PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# PaIRKAT: A pathway integrated regression-based kernel association test with applications to metabolomics and COPD phenotypes

**Charlie M. Carpenter**[1]*, **Weiming Zhang**[2‡], **Lucas Gillenwater**[3], **Cameron Severn**[1], **Tusharkanti Ghosh**[1], **Russell Bowler**[4], **Katerina Kechris**[1], **Debashis Ghosh**[1]

**1** Department of Biostatistics and Informatics, University of Colorado Denver, Anschutz Medical campus, Denver, Colorado, United States of America, **2** Syneos Health, Morrisville, North Carolina, United States of America, **3** Computational Bioscience Program, University of Colorado Denver, Anschutz medical campus, Denver, Colorado, United States of America, **4** Department of Medicine, National Jewish Health, Denver; University of Colorado Denver, Anschutz Medical Campus, Denver, Colorado, United States of America

‡ Previously affiliated with Anschutz Medial Campus, Denver, Colorado United States of America
* charles.carpenter@cuanschutz.edu

## Abstract

High-throughput data such as metabolomics, genomics, transcriptomics, and proteomics have become familiar data types within the "-omics" family. For this work, we focus on subsets that interact with one another and represent these "pathways" as graphs. Observed pathways often have disjoint components, i.e., nodes or sets of nodes (metabolites, etc.) not connected to any other within the pathway, which notably lessens testing power. In this paper we propose the Pathway Integrated Regression-based Kernel Association Test (PaIRKAT), a new kernel machine regression method for incorporating known pathway information into the semi-parametric kernel regression framework. This work extends previous kernel machine approaches. This paper also contributes an application of a graph kernel regularization method for overcoming disconnected pathways. By incorporating a regularized or "smoothed" graph into a score test, PaIRKAT can provide more powerful tests for associations between biological pathways and phenotypes of interest and will be helpful in identifying novel pathways for targeted clinical research. We evaluate this method through several simulation studies and an application to real metabolomics data from the COPDGene study. Our simulation studies illustrate the robustness of this method to incorrect and incomplete pathway knowledge, and the real data analysis shows meaningful improvements of testing power in pathways. PaIRKAT was developed for application to metabolomic pathway data, but the techniques are easily generalizable to other data sources with a graph-like structure.

OPEN ACCESS

**Citation:** Carpenter CM, Zhang W, Gillenwater L, Severn C, Ghosh T, Bowler R, et al. (2021) PaIRKAT: A pathway integrated regression-based kernel association test with applications to metabolomics and COPD phenotypes. PLoS Comput Biol 17(10): e1008986. https://doi.org/10.1371/journal.pcbi.1008986

**Editor:** Carl Herrmann, Heidelberg University, GERMANY

**Received:** April 23, 2021

**Accepted:** October 13, 2021

**Published:** October 22, 2021

**Peer Review History:** PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1008986

**Data Availability Statement:** The R code for simulations and an example workflow with source code is available at https://github.com/

CharlieCarpenter/PaIRKAT. A shiny app can be found at https://csevern.shinyapps.io/pairkat/. The metabolomics data set from the COPDGene Study can be found through PMID: 20214461 or DOI: 10.3109/15412550903499522.

**Funding:** RB was awarded U01 HL089897 and U01 HL089856 from the National Heart, Lung, and Blood Institute, https://www.nhlbi.nih.gov/. KK and DG were awarded U01 CA235488 from from the National Cancer Institute, https://www.cancer.gov/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Author summary

PaIRKAT is a tool for improving testing power on high dimensional data by including graph topography in the kernel machine regression setting. Studies on high-dimensional data can struggle to include the complex relationships between variables. The semiparametric kernel machine regression model is a powerful tool for capturing these types of relationships. They provide a framework for testing for relationships between outcomes of interest and high dimensional data such as metabolomic, genomic, or proteomic pathways. Our paper proposes a kernel machine method for including known biological connections between high dimensional variables by representing them as edges of 'graphs' or 'networks.' It is common for nodes (e.g., metabolites) to be disconnected from all others within the graph, which leads to meaningful decreases in testing power when graph information is ignored. We include a graph regularization or 'smoothing' approach for managing this issue. We demonstrate the benefits of this approach through simulation studies and an application to the metabolomic data from the COPDGene study.

This is a *PLOS Computational Biology* Methods paper.

## Introduction

Metabolomics is the study of the metabolite composition of a cell, tissue, or biological fluid. Leading metabolomic experimental techniques such as liquid or gas chromatography coupled with mass spectrometry (LC-MS or GC-MS) and nuclear magnetic resonance (NMR) spectroscopy can capture the abundance of all metabolites within a cell (the metabolome). These technologies provide high-throughput data similar to other familiar -omics datatypes such as genomics, transcriptomics, and proteomics. An important advantage of metabolomics over other -omics data is its proximity to biological phenotypes[1]. While genomic or proteomic data are vital pieces for understanding the progression from DNA to phenotype, the metabolites are the end products of the enzymatic reactions of a cell[2]. The metabolome is comprised of exogenous (environmentally derived) and endogenous (genetically regulated) metabolites which can be used as biomarkers for the current phenotypic state of a cell or organism.

Like other -omics data, careful considerations of the metabolome's unique characteristics are required to fully leverage it for biological insights. Specifically, metabolites are known to be related directly and indirectly by enzymatic reactions within a metabolomic pathway. Clustering methods have been developed to incorporate this connectivity into the primary analysis to avoid this two-step approach. These include Bayesian methods for metabolite clustering based on peak detection[3,4] and *ad hoc* methods based on singleton metabolite presence[5]. For this work, we choose to group subsets of metabolites that interact with one another and represent these pathways as *graphs* or *networks*. Throughout this paper we will use the term *graph* and *network* interchangeably. Open source databases with metabolomic pathway documentation such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), the Human Metabolome Database (HMDB), Reactome, OmniPath, and WikiPathways are growing resources[6–10], and the pathways within these databases are easily translated to graphs to be used in downstream analyses.

