi-omics network, comprising metabolomics, PPI, signaling and transcription factor regulation. We did this by developing an atom-tracing based algorithm to construct our metabolic network and applied stringent filtering criteria for our gene interaction network to exclude misleading shortcuts (Chapter 2). This multi-omics network was the basis for our piTracer tool.

- A new R Shiny-based application, piTracer, that enables the rapid and automatic reconstruction of biological cascades, was presented (Chapter 3). to recover molecular traces from our multi-omics network (Chapter 2), we implemented Yen's k-shortest path algorithm. To group biologically similar paths, we additionally implemented a path clustering algorithm. We then demonstrated that piTracer, unlike existing tools, accurately constructed well-known gene-metabolite, metabolite-metabolite, and gene-gene cascades.

- With the ability of piTracer to reconstruct biological pathways, we demonstrated its utility in finding compensatory survival metabolic pathways in a breast cancer cell line (Chapter 4). Additionally, based on these reconstructions, piTracer enabled us to predict and prioritize the correct gene targets out of thousands of potential druggable genes. Remarkably, we were able to experimentally validate the deleterious effects of drugging any of our selected genes on the breast cancer cell line.

- VAEs are well-suited in modeling nonlinearities in data by learning feature interactions. To find a low-dimensional representation of metabolomics data using the nonlinearities that arise from metabolic interactions, we trained a VAE on TwinsUK, a large-scale metabolomics population cohort of human blood samples. We then interpreted learned VAE latent representations by calculating a global feature importance score, i.e. SAGE scores, which showed that representations signify specific cellular mechanisms.

- Furthermore, we demonstrated that VAE latent representations significantly correlate with unseen and very different patient groups, implying that, leveraging nonlinearities in metabolomics data, these VAE representations capture disease-associated biological mechanisms.

6.2 Extensions and future directions

From a methodological perspective, the results presented in this thesis could be extended in several directions, which are discussed below.

piTracer app improvements

(1) In the future, SNP to gene interactions could be integrated in our multi-omics network. This would allow users to construct biological paths based on genetic influences. Currently, SNPs have to be manually mapped to genes first, and these genes are used as a proxy in piTracer. Having SNPs directly linked to genes in the future will help bypass this step.

(2) Metabolite-to-gene interactions are not integrated in our multi-omics network. These interactions include end-product inhibition to regulate pathway output, nuclear receptor mechanisms, and epigenetic processes that depend on metabolite pools. These must be included due to the increasing body of evidence demonstrating the central role of the metabolome in biological processes such as signal transduction, proteostasis, and regulation of gene expression on a systemic level [119]. In future work, these interactions could be parsed from published studies that demonstrate regulatory effects of metabolites on other omes and included in our multi-omics network, since there are currently no metabolite-to-gene databases.

(3) A further extension could be the curation of our gene interaction network. That is, by performing network analysis to find high degree gene-nodes, i.e. genes connected to many other genes, and verifying whether they represent true hub genes or network database artifacts. Despite selecting specific interactions from STRING, OmniPath, and GTEx, there are still some shortcut interactions that persist in the network. Future efforts to delete these shortcut nodes in the gene interaction region of our network could further improve the gene-gene paths generated by our app.

(4) Currently, in our algorithm for predicting druggable genes from metabolomics data, some simplifications were made. For example, the gene scores were calculated primarily from binary metabolic scores, i.e. whether a metabolite was statistically significant or not between two experimental conditions. An extension of the prediction method could be done by incorporating metabolite fold changes or gene expression data to generate more quantitative predictions. If successful, this could allow for the prediction of a specific quantitative metabolic output in an experiment, e.g. how much the metabolome of a biological system would be impacted by perturbing a certain enzyme.

VAE model improvements

(5) The main limitation of our metabolomics-based VAE model is the size of the dataset it was trained on, namely the TwinsUK training dataset with n=4,644. This is a general issue with human subject metabolomics studies, where even the largest cohorts reach only about n=15,000 [120]. Deep learning models are currently more popular in larger datasets of n=60,000 samples and more, e.g. from e.g. single cell transcriptomics [43, 121–123], images [124–126], and text sources [127]. Learning the variation in such large datasets allows these models to significantly outperform any linear and other nonlinear. Interestingly, large metabolomics datasets, such as from the UK BioBank with a sample size of up to n=500,000 [128], will be available in the near future, and will enable the creation of more expressive and deeper VAE models.

(6) Another limitation in the construction of our VAE model was that it was solely trained on common metabolites between all of our datasets, despite the intersection of each dataset pair having more metabolites. These discarded metabolites could contain important nonlinearities that would increase the expressivity of the VAE model. Other studies have demonstrated that it is possible to train VAEs that are capable of handling missing data [129, 130]. Training such a model on TwinsUK data could (1) potentially increase the magnitude of statistical association of latent dimensions with disease, (2) allow for the discovery of new disease mechanisms, and (3) enable better and faster data imputation, as current imputation methods are computationally expensive, especially for large datasets.

6.3 Conclusion

In this thesis, I presented novel methods in the interpretation of high-throughput biological data, specifically untargeted metabolomics data. Compared to state-of-the-art techniques, I demonstrated that all of my methods had significantly superior performance in automatically and rapidly reconstructing true biological cascades, finding correct disease-essential and druggable genes, and constructing a universal low-dimensional representation of metabolics data. By correctly constructing molecular mechanisms and enabling disease-relevant gene prioritization, piTracer enables the interpretation of complex statistical results and proves invaluable in the streamlining of drug screening and repurposing efforts. Moreover, VAEs could replace common dimensionality reduction approaches, such as PCA, and enable a better systems-level comprehension of metabolism. Taken together, these network-based methods enable the efficient utilization and lead to significant advancements in our understanding of high-throughput multi-omics data.