The semiparametric kernel machine regression method[11,12] has gained popularity in many areas of biomedical research such as genomics, microbiome analysis, and neuroimaging

[13–15]. One reason for its popularity is that it provides a computationally scalable method of classification and regression through the introduction of a *kernel* function. Another is that it provides a setting for formal statistical estimation and testing procedures for high-dimensional data sources, often using a score statistic. Formal statistical tests are useful for metabolomic research, as a goal is often identifying specific metabolites and pathways for further inquiry. At a high level, kernel machines test for relationships between an outcome and a set of predictors by testing if variation between the two correspond with one another.

A hurdle more unique to metabolomics is the high levels of sparsity in individual metabolites and pathway connectivity. While metabolomic databases (e.g., KEGG, HMDB) are growing, none are considered complete. Data generating techniques like LC-MS and GC-MS are also imperfect technologies that may miss metabolite abundances that are too low[16]. Thus, pathway representations of metabolomic data are often sparse and disconnected, i.e., nodes or sets of nodes are not connected to any other within the pathway.

Disjoint nodes are of concern for graph-structured data. Techniques that force graphs to be fully connected by making small, uniform changes to the structure have been suggested for handling this issue[17,18]. However, it is understood that these alterations impose new challenges by changing the subspaces spanned by the graph. Works by Schaid [19] as well as Freytag et al. [20] developed a network-based kernel where similarity is defined directly from the network structure. These methods and others like it are tailored to genome-wide association studies and not applicable to other omics data. Freytag also imposes "as much noise as necessary" within the network to ensure positive semidefinite matrices which is something we aim to avoid. In fact, our proposal dampens out noisy features of the graph. The *PIMKL* method works with pathways within the metabolome by combining them through a weighted summed kernel[21]. These weights provide insight into the importance of each sub-pathway, but this does not surmount to the level of evidence gathered from a direct comparison between specific pathways and phenotype.

In this paper we propose the Pathway Integrated Regression-based Kernel Association Test (PaIRKAT), a new kernel machine regression method for incorporating known pathway information into the semi-parametric kernel regression framework. In addition, PaIRKAT contributes an application of a graph kernel regularization method for overcoming sparse connectivity and disjoint pathways. To our knowledge, this is the first method to incorporate graph regularization into a kernel regression test. PaIRKAT allows for tests of association with phenotypes and the specific pathways while integrating pathway structure, and, instead of adding small amounts of noise, this approach dampens noisy components of a pathway while preserving biologically relevant signals. This leads to improved testing power and better overall biomarker detection. We evaluate these methods through several simulation studies and an application to real metabolomics data from the COPDGene[22] study.

## Results

### Method overview

Here we provide the main steps of PaIRKAT and provide an overview of the ideas behind them. The method is described in full in **Methods and Models**. The primary goal of PaIRKAT is to include the topographical information of graph structured data into the kernel machine regression model. We use the semiparametric kernel machine model[11,12,23] to test for relationships between the phenotype of interest, $\boldsymbol{Y}$, and a high dimensional set, $\boldsymbol{Z}$, while controlling for important covariates, $\boldsymbol{X}$, in the model $g(\boldsymbol{Y}) = \mathbf{X}\boldsymbol{\beta}+h(\boldsymbol{Z})+\epsilon$. In this model $h(\cdot)$ is a positive semidefinite kernel function that transforms $\boldsymbol{Z}$ to an appropriate feature space.

Omics data (metabolomics, genomics, etc.) can often be represented as a graph with edges representing biological interactions between the nodes (metabolites, etc.). Freytag et al. and Schaid both define a kernel directly from the graph structure where higher proximity within the pathway gives a higher similarity score [19,20]. This has been coined a 'guilt by association' approach [24] and has been proven effective empirically. These methods use a map from SNPs to genes to formulate similarity matrices, making them unapplicable to other types of studies. PaIRKAT also uses the 'guilt by association' paradigm but relies on a graph's *regularized normalized Laplacian* as the measure of proximity within the pathway. Then any appropriate kernel can be applied for testing making it more generally applicable than other similar approaches.

We explored the utility of incorporating the Laplacian directly into the kernel machine but found it to be ineffective using simulation studies. Instead, we transform $\tilde{\boldsymbol{L}}$ using methods designed to dampen noisy aspects of a graph while preserving its biologically relevant features [25,26]. The PaIRKAT method is to include this *regularized normalized Laplacian*, $\tilde{\boldsymbol{L}}_R$, in the model through the kernel function as $g(\boldsymbol{Y}) = \mathbf{X}\boldsymbol{\beta} + h(\boldsymbol{Z}\tilde{\boldsymbol{L}}_R) + \epsilon$. Tests for relationships between $\boldsymbol{Y}$ and $h(\boldsymbol{Z}\tilde{\boldsymbol{L}}_R)$ are performed using an adjusted score statistic[23] and Davies' method for estimating distributions of linear combinations of $\chi^2$ variables[27].

## Simulation results

A complete description of our simulation study can be found in **Methods and Models**, but we give a brief synopsis of the simulation scheme. We first randomly generated a graph. Second, we randomly generated features, $\boldsymbol{Z}$, from multivariate normal distribution with a covariance structure derived from the graph. Lastly, we randomly generated a normally distributed outcome, $\boldsymbol{Y}$, with a mean based on a linear relationship between the columns of $\boldsymbol{Z}$. We performed tests ignoring graph topography, including graph topography in the kernel function via the normalized Laplacian ($\tilde{\boldsymbol{L}}$), and our proposed method PaIRKAT of including graph topography in the kernel function via the regularized Laplacian ($\tilde{\boldsymbol{L}}_R$). Our simulations aimed to assess how sensitive our method is to incomplete and/or incorrect graph information. We also compare the power of our method to two simple competing approaches: an F-test on all principal components (PCs) of $\boldsymbol{Z}$ [28] and the minimum Simes' adjusted p-value[29] from univariate tests on $\boldsymbol{Z}$ (Univariate Simes).

Type I error rates for PaIRKAT are summarized in Tables 1, 2, 3 and 4. The type I error rates for tests using a graph's normalized Laplacian, $\tilde{\boldsymbol{L}}$ (see **Methods and Models** section for

**Table 1. Type 1 error rates using all pathway information, i.e., no nodes or edges were dropped for these simulations.** "*Perfect*" indicates calculating $\tilde{L}_R$ from the graph used to generate the data. "*Mismatch*" indicates the percentage of direct edges that were incorrect. Error rates were calculated from score tests on 1000 simulated data sets. All simulations used graphs with 15, 30, or 45 nodes. "*Complete Mismatch*" indicates 100% mismatch.

| | **Pathway size** | | |
|---|---|---|---|
| | 15 | 30 | 45 |
| *Perfect* | 0.0482 | 0.0529 | 0.0568 |
| *10% Mismatch* | 0.0498 | 0.0494 | 0.0474 |
| *40% Mismatch* | 0.0487 | 0.0525 | 0.0464 |
| *70% Mismatch* | 0.0502 | 0.0512 | 0.0511 |
| *Complete Mismatch* | 0.0487 | 0.0511 | 0.0494 |
| *No Pathway* | 0.0580 | 0.0540 | 0.0530 |
| *Principal Component* | 0.0484 | 0.0543 | 0.0558 |
| *Univariate Simes* | 0.0490 | 0.0513 | 0.0507 |

https://doi.org/10.1371/journal.pcbi.1008986.t001

**Table 2. Type 1 error rates using pathways with 5% missing edges.** Error rates were calculated from score tests on 1000 simulated data sets using graphs with 15, 30, or 45 nodes. The graph used to simulate $\boldsymbol{Z}$ and $\boldsymbol{Y}$ was of medium edge density, while the graph used to test was of low density. The low-density graphs are drawn from the Barabasi-Albert model with edge density 0.13, 0.07, and 0.04 for graphs with 15, 30, and 45 nodes, respectively. Medium edge density graphs are created by giving any 2 nodes without a direct edge between them a 5% chance of becoming directly connected. This creates graphs with an average edge density of 0.18, 0.12, and 0.09 for graphs with 15, 30, and 45 nodes, respectively. "*Perfect*" indicates calculating $\tilde{\boldsymbol{L}}_R$ from the graph without changing remaining edges. "*Mismatch*" indicates the percentage of remaining direct edges that were incorrect. "*Complete Mismatch*" indicates 100% mismatch.

| | **Pathway size** | | |
|---|---|---|---|
| | 15 | 30 | 45 |
| *Perfect Network* | 0.0497 | 0.0491 | 0.0463 |
| *10% Mismatch* | 0.0478 | 0.0479 | 0.0485 |
| *40% Mismatch* | 0.0465 | 0.0510 | 0.0536 |
| *70% Mismatch* | 0.0518 | 0.0523 | 0.0486 |
| *Complete Mismatch* | 0.0491 | 0.0539 | 0.0463 |
| *No Network* | 0.0480 | 0.0440 | 0.0390 |
| *Principal Component* | 0.0507 | 0.0489 | 0.0494 |
| *Univariate Simes* | 0.0515 | 0.0494 | 0.0477 |

https://doi.org/10.1371/journal.pcbi.1008986.t002

definition), are summarized in S1, S2, S3, and S4 Tables. The type I error rate of ≈0.05 is maintained throughout all simulation scenarios.

The power curves for all pathway structures and competing methods while simulating complete knowledge, missing edges, and missing nodes are displayed in Fig 1. Having a perfect pathway structure provides the most power. Relationships between an outcome and pathway are easier to detect in larger pathways. The more incorrect direct edges in the pathway, the lower the overall power. The univariate Simes was improved by including $\tilde{\boldsymbol{L}}_R$. Using the PCs of $\boldsymbol{Z}$ and $\boldsymbol{Z}\tilde{\boldsymbol{L}}_R$ gave the exact same power, which is expected from a basis transformation, and performed similarly to a completely incorrect edge structure. Clearly, any correct information from the graph improved power overall. We also see that increasing the overall signal to noise ratio improves power for all pathway structures (Fig 2). PaIRKAT ($\tilde{\boldsymbol{L}}_R$) achieves approximately

**Table 3. Type 1 error rates using pathways with 15% missing edges.** Error rates were calculated from score tests on 1000 simulated data sets using graphs with 15, 30, or 45 nodes. The graph used to simulate $\boldsymbol{Z}$ and $\boldsymbol{Y}$ was of high edge density, while the graph used to test was of low density. The low-density graphs are drawn from the Barabasi-Albert model with edge density 0.13, 0.07, and 0.04 for graphs with 15, 30, and 45 nodes, respectively. High edge density graphs are created by giving any 2 nodes without a direct edge between them a 15% chance of becoming directly connected. This creates graphs with an average edge density of 0.26, 0.21, and 0.19 for graphs with 15, 30, and 45 nodes, respectively. "*Perfect*" indicates calculating $\tilde{\boldsymbol{L}}_R$ from the graph without changing remaining edges. "*Mismatch*" indicates the percentage of remaining direct edges that were incorrect. "*Complete Mismatch*" indicates 100% mismatch.