Appendix: Other work

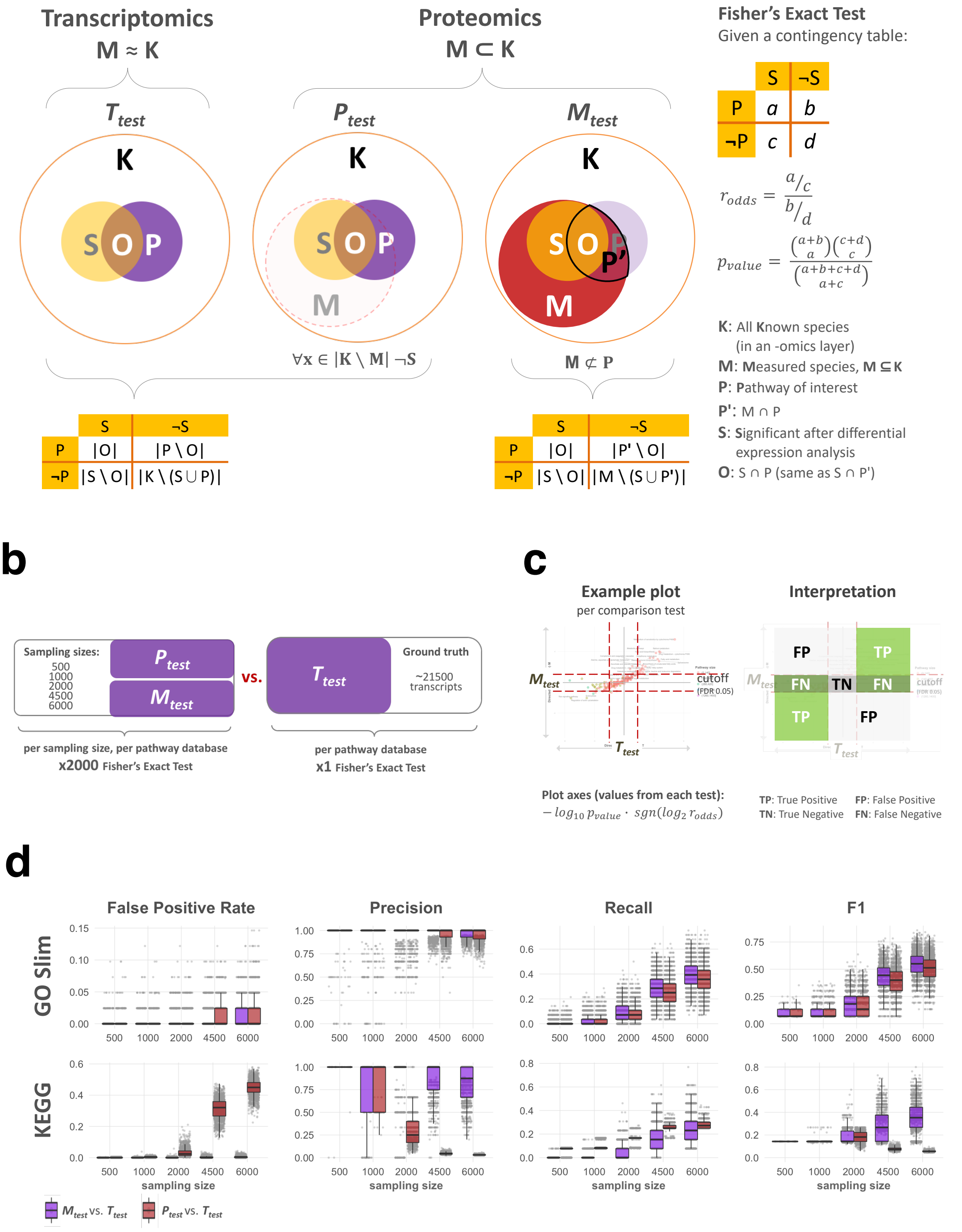
During the course of my PhD, I undertook several projects which are not discussed in this thesis.

The Effects of Background Set on Pathway Enrichment Analysis

To reveal important biological processes from a high-throughput experiment, pathway enrichment (PE) analysis identifies pathways that contain a higher number of regulated proteins than expected by chance. A well-known PE technique is to compare the differentially expressed proteins with a chosen background, which usually comprises the proteome complement set. Enrichment can then be quantified by statistical methods such as Fisher's exact test. However, the choice of background can drastically affect PE results. For instance, when the experiment only captures a small subset of the proteome, the complement set will contain many unmeasured proteins that are automatically considered differentially unaffected. This assumption then inflates the number of unperturbed proteins used in the calculation of the PE significance, leading to an artificial increase in the number of enriched pathways (Appendix Figure 1a).

To quantify how the background choice affects PE analyses, we carried out a simulation study (Appendix Figure 1b) using whole genome expression data from diabetic and healthy mice. Subsets of the transcript data were taken and enrichment analysis on Gene Ontology (GO) Slim and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways using Fisher's test were done. The results were subsequently compared to the results of the same analysis using the entirety of the genome expression data (ground truth) (Appendix Figure 1c).

We found that when the proteome background is chosen, the false positive rates for the enriched pathways are higher (Appendix Figure 1d). Moreover, precision tends to be lower, in contrast to simulations using the measured list background. Therefore, when performing PE analysis, it is best to select a background list consisting exclusively of the measured proteins.



Appendix Figure 1. (**a**) Comparison between common pathway enrichment assumptions and our proposed assumption. (**b**) The parameters in our simulation study using whole genome expression data from diabetic and healthy mice. (**c**) Using the entirety of the genome expression data as our ground truth, we scored pathway enrichment methods with different background assumptions. (**d**) The results of our simulation study, where we found that our proposed method performs better than methods where it is assumed that what is not measured cannot be differentially regulated.