| | **Pathway size** | | |
|---|---|---|---|
| | 15 | 30 | 45 |
| *Perfect Network* | 0.0508 | 0.0538 | 0.0456 |
| *10% Mismatch* | 0.0541 | 0.0519 | 0.0521 |
| *40% Mismatch* | 0.0495 | 0.0486 | 0.0478 |
| *70% Mismatch* | 0.0514 | 0.0506 | 0.0524 |
| *Complete Mismatch* | 0.0504 | 0.0523 | 0.0490 |
| *No Network* | 0.0430 | 0.0530 | 0.0510 |
| *Principal Component* | 0.0525 | 0.0509 | 0.0481 |
| *Univariate Simes* | 0.0499 | 0.0491 | 0.0459 |

https://doi.org/10.1371/journal.pcbi.1008986.t003

**Table 4. Type 1 error rates using pathways with dropped nodes.** Error rates were calculated from score tests on 1000 simulated data sets using graphs 15, 30, or 45 nodes initially. The graph used to simulate $\boldsymbol{Z}$ and $\boldsymbol{Y}$ contained all nodes. Nodes with degree below the 25th percentile within a graph had a 25% chance of being dropped before testing. "*Perfect*" indicates calculating $\tilde{\boldsymbol{L}}_R$ from the graph without changing edges between remaining nodes. "*Mismatch*" indicates the percentage of direct edges between remaining nodes that were incorrect. "*Complete Mismatch*" indicates 100% mismatch.

| | Pathway size | | |
|---|---|---|---|
| | 15 | 30 | 45 |
| *Perfect Network* | 0.0480 | 0.0513 | 0.0494 |
| *10% Mismatch* | 0.0499 | 0.0489 | 0.0476 |
| *40% Mismatch* | 0.0492 | 0.0495 | 0.0501 |
| *70% Mismatch* | 0.0522 | 0.0511 | 0.0500 |
| *Complete Mismatch* | 0.0481 | 0.0488 | 0.0483 |
| *No Network* | 0.0420 | 0.0490 | 0.0530 |
| *Principal Component* | 0.0505 | 0.0476 | 0.0501 |
| *Univariate Simes* | 0.0481 | 0.0502 | 0.0502 |

https://doi.org/10.1371/journal.pcbi.1008986.t004

80% power at a signal to noise ratio around 0.32, whereas ignoring network information requires a signal to noise ratio over twice that, about 0.70 and only including the Laplacian never achieves 80% power (Fig 2). The univariate Simes' test performed as well as PaIRKAT