Cross-omics imputation using Variational Autoencoders

Multi-omics data is measured from the same underlying biological system. Consequently, if it were possible to construct an accurate model of such a biological system, it would be possible to generate all possible omics of the system. This is highly desirable, since having all possible omics measured would allow for a holistic systems-level analysis and interpretation of data. Such a model would take one omics layer as an input and would predict/infer other types of omics data. Among its many applications, this would lead to a reduction of costs and time associated with measuring multi-omics data. Naturally, such a model will increase our understanding of which omics layer is the most informative in predicting other omics types or whether there are specific combinations of subsets of omics that can predict all other omics. One way such a model could be created is by training VAEs on multi-omics data, and aligning the learned latent spaces of all omics types.

To this end, we used The Cancer Genome Atlas (TCGA) data, consisting of transcripts, proteins, and single- nucleotide variants and simultaneously trained a VAE for each data type. While training, we introduced a loss function that aligned the latent spaces of the VAEs. We generated preliminary results that showed better cancer type separation in the VAE latent space than our baseline PCA. Furthermore, compared to PCA, the VAE had a marginally better performance at predicting one omics from another.

We also trained VAEs on TwinsUK transcriptomics and metabolomics data. However, we could not show any improvements compared to PCA in the omics prediction task, due the low sample overlap (384 samples) and the large number of variables for each dataset (~15,000 genes). We also attempted to first train separate VAEs for each omics layer, since each data type had more samples than their overlap, and then align the latent spaces of the VAEs. However, this resulted in a similar performance as the PCA baseline, which is why we did not follow up on this project any further.

Currently, given existing datasets, it is difficult to assess whether our approach of training VAEs on multi-omics data allows for the modeling of a biological system capable of generating different omics types. This is due to the absence of large (sample sizes >20,000) and matched multi-omics data. Another study [131] attempted to show the feasibility of VAEs in a similar omics prediction task, but were limited to data of the same omics type (single-cell transcriptomics) measured on different platforms. Despite the absence of large and matched multi-omics data our preliminary TCGA results are encouraging. Therefore, this project should be pursued further in the future.

Bibliography

[1] Y. Hasin, M. Seldin, and A. Lusis, “Multi-omics approaches to disease,” *Genome Biol*, vol. 18, no. 1, p. 83, May 2017, doi: 10.1186/s13059-017-1215-1.

[2] M. Civelek and A. J. Lusis, “Systems genetics approaches to understand complex traits,” *Nat Rev Genet*, vol. 15, no. 1, pp. 34–48, Jan. 2014, doi: 10.1038/nrg3575.

[3] K. Suhre *et al.*, “Human metabolic individuality in biomedical and pharmaceutical research,” *Nature*, vol. 477, no. 7362, pp. 54–60, Aug. 2011, doi: 10.1038/nature10354.

[4] R. Apweiler *et al.*, “Whither systems medicine?,” *Exp Mol Med*, vol. 50, no. 3, p. e453, Mar. 2018, doi: 10.1038/emm.2017.290.

[5] K. A. Marijt *et al.*, “Metabolic stress in cancer cells induces immune escape through a PI3K-dependent blockade of IFNγ receptor signaling,” *J Immunother Cancer*, vol. 7, no. 1, p. 152, Jun. 2019, doi: 10.1186/s40425-019-0627-8.

[6] J. A. Hartiala *et al.*, “Genome-wide association study and targeted metabolomics identifies sex-specific association of CPS1 with coronary artery disease,” *Nat Commun*, vol. 7, p. 10558, Jan. 2016, doi: 10.1038/ncomms10558.

[7] E. Bobrovnikova-Marjon and J. B. Hurov, “Targeting metabolic changes in cancer: novel therapeutic approaches,” *Annu Rev Med*, vol. 65, pp. 157–170, 2014, doi: 10.1146/annurev-med-092012-112344.

[8] K. Inoue *et al.*, “Metabolic profiling of Alzheimer’s disease brains,” *Sci Rep*, vol. 3, p. 2364, 2013, doi: 10.1038/srep02364.

[9] V. De Preter *et al.*, “Metabolic profiling of the impact of oligofructose-enriched inulin in Crohn’s disease patients: a double-blinded randomized controlled trial,” *Clin Transl Gastroenterol*, vol. 4, p. e30, Jan. 2013, doi: 10.1038/ctg.2012.24.

[10] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, “Data, information, knowledge and principle: back to metabolism in KEGG,” *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D199-205, Jan. 2014, doi: 10.1093/nar/gkt1076.

[11] Gene Ontology Consortium, “Gene Ontology Consortium: going forward,” *Nucleic Acids Res*, vol. 43, no. Database issue, pp. D1049-1056, Jan. 2015, doi: 10.1093/nar/gku1179.

[12] T. Jewison *et al.*, “SMPDB 2.0: big improvements to the Small Molecule Pathway Database,” *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D478-484, Jan. 2014, doi: 10.1093/nar/gkt1067.

[13] E. Brunk *et al.*, “Recon3D enables a three-dimensional view of gene variation in human metabolism,” *Nat Biotechnol*, vol. 36, no. 3, pp. 272–281, Mar. 2018, doi: 10.1038/nbt.4072.

[14] D. Szklarczyk *et al.*, “STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D607–D613, Jan. 2019, doi: 10.1093/nar/gky1131.

[15] D. Türei, T. Korcsmáros, and J. Saez-Rodriguez, “OmniPath: guidelines and gateway for literature-curated signaling pathway resources,” *Nat Methods*, vol. 13, no. 12, pp. 966–967, Nov. 2016, doi: 10.1038/nmeth.4077.

[16] A. Fabregat *et al.*, “Reactome graph database: Efficient access to complex pathway data,” *PLoS Comput Biol*, vol. 14, no. 1, p. e1005968, Jan. 2018, doi: 10.1371/journal.pcbi.1005968.

[17] J. Reimand *et al.*, “Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap,” *Nat Protoc*, vol. 14, no. 2, pp. 482–517, Feb. 2019, doi: 10.1038/s41596-018-0103-9.