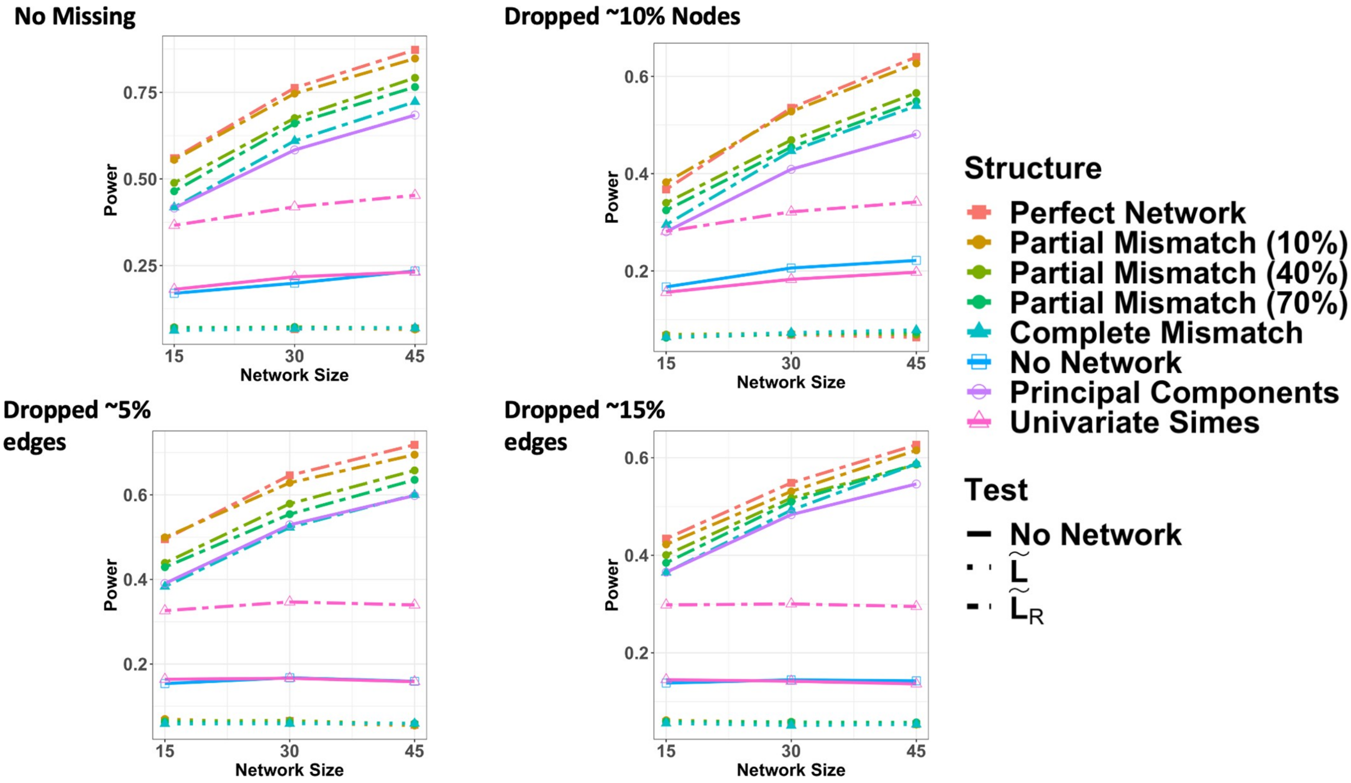

**Fig 1. Power curves from the four pathway *knowledge* and 6 pathway *structure* simulation scenarios.** Power curves were all calculated from score tests on 1000 simulated data sets using graphs with 15, 30, or 45 nodes. Power curves assuming complete pathway knowledge with no dropped edges or nodes are displayed in a). For (b) and (c), the graph used to simulate $\boldsymbol{Z}$ and $\boldsymbol{Y}$ was of medium or high density, respectively, while the graph used to test was of low density. Medium and high edge density graphs used for data generation had ~5% and ~15% more edges, respectively, than the low-density graph used for testing. The power curve generated assuming missing nodes (d) used all graph nodes to generate $\boldsymbol{Z}$ and $\boldsymbol{Y}$. Then nodes (and corresponding columns of $\boldsymbol{Z}$) with degree below the 25th percentile within a graph had a 25% chance of being dropped before testing.

https://doi.org/10.1371/journal.pcbi.1008986.g001

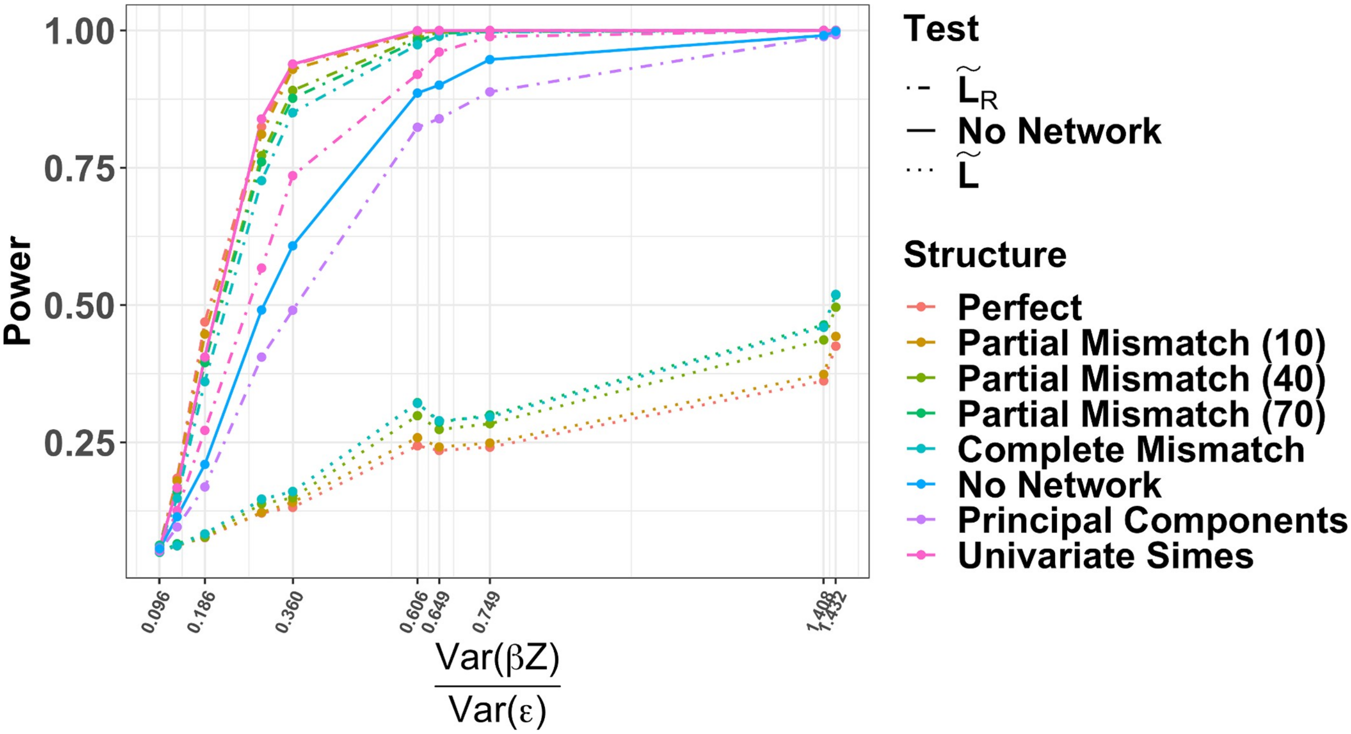

**Fig 2. Signal to Noise Ratio. Power curves from increasing the signal to noise ratio while assuming complete pathway knowledge.** The signal to noise ratio was calculated as the as the ratio between the overall variance in $\boldsymbol{Y}$, $Var(\beta_0 + \sum_{j=1}^{p} \beta_j Z_{ij})$, and the overall residual variance, $Var[Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j Z_{ij})]$. Each power calculation comes from score tests on 1000 simulated data sets using graphs with 30 nodes.

https://doi.org/10.1371/journal.pcbi.1008986.g002

with perfect pathway knowledge. This is unsurprising since all $\boldsymbol{z}_i$ are related to the outcome in our simulations.