[18] E. Y. Chen *et al.*, “Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool,” *BMC Bioinformatics*, vol. 14, p. 128, Apr. 2013, doi: 10.1186/1471-2105-14-128.

[19] W. T. Barry, A. B. Nobel, and F. A. Wright, “Significance analysis of functional categories in gene expression studies: a structured permutation approach,” *Bioinformatics*, vol. 21, no. 9, pp. 1943–1949, May 2005, doi: 10.1093/bioinformatics/bti260.

[20] W. Luo, M. S. Friedman, K. Shedden, K. D. Hankenson, and P. J. Woolf, “GAGE: generally applicable gene set enrichment for pathway analysis,” *BMC Bioinformatics*, vol. 10, p. 161, May 2009, doi: 10.1186/1471-2105-10-161.

[21] A. Subramanian *et al.*, “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles,” *Proc Natl Acad Sci U S A*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005, doi: 10.1073/pnas.0506580102.

[22] D. S. Himmelstein *et al.*, “Systematic integration of biomedical knowledge prioritizes drugs for repurposing,” *Elife*, vol. 6, Sep. 2017, doi: 10.7554/eLife.26726.

[23] M. A. Reyna, M. D. M. Leiserson, and B. J. Raphael, “Hierarchical HotNet: identifying hierarchies of altered subnetworks,” *Bioinformatics*, vol. 34, no. 17, pp. i972–i980, Sep. 2018, doi: 10.1093/bioinformatics/bty613.

[24] J. D. Orth, I. Thiele, and B. Ø. Palsson, “What is flux balance analysis?,” *Nat Biotechnol*, vol. 28, no. 3, pp. 245–248, Mar. 2010, doi: 10.1038/nbt.1614.

[25] Z. Abd Algfoor, M. Shahrizal Sunar, A. Abdullah, and H. Kolivand, “Identification of metabolic pathways using pathfinding approaches: a

systematic review," *Brief Funct Genomics*, vol. 16, no. 2, pp. 87–98, Mar. 2017, doi: 10.1093/bfgp/elw002.

[26] D. Croes, F. Couche, S. J. Wodak, and J. van Helden, "Inferring meaningful pathways in weighted metabolic networks," *J Mol Biol*, vol. 356, no. 1, pp. 222–236, Feb. 2006, doi: 10.1016/j.jmb.2005.09.079.

[27] I. T. Jolliffe, *Principal component analysis*, 2nd ed. New York: Springer, 2002.

[28] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1–10, Jul. 1991, doi: 10.1016/0165-1684(91)90079-X.

[29] J. Krumsiek, K. Suhre, T. Illig, J. Adamski, and F. J. Theis, "Bayesian Independent Component Analysis Recovers Pathway Signatures from Blood Metabolomics Data," *J. Proteome Res.*, vol. 11, no. 8, pp. 4120–4131, Aug. 2012, doi: 10.1021/pr300231n.

[30] M. C. Walsh, L. Brennan, J. P. G. Malthouse, H. M. Roche, and M. J. Gibney, "Effect of acute dietary standardization on the urinary, plasma, and salivary metabolomic profiles of healthy humans," *Am J Clin Nutr*, vol. 84, no. 3, pp. 531–539, Sep. 2006, doi: 10.1093/ajcn/84.3.531.

[31] G. Nyamundanda, L. Brennan, and I. C. Gormley, "Probabilistic principal component analysis for metabolomic data," *BMC Bioinformatics*, vol. 11, p. 571, Nov. 2010, doi: 10.1186/1471-2105-11-571.

[32] H. Yamamoto *et al.*, "Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables," *Chemometrics and Intelligent Laboratory Systems*, vol. 98, no. 2, pp. 136–142, Oct. 2009, doi: 10.1016/j.chemolab.2009.05.006.

[33] Y. Liu, K. Smirnov, M. Lucio, R. D. Gougeon, H. Alexandre, and P. Schmitt-Kopplin, "MetICA: independent component analysis for high-

resolution mass-spectrometry based non-targeted metabolomics," *BMC Bioinformatics*, vol. 17, p. 114, Mar. 2016, doi: 10.1186/s12859-016-0970-4.

[34] X. W. Zhang, Y. L. Yap, D. Wei, F. Chen, and A. Danchin, "Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis," *Eur J Hum Genet*, vol. 13, no. 12, pp. 1303–1311, Dec. 2005, doi: 10.1038/sj.ejhg.5201495.

[35] A. E. Teschendorff, M. Journée, P. A. Absil, R. Sepulchre, and C. Caldas, "Elucidating the altered transcriptional programs in breast cancer using independent component analysis," *PLoS Comput Biol*, vol. 3, no. 8, p. e161, Aug. 2007, doi: 10.1371/journal.pcbi.0030161.

[36] J. M. Engreitz, B. J. Daigle, J. J. Marshall, and R. B. Altman, "Independent component analysis: mining microarray data for fundamental human gene expression modules," *J Biomed Inform*, vol. 43, no. 6, pp. 932–944, Dec. 2010, doi: 10.1016/j.jbi.2010.07.001.

[37] R. Gaujoux and C. Seoighe, "CellMix: a comprehensive toolbox for gene expression deconvolution," *Bioinformatics*, vol. 29, no. 17, pp. 2211–2212, Sep. 2013, doi: 10.1093/bioinformatics/btt351.

[38] K. Schwahn, R. Beleggia, N. Omranian, and Z. Nikoloski, "Stoichiometric Correlation Analysis: Principles of Metabolic Functionality from Metabolomics Data," *Front Plant Sci*, vol. 8, p. 2152, 2017, doi: 10.3389/fpls.2017.02152.

[39] H.-S. Song, F. DeVilbiss, and D. Ramkrishna, "Modeling metabolic systems: the need for dynamics," *Current Opinion in Chemical Engineering*, vol. 2, no. 4, pp. 373–382, Nov. 2013, doi: 10.1016/j.coche.2013.08.004.