## COPDGene analysis results

A complete description of these analyses can be found in **Methods and Models**, but here we give a brief description of the outcome variables we analyzed. We create models for two phenotypes from the COPDGene study[22]: (1) percent emphysema and (2) the ratio of post-bronchodilator forced expiratory volume at one second divided by forced vital capacity ($FEV_1$/FVC). To normalize $FEV_1$/FVC, we use the following log ratio transformation, $\log\left(\frac{FEV_1/FVC}{1-FEV_1/FVC}\right)$. This is referred to as the "log $FEV_1$/FVC ratio" for simplicity. We test for associations between 28 pathways and each outcome under the same three conditions in the simulation study: ignoring graph topography, including graph topography via the normalized Laplacian ($\tilde{\boldsymbol{L}}$), and our proposed method PaIRKAT of including graph topography via the regularized Laplacian ($\tilde{\boldsymbol{L}}_R$).

Including the metabolites' regularized graphs had large impacts on the associations between the log FEV1/FVC ratio and several subsets of metabolites. For the 28 pathways tested, power was improved for 17 pathways when using PaIRKAT vs. using $\tilde{\boldsymbol{L}}$ or ignoring pathway information. Of note, the strength of the associations between the log FEV1/FVC ratio and the *ABC transporters*, the *arginine and proline metabolism*, the *cysteine and methionine metabolism*, the *pyrimidine metabolism*, the *glycine, serine, and threonine metabolism*, and the *neuroactive ligand-receptor interaction* metabolite subsets increased dramatically. The average p-value was

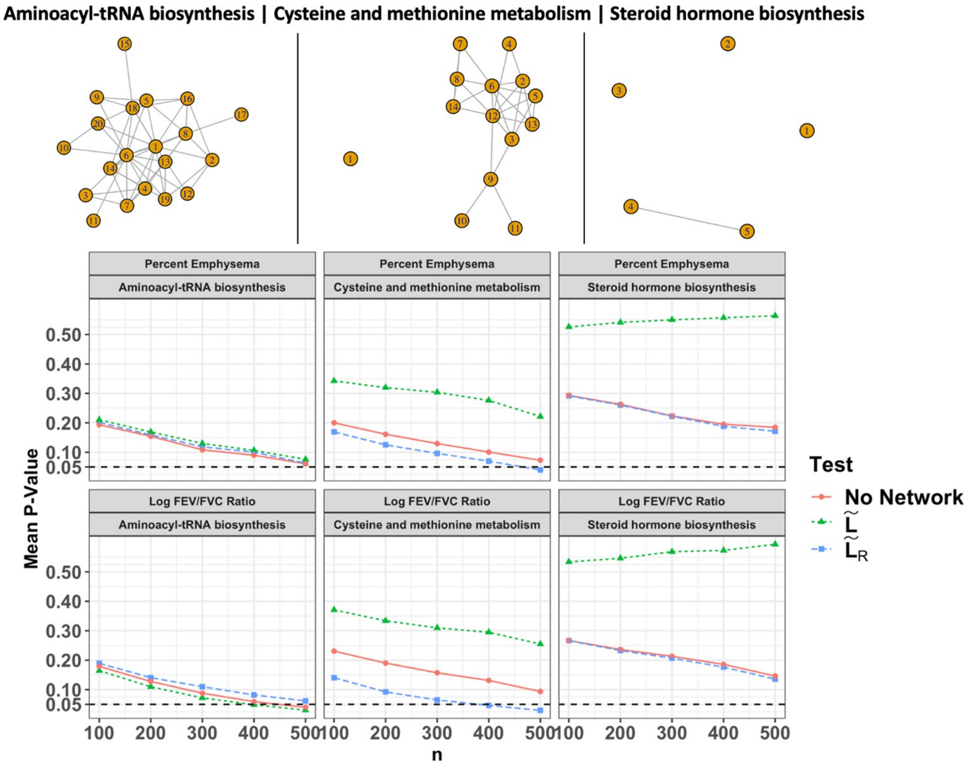

**Fig 3. Selected results from COPDGene subset analysis.** Average p-values from kernel regressing tests that do not include pathway information (No Laplacian, red circles), include pathway information through a normalized Laplacian ($\tilde{L}$, green triangles), and include pathway information through a regularized normalized Laplacian ($\tilde{L}_R = (I + \tau\tilde{L})^{-1}$, blue squares) are displayed. P-values were averaged over 100 random subsets of size 100, 200, 300, 400, and 500 from the COPDGene dataset. $\tau$ was set to 1 for all tests that used $\tilde{L}_R$. The 3 pathways selected illustrate the expected results in fully connected (left), partially disconnected (middle), and sparse (right) graphs.

https://doi.org/10.1371/journal.pcbi.1008986.g003

also lower for 12 pathways with using $\tilde{L}$ vs. ignoring pathway information. S1 Fig displays the p-values from the kernel regression tests for associations between the log FEV1/FVC and the 28 pathways of interest for each subsample size.

Including the metabolites' regularized graphs also had impacts on the associations between percent emphysema and several subsets of metabolites. For the 28 subsets of metabolites tested, power was improved for 17 pathways when using PaIRKAT vs. including $\tilde{L}$ or ignoring pathway information. Of note, the strength of the associations between percent emphysema and the *ABC transporters*, the *β-alanine metabolism*, the *neuroactive ligand-receptor interaction*, the *glycine*, *serine and threonine metabolism*, and the *histidine metabolism* metabolite subsets increased dramatically when using PaIRKAT vs. ignoring pathway information. The average p-value was also lower for the same 5 pathways with using $\tilde{L}$ vs. ignoring pathway information. However, there was still not a significant result from any method for 4 of these pathways, and PaIRKAT provided similar power for the fifth. S2 Fig displays the p-values from the kernel regression tests for associations between percent emphysema and the 28 pathways of interest for each subsample size.

Fig 3 displays results from 3 pathways selected to illustrate PaIRKAT's impact on power for fully connected (left column), partially disconnected (middle column), and sparse (right column) graphs. For the *steroid hormone biosynthesis* pathway, an almost completely sparse pathway, we see virtually no differences between PaIRKAT and ignoring pathway connectivity. We also see relatively small differences between all three methods for the fully connected *aminoacyl-tRNA biosynthesis* pathway. The major impacts from PaIRKAT come when there are a few nodes or node subsets disjoint from the rest of the graph, as we see in the *cysteine and methionine metabolism*.

## Discussion

We have developed PaIRKAT, a method for incorporating pathway information under a kernel regression framework. Other methods to incorporate pathway connectivity via graph operations have been developed[20,21,26,30–32]. PaIRKAT enables the researcher to test on specified pathways instead of aggregating all pathways through a weighted kernel as in[21,30]. It can also handle disjointed pathways without adding in artificial noise to the network as in [17,18,20]. This allows the investigator to compile information from multiple sources, e.g., KEGG and HMDB. The regression framework also expands upon a method developed for classification[26]. It should be noted that the kernel framework is testing a global null, i.e., if any node covaries with the outcome the null hypothesis is rejected. See Goeman and Buhlmann[33] for a full discussion on whether or not this approach is appropriate for pathway based hypotheses.

Pathway misspecifications from incomplete data collection or imperfect canonical pathways within databases are common hurdles in -omics studies. We explored the sensitivity of the method by simulating data assuming incorrect pathway structures and incomplete pathway knowledge. These studies show that our method is highly robust to pathway misspecifications. In smaller pathways, we see that the partially mismatched structure with ~10% of direct edges being incorrect does as well as the perfect network structure. This is likely due to the very small change from the perfect structure in these cases, as a graph with only 15 nodes could easily be unchanged with only a 10% chance to change an edge. Furthermore, even with incorrect or incomplete pathway information, our method provides significantly improved power over ignoring pathway information while maintaining an appropriate type I error rate. We believe this is because many indirect connections between nodes are preserved, and these connections still provide more accurate information than incorrectly assuming independence among nodes.

One benefit of using PaIRKAT is improved power to identify pathways that are associated with clinical phenotypes. For example, an application to the COPDGene dataset using KEGG's database of metabolic pathways also illustrated PaIRKAT's ability to improve testing power over simply treating metabolites as independent (Figs 3, S1, and S2). The regularization technique was also able to handle pathways with few metabolites and/or disjoint components. Several tests had a notables boost in power from including pathway connectivity for both percent emphysema and the log $FEV_1$/FVC ratio, and most pathways have been previously associated with COPD and lung function. Huang et al. linked environmental exposures, COPD risk, and metabolomic pathways, and found associations between COPD and the *histidine metabolism*, *cysteine and methionine metabolism*, and *β-alanine metabolism* pathways[34]. The *glycine*, *serine and threonine metabolism*, *aminoacyl-tRNA biosynthesis*, *pyrimidine metabolism*, *pantothenate and CoA biosynthesis*, pathways have all previously been associated with asthma[35]. The *β*-Alanine metabolism, ABC transporters, purine metabolism, pantothenate and CoA biosynthesis pathways were all differentially associated with COPD subclasses for patients with lung cancer[36]. Another study of the COPDGene dataset[22] using a two-step pathway enrichment approach found that the *purine metabolism*, *mineral absorption*, *arginine biosynthesis*, *aminoacyl-tRNA biosynthesis*, *ABC transporters* and *glycine*, *serine and threonine metabolism* pathways were all associated with various measures of lung function and increased COPD exacerbations[37]. The three *ABC transporters* have also been shown to be related to COPD in several murine knockout and human studies (see Chai et al.[38]). Finally, the *arginine biosynthesis* pathway has also been associated with COPD in multiple studies [39,40].

### Graph information

We used a non-proprietary version of KEGG available in R. The proprietary version of this database has more up to date information and could have resulted in different pathway structures for the COPDGene data set. There is also a substantial literature on data driven methods for deriving networks from omics data [41–47]. Chai et al. provide a nice review[48]. We leave the investigation of how these data-driven methods interact with ours to future research.

### Impacts of regularization

In simulation studies and real data analyses we saw meaningful improvements in power by including pathway information through a graph's regularized normalized Laplacian, (PaIRKAT) when compared to ignoring the pathway information or using $\tilde{\boldsymbol{L}}$. PaIRKAT was essential to maintaining testing power when graphs had disjoint nodes or sub-graphs. Using the normalized Laplacian, $\tilde{\boldsymbol{L}}$, hindered testing performance compared to using PaIRKAT or ignoring the pathway information when a graph was disconnected. In connected graphs PaIRKAT, using $\tilde{\boldsymbol{L}}$, and ignoring the pathway information all performed similarly in the real data analyses (Fig 3).

It is well established that $\tilde{\boldsymbol{L}}$ is a symmetric and positive semidefinite matrix with eigen values $0 \leq \lambda_1, \lambda_2, \ldots, \lambda_p \leq 2$, where the number of $\lambda_i = 0$ is the number of disjoint components of the undirected graph $G$ (see **Methods and Models**). Therefore, graphs with very low connectivity, meaning many $\lambda_i = 0$, will not be as impacted by regularization since all $r^{-1}$ $(\lambda_i = 0) = a$ for some scalar $a$. In words, there is no extra information from a graph when most nodes are disconnected from one another (e.g., Fig 3, right column).