[40] T. Illig *et al.*, "A genome-wide perspective of genetic variation in human metabolism," *Nat Genet*, vol. 42, no. 2, pp. 137–141, Feb. 2010, doi: 10.1038/ng.507.

[41] A.-K. Petersen *et al.*, "On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies," *BMC Bioinformatics*, vol. 13, p. 120, Jun. 2012, doi: 10.1186/1471-2105-13-120.

[42] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE J.*, vol. 37, no. 2, pp. 233–243, Feb. 1991, doi: 10.1002/aic.690370209.

[43] G. P. Way and C. S. Greene, "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders," *Pac Symp Biocomput*, vol. 23, pp. 80–91, 2018.

[44] M. Chen *et al.*, "Multifaceted protein-protein interaction prediction based on Siamese residual RCNN," *Bioinformatics*, vol. 35, no. 14, pp. i305–i314, Jul. 2019, doi: 10.1093/bioinformatics/btz328.

[45] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv:1312.6114 [cs, stat]*, May 2014, Accessed: Jan. 15, 2021. [Online]. Available: http://arxiv.org/abs/1312.6114.

[46] A. B. Dincer, S. Celik, N. Hiranuma, and S.-I. Lee, "DeepProfile: Deep learning of cancer molecular profiles for precision medicine," Bioinformatics, preprint, Mar. 2018. doi: 10.1101/278739.

[47] M. Lotfollahi, F. A. Wolf, and F. J. Theis, "scGen predicts single-cell perturbation responses," *Nat Methods*, vol. 16, no. 8, pp. 715–721, Aug. 2019, doi: 10.1038/s41592-019-0494-8.

[48] L. Rampasek, D. Hidru, P. Smirnov, B. Haibe-Kains, and A. Goldenberg, "Dr.VAE: Drug Response Variational Autoencoder," *arXiv:1706.08203 [stat]*, Jul. 2017, Accessed: Jan. 15, 2021. [Online]. Available: http://arxiv.org/abs/1706.08203.

[49] Y. Pomyen, K. Wanichthanarak, P. Poungsombat, J. Fahrmann, D. Grapov, and S. Khoomrung, "Deep metabolome: Applications of deep learning in

metabolomics," *Comput Struct Biotechnol J*, vol. 18, pp. 2818–2825, 2020, doi: 10.1016/j.csbj.2020.09.033.

[50] A. Noronha *et al.*, "The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease," *Nucleic Acids Res*, vol. 47, no. D1, pp. D614–D624, Jan. 2019, doi: 10.1093/nar/gky992.

[51] A. Dalby *et al.*, "Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited," *J. Chem. Inf. Model.*, vol. 32, no. 3, pp. 244–255, May 1992, doi: 10.1021/ci00007a012.

[52] S. A. Rahman *et al.*, "Reaction Decoder Tool (RDT): extracting features from chemical reactions," *Bioinformatics*, vol. 32, no. 13, pp. 2065–2066, Jul. 2016, doi: 10.1093/bioinformatics/btw096.

[53] A. R. Sonawane *et al.*, "Understanding Tissue-Specific Gene Regulation," *Cell Rep*, vol. 21, no. 4, pp. 1077–1088, Oct. 2017, doi: 10.1016/j.celrep.2017.10.001.

[54] K. Glass, C. Huttenhower, J. Quackenbush, and G.-C. Yuan, "Passing messages between biological networks to refine predicted interactions," *PLoS One*, vol. 8, no. 5, p. e64832, 2013, doi: 10.1371/journal.pone.0064832.

[55] GTEx Consortium, "Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans," *Science*, vol. 348, no. 6235, pp. 648–660, May 2015, doi: 10.1126/science.1262110.

[56] D. Szklarczyk *et al.*, "STRING v10: protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Res*, vol. 43, no. Database issue, pp. D447-452, Jan. 2015, doi: 10.1093/nar/gku1003.

[57] R. Herwig, C. Hardt, M. Lienhard, and A. Kamburov, "Analyzing and interpreting genome data at the network level with ConsensusPathDB," *Nat*

Protoc, vol. 11, no. 10, pp. 1889–1907, Oct. 2016, doi: 10.1038/nprot.2016.117.

[58] A. Krämer, J. Green, J. Pollard, and S. Tugendreich, “Causal analysis approaches in Ingenuity Pathway Analysis,” *Bioinformatics*, vol. 30, no. 4, pp. 523–530, Feb. 2014, doi: 10.1093/bioinformatics/btt703.

[59] J. Y. Yen, “Finding the *K* Shortest Loopless Paths in a Network,” *Management Science*, vol. 17, no. 11, pp. 712–716, Jul. 1971, doi: 10.1287/mnsc.17.11.712.

[60] M. Bockholt and K. A. Zweig, “Clustering of Paths in Complex Networks,” in *Complex Networks & Their Applications V*, Cham, 2017, pp. 183–195, doi: 10.1007/978-3-319-50901-3_15.

[61] R Core Team (2020), *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* .

[62] P. Langfelder, B. Zhang, and S. Horvath, “Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R,” *Bioinformatics*, vol. 24, no. 5, pp. 719–720, Mar. 2008, doi: 10.1093/bioinformatics/btm563.

[63] W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson, *shiny: Web Application Framework for R. R package version 1.5.0.* 2020.

[64] S.-Y. Shin *et al.*, “An atlas of genetic influences on human blood metabolites,” *Nat Genet*, vol. 46, no. 6, pp. 543–550, Jun. 2014, doi: 10.1038/ng.2982.

[65] M. M. Heemskerk, V. J. A. van Harmelen, K. W. van Dijk, and J. B. van Klinken, “Reanalysis of mGWAS results and in vitro validation show that lactate dehydrogenase interacts with branched-chain amino acid metabolism,” *Eur J Hum Genet*, vol. 24, no. 1, pp. 142–145, Jan. 2016, doi: 10.1038/ejhg.2015.106.

[66] K. Matejka *et al.*, “Dynamic modelling of an ACADS genotype in fatty acid oxidation - Application of cellular models for the analysis of common genetic variants,” *PLoS One*, vol. 14, no. 5, p. e0216110, 2019, doi: 10.1371/journal.pone.0216110.

[67] A. El-Gharbawy and J. Vockley, “Inborn Errors of Metabolism with Myopathy: Defects of Fatty Acid Oxidation and the Carnitine Shuttle System,” *Pediatr Clin North Am*, vol. 65, no. 2, pp. 317–335, Apr. 2018, doi: 10.1016/j.pcl.2017.11.006.

[68] A. Bevilacqua and M. Bizzarri, “Inositols in Insulin Signaling and Glucose Metabolism,” *International Journal of Endocrinology*, vol. 2018, pp. 1–8, Nov. 2018, doi: 10.1155/2018/1968450.

[69] J. M. Berg, J. L. Tzmoczko, and L. Stryer, “Section 26.1, Phosphatidate Is a Common Intermediate in the Synthesis of Phospholipids and Triacylglycerols,” in *Biochemistry*, 5th ed., New York: W.H. Freeman, 2002.

[70] L. Sun, M. Bartlam, Y. Liu, H. Pang, and Z. Rao, “Crystal structure of the pyridoxal-5’-phosphate-dependent serine dehydratase from human liver,” *Protein Sci*, vol. 14, no. 3, pp. 791–798, Mar. 2005, doi: 10.1110/ps.041179105.

[71] J.-H. Park *et al.*, “RICK/RIP2 mediates innate immune responses induced through Nod1 and Nod2 but not TLRs,” *J Immunol*, vol. 178, no. 4, pp. 2380–2386, Feb. 2007, doi: 10.4049/jimmunol.178.4.2380.

[72] M. Hasegawa *et al.*, “A critical role of RICK/RIP2 polyubiquitination in Nod-induced NF-kappaB activation,” *EMBO J*, vol. 27, no. 2, pp. 373–383, Jan. 2008, doi: 10.1038/sj.emboj.7601962.

[73] M. M. Shen, “Nodal signaling: developmental roles and regulation,” *Development*, vol. 134, no. 6, pp. 1023–1034, Mar. 2007, doi: 10.1242/dev.000166.

[74] J.-S. Mo, F.-X. Yu, R. Gong, J. H. Brown, and K.-L. Guan, “Regulation of the Hippo-YAP pathway by protease-activated receptors (PARs),” *Genes Dev*, vol. 26, no. 19, pp. 2138–2143, Oct. 2012, doi: 10.1101/gad.197582.112.

[75] D. Ready *et al.*, “Mapping the STK4/Hippo signaling network in prostate cancer cell,” *PLoS One*, vol. 12, no. 9, p. e0184590, 2017, doi: 10.1371/journal.pone.0184590.

[76] P. C. Calses, J. J. Crawford, J. R. Lill, and A. Dey, “Hippo Pathway in Cancer: Aberrant Regulation and Therapeutic Opportunities,” *Trends Cancer*, vol. 5, no. 5, pp. 297–307, May 2019, doi: 10.1016/j.trecan.2019.04.001.

[77] C. Holohan, S. Van Schaeybroeck, D. B. Longley, and P. G. Johnston, “Cancer drug resistance: an evolving paradigm,” *Nat Rev Cancer*, vol. 13, no. 10, pp. 714–726, Oct. 2013, doi: 10.1038/nrc3599.

[78] X. Wang, H. Zhang, and X. Chen, “Drug resistance and combating drug resistance in cancer,” *CDR*, 2019, doi: 10.20517/cdr.2019.10.

[79] D. Longley and P. Johnston, “Molecular mechanisms of drug resistance,” *J. Pathol.*, vol. 205, no. 2, pp. 275–292, Jan. 2005, doi: 10.1002/path.1706.

[80] D. S. Wishart, “Emerging applications of metabolomics in drug discovery and precision medicine,” *Nat Rev Drug Discov*, vol. 15, no. 7, pp. 473–484, Jul. 2016, doi: 10.1038/nrd.2016.32.

[81] J. Krumsiek, J. Bartel, and F. J. Theis, “Computational approaches for systems metabolomics,” *Current Opinion in Biotechnology*, vol. 39, pp. 198–206, Jun. 2016, doi: 10.1016/j.copbio.2016.04.009.

[82] G. Housman *et al.*, “Drug Resistance in Cancer: An Overview,” *Cancers*, vol. 6, no. 3, pp. 1769–1792, Sep. 2014, doi: 10.3390/cancers6031769.

[83] A. Halama *et al.*, “Accelerated lipid catabolism and autophagy are cancer survival mechanisms under inhibited glutaminolysis,” *Cancer Lett*, vol. 430, pp. 133–147, Aug. 2018, doi: 10.1016/j.canlet.2018.05.017.

[84] L. M. D. Reis *et al.*, “Dual inhibition of glutaminase and carnitine palmitoyltransferase decreases growth and migration of glutaminase inhibition-resistant triple-negative breast cancer cells,” *J Biol Chem*, vol. 294, no. 24, pp. 9342–9357, Jun. 2019, doi: 10.1074/jbc.RA119.008180.

[85] S. Pushpakom *et al.*, “Drug repurposing: progress, challenges and recommendations,” *Nat Rev Drug Discov*, vol. 18, no. 1, pp. 41–58, Jan. 2019, doi: 10.1038/nrd.2018.168.

[86] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott, “Principles of early drug discovery,” *Br J Pharmacol*, vol. 162, no. 6, pp. 1239–1249, Mar. 2011, doi: 10.1111/j.1476-5381.2010.01127.x.

[87] D. Cook *et al.*, “Lessons learned from the fate of AstraZeneca’s drug pipeline: a five-dimensional framework,” *Nat Rev Drug Discov*, vol. 13, no. 6, pp. 419–431, Jun. 2014, doi: 10.1038/nrd4309.

[88] A. Moayyeri, C. J. Hammond, D. J. Hart, and T. D. Spector, “The UK Adult Twin Registry (TwinsUK Resource),” *Twin Res Hum Genet*, vol. 16, no. 1, pp. 144–149, Feb. 2013, doi: 10.1017/thg.2012.89.

[89] I. Covert, S. Lundberg, and S.-I. Lee, “Understanding Global Feature Contributions With Additive Importance Measures,” *arXiv:2004.00668 [cs, stat]*, Oct. 2020, Accessed: Jan. 15, 2021. [Online]. Available: http://arxiv.org/abs/2004.00668.

[90] D. O. Mook-Kanamori *et al.*, “1,5-Anhydroglucitol in saliva is a noninvasive marker of short-term glycemic control,” *J Clin Endocrinol Metab*, vol. 99, no. 3, pp. E479-483, Mar. 2014, doi: 10.1210/jc.2013-3596.

[91] S. Hammoudeh *et al.*, “The prevalence of metabolic syndrome in patients receiving antipsychotics in Qatar: a cross sectional comparative study,” *BMC Psychiatry*, vol. 18, no. 1, p. 81, Mar. 2018, doi: 10.1186/s12888-018-1662-6.

[92] Eastern Cooperative Oncology Group, “A Phase III Trial in Adult Acute Myeloid Leukemia: Daunorubicin Dose-Intensification Prior to Risk-Allocated Autologous Stem Cell Transplantation,” clinicaltrials.gov, Clinical trial registration NCT00049517, Oct. 2020. Accessed: Jan. 14, 2021. [Online]. Available: https://clinicaltrials.gov/ct2/show/NCT00049517.

[93] A. M. Evans, C. D. DeHaven, T. Barrett, M. Mitchell, and E. Milgram, “Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems,” *Anal Chem*, vol. 81, no. 16, pp. 6656–6667, Aug. 2009, doi: 10.1021/ac901536h.

[94] Z. Yu *et al.*, “Differences between human plasma and serum metabolite profiles,” *PLoS One*, vol. 6, no. 7, p. e21230, 2011, doi: 10.1371/journal.pone.0021230.

[95] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn, “Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics,” *Anal Chem*, vol. 78, no. 13, pp. 4281–4290, Jul. 2006, doi: 10.1021/ac051632c.

[96] C. S. Schmaljohn, S. E. Hasty, L. Rasmussen, and J. M. Dalrymple, “Hantaan virus replication: effects of monensin, tunicamycin and endoglycosidases on the structural glycoproteins,” *J Gen Virol*, vol. 67 (Pt 4), pp. 707–717, Apr. 1986, doi: 10.1099/0022-1317-67-4-707.

[97] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” p. 8.

[98] A. L. Maas, A. Y. Hannun, and A. Y. Ng, “Rectifier Nonlinearities Improve Neural Network Acoustic Models,” p. 6.

[99] *keras-team/keras-tuner*. Keras, 2021.

[100] R. Obeid, J. Geisel, and W. A. Nix, “4-Pyridoxic Acid/Pyridoxine Ratio in Patients with Type 2 Diabetes is Related to Global Cardiovascular Risk Scores,” *Diagnostics (Basel)*, vol. 9, no. 1, Mar. 2019, doi: 10.3390/diagnostics9010028.

[101] W. A. Nix *et al.*, “Vitamin B status in patients with type 2 diabetes mellitus with and without incipient nephropathy,” *Diabetes Res Clin Pract*, vol. 107, no. 1, pp. 157–165, Jan. 2015, doi: 10.1016/j.diabres.2014.09.058.

[102] B. Batchuluun *et al.*, “Elevated Medium-Chain Acylcarnitines Are Associated With Gestational Diabetes Mellitus and Early Progression to Type 2 Diabetes and Induce Pancreatic β-Cell Dysfunction,” *Diabetes*, vol. 67, no. 5, pp. 885–897, May 2018, doi: 10.2337/db17-1150.

[103] J. Bene, K. Hadzsiev, and B. Melegh, “Role of carnitine and its derivatives in the development and management of type 2 diabetes,” *Nutr Diabetes*, vol. 8, no. 1, p. 8, Mar. 2018, doi: 10.1038/s41387-018-0017-1.

[104] International Expert Committee, “International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes,” *Diabetes Care*, vol. 32, no. 7, pp. 1327–1334, Jul. 2009, doi: 10.2337/dc09-9033.

[105] American Diabetes Association, “Diagnosis and classification of diabetes mellitus,” *Diabetes Care*, vol. 33 Suppl 1, pp. S62-69, Jan. 2010, doi: 10.2337/dc10-S062.

[106] B. Cao *et al.*, “Characterizing acyl-carnitine biosignatures for schizophrenia: a longitudinal pre- and post-treatment study,” *Transl Psychiatry*, vol. 9, no. 1, p. 19, Jan. 2019, doi: 10.1038/s41398-018-0353-x.

[107] B. Cao, Y. Chen, R. S. McIntyre, and L. Yan, “Acyl-Carnitine plasma levels and their association with metabolic syndrome in individuals with schizophrenia,” *Psychiatry Res*, vol. 293, p. 113458, Nov. 2020, doi: 10.1016/j.psychres.2020.113458.

[108] M. Miyashita *et al.*, “Clinical features of schizophrenia with enhanced carbonyl stress,” *Schizophr Bull*, vol. 40, no. 5, pp. 1040–1046, Sep. 2014, doi: 10.1093/schbul/sbt129.

[109] M. Arai *et al.*, “Enhanced carbonyl stress in a subpopulation of schizophrenia,” *Arch Gen Psychiatry*, vol. 67, no. 6, pp. 589–597, Jun. 2010, doi: 10.1001/archgenpsychiatry.2010.62.

[110] J. Kreitz *et al.*, “Metabolic Plasticity of Acute Myeloid Leukemia,” *Cells*, vol. 8, no. 8, Jul. 2019, doi: 10.3390/cells8080805.

[111] N. Chapuis, L. Poulain, R. Birsen, J. Tamburini, and D. Bouscary, “Rationale for Targeting Deregulated Metabolic Pathways as a Therapeutic Strategy in Acute Myeloid Leukemia,” *Front Oncol*, vol. 9, p. 405, 2019, doi: 10.3389/fonc.2019.00405.

[112] C. Bosc *et al.*, “Autophagy regulates fatty acid availability for oxidative phosphorylation through mitochondria-endoplasmic reticulum contact sites,” *Nat Commun*, vol. 11, no. 1, p. 4056, Aug. 2020, doi: 10.1038/s41467-020-17882-2.

[113] A. Vitkevičienė *et al.*, “Oxidative phosphorylation inhibition induces anticancerous changes in therapy-resistant-acute myeloid leukemia patient cells,” *Mol Carcinog*, vol. 58, no. 11, pp. 2008–2016, Nov. 2019, doi: 10.1002/mc.23092.

[114] F. Picou *et al.*, “n-3 Polyunsaturated fatty acids induce acute myeloid leukemia cell death associated with mitochondrial glycolytic switch and Nrf2 pathway activation,” *Pharmacol Res*, vol. 136, pp. 45–55, Oct. 2018, doi: 10.1016/j.phrs.2018.08.015.

[115] J. E. Slagsvold, C. H. H. Pettersen, T. Follestad, H. E. Krokan, and S. A. Schønberg, “The antiproliferative effect of EPA in HL60 cells is mediated by

alterations in calcium homeostasis," *Lipids*, vol. 44, no. 2, pp. 103–113, Feb. 2009, doi: 10.1007/s11745-008-3263-5.

[116] T. Yamagami, C. D. Porada, R. S. Pardini, E. D. Zanjani, and G. Almeida-Porada, "Docosahexaenoic acid induces dose dependent cell death in an early undifferentiated subtype of acute myeloid leukemia cell line," *Cancer Biol Ther*, vol. 8, no. 4, pp. 331–337, Feb. 2009, doi: 10.4161/cbt.8.4.7334.

[117] A. Loew, T. Köhnke, E. Rehbeil, A. Pietzner, and K.-H. Weylandt, "A Role for Lipid Mediators in Acute Myeloid Leukemia," *Int J Mol Sci*, vol. 20, no. 10, May 2019, doi: 10.3390/ijms20102425.

[118] L. Rampášek, D. Hidru, P. Smirnov, B. Haibe-Kains, and A. Goldenberg, "Dr.VAE: improving drug response prediction via modeling of drug perturbation effects," *Bioinformatics*, vol. 35, no. 19, pp. 3743–3751, Oct. 2019, doi: 10.1093/bioinformatics/btz158.

[119] M. M. Rinschen, J. Ivanisevic, M. Giera, and G. Siuzdak, "Identification of bioactive metabolites using activity metabolomics," *Nat Rev Mol Cell Biol*, vol. 20, no. 6, pp. 353–367, Jun. 2019, doi: 10.1038/s41580-019-0108-4.

[120] B. Yu *et al.*, "The Consortium of Metabolomics Studies (COMETS): Metabolomics in 47 Prospective Cohort Studies," *Am J Epidemiol*, vol. 188, no. 6, pp. 991–1012, Jun. 2019, doi: 10.1093/aje/kwz028.

[121] H. M. Kang *et al.*, "Multiplexed droplet single-cell RNA-sequencing using natural genetic variation," *Nat Biotechnol*, vol. 36, no. 1, pp. 89–94, Jan. 2018, doi: 10.1038/nbt.4042.

[122] T. Hagai *et al.*, "Gene expression variability across cells and species shapes innate immunity," *Nature*, vol. 563, no. 7730, pp. 197–202, Nov. 2018, doi: 10.1038/s41586-018-0657-2.

[123] G. X. Y. Zheng *et al.*, “Massively parallel digital transcriptional profiling of single cells,” *Nat Commun*, vol. 8, p. 14049, Jan. 2017, doi: 10.1038/ncomms14049.

[124] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[125] A. Krizhevsky, “Learning Multiple Layers of Features from Tiny Images,” p. 60.

[126] T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context,” *arXiv:1405.0312 [cs]*, Feb. 2015, Accessed: Jan. 15, 2021. [Online]. Available: http://arxiv.org/abs/1405.0312.

[127] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, “A large annotated corpus for learning natural language inference,” *arXiv:1508.05326 [cs]*, Aug. 2015, Accessed: Jan. 15, 2021. [Online]. Available: http://arxiv.org/abs/1508.05326.

[128] “Nightingale Health and UK Biobank announces major initiative to analyse half a million blood samples to facilitate global medical research.” https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/nightingale-health-and-uk-biobank-announces-major-initiative-to-analyse-half-a-million-blood-samples-to-facilitate-global-medical-research (accessed Jan. 15, 2021).

[129] J. T. McCoy, S. Kroon, and L. Auret, “Variational Autoencoders for Missing Data Imputation with Application to a Simulated Milling Circuit,” *IFAC-PapersOnLine*, vol. 51, no. 21, pp. 141–146, 2018, doi: 10.1016/j.ifacol.2018.09.406.

[130] L. Persico, “[Proteus infections: a current problem],” *Policlinico Prat*, vol. 72, no. 49, pp. 1665–1682, Dec. 1965.

[131] K. D. Yang and C. Uhler, “Multi-Domain Translation by Learning Uncoupled Autoencoders,” *arXiv:1902.03515 [cs, stat]*, Feb. 2019, Accessed: Jan. 15, 2021. [Online]. Available: http://arxiv.org/abs/1902.03515.