One limitation of this study is our focus on the Gaussian kernel. There has been success with other kernels for high dimensional data such as ones tailored to the data type [14,20] or simple linear and weighted linear kernels [23,49–51]. We have shown that including that including pathway information can improve the power of the Gaussian kernel and leave the impacts on other kernels to future work.

### Summary

In summary, our proposed method serves as a framework for including pathway information into a kernel machine regression test. We developed this method for application to metabolomic pathway data, but the techniques are easily generalizable to other data sources with a graph-like structure. It is important to examine the structure of a graph before applying a regularization step. Unique challenges arose from the sparsity present in many metabolomic pathways which can greatly hinder performance. We implement a graph regularization kernel to handle disconnected pathways. This regularization step is novel in the application of graph-based kernel machine regression to biological data. Our simulation studies illustrate the robustness of this method to improper and incomplete pathway knowledge. The method presented can provide powerful tests for associations between biological pathways and phenotypes of interest and will be helpful in identifying novel pathways for targeted clinical research.

## Methods and models

### The Kernel machine model

We assume that the data are properly filtered, imputed, and normalized for the methods described in this paper. Consider a dataset with observations from $n$ subjects. Let $\boldsymbol{Y}$ be an $n \times 1$

vector representing a continuous or discrete phenotype of interest. Also let $\boldsymbol{X}$ be a $n\times q$ matrix of clinical covariates and $\boldsymbol{Z}$ be an $n\times p$ matrix of graph structure data. The phenotype can then be modeled through the following semiparametric model

$$g(\boldsymbol{Y}) = \mathbf{X}\boldsymbol{\beta} + h(\boldsymbol{Z}) + \epsilon, \qquad 1)$$

where $g$ is either the identity or *logit* link function, $\boldsymbol{\beta}$ is a $q\times1$ vector of regression coefficients, $\epsilon$ is an $n\times1$ vector of normally distributed error terms, and $h$ is a kernel function. There are no parametric assumptions placed on $h$ except that it lies in some feature space. This more relaxed requirement from the kernel regression provides flexibility and robustness to model misspecification. Another key advantage of introducing the kernel function is its ability to capture nonlinear relationships between the phenotype ($\boldsymbol{Y}$) and the metabolome ($\boldsymbol{Z}$) in a computationally tractable manner.

These relationships are assumed to exist in some feature space that is generated by a positive definite kernel function $K(\cdot,\cdot)$. The kernel function can be understood as a feature map that delivers the dot product between $z_i$ and $z_j$ within the features space, i.e., $K(z_i, z_j) = \langle\phi(z_i),\phi(z_j)\rangle$, where $\phi(\cdot)$ is the transformation to the feature space and $\langle\cdot,\cdot\rangle$ is the dot product. The representer theorem allows $h(\boldsymbol{Z})$ to be represented through the kernel function $K(\cdot,\cdot)$ as $h(\cdot) = \sum_{i=1}^{n} \alpha_i K(\cdot, \boldsymbol{z}_i, \rho)$ for some coefficients $\alpha_i \in \mathbb{R}$. More detailed derivations can be found in texts by Schölkopf and Smola[52] as well as Cristianini and Shawe-Taylor[53].

The kernel function $K$ can be thought of as a measurement of similarity between two individuals. Common choices for kernel functions are the *Linear Kernel*: $K(z_i, z_j) = z_i^T z_j$ (the dot product), the *dth Polynomial Kernel*: $K(z_i, z_j, \rho) = (z_i^T z_j + \rho)^d$, and the *Gaussian Kernel*: $K(z_i, z_j, \rho) = \exp\{-||z_i - z_j||^2/\rho\}$, where $||\cdot||$ is the Euclidean ($L_2$) norm. For this work, we employ the Gaussian kernel and use the median of all pairwise Euclidean distances between all $z_i$ and $z_j$ as an empirical estimate of $\rho$. We choose to work with the Gaussian kernel since it is a *characteristic* kernel, a desirable property meaning that probability measures embedded through the kernel function are unique.

**Kernel-based score test.** Liu et al. show a connection between kernel machine regression and linear mixed models for semiparametric modeling of high dimensional data [11,12]. The parameters $\boldsymbol{\beta}$ and $h(\boldsymbol{Z})$ can be estimated by maximizing the scale penalized likelihood

$$L(\boldsymbol{\beta}, h) = -\frac{1}{2}\sum_{i=1}^{n}[y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - h(z_i)]^2 - \frac{1}{2}\lambda||h||^2 \qquad 2)$$

$$= -\frac{1}{2}\sum_{i=1}^{n}[y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - \sum_{j=1}^{n}\alpha_j K(z_i, z_j)]^2 - \frac{1}{2}\lambda\boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha}, \qquad 3)$$

where $\boldsymbol{K} = K(z_i, z_j, \rho)$ is the semi-positive definite kernel function of choice. $h(\boldsymbol{Z})$ can then be understood as subject specific random effects with mean 0 and variance $\tau\boldsymbol{K}$. Testing for an association between phenotype and pathway is then equivalent to testing the null hypothesis $H_0$:$\tau = 0$ vs $H_1$:$\tau>0$. We adopt Chen et al.'s adjusted kernel association test adjusted for small samples, which is common for many omics studies [23]. The standard quadratic score statistic for kernel association tests,

$$Q(\boldsymbol{\beta}, \sigma, \rho) = \frac{1}{\sigma^2}(\boldsymbol{Y} - \boldsymbol{X\beta})^T\boldsymbol{K}(\boldsymbol{Y} - \boldsymbol{X\beta}), \qquad 4)$$

is adjusted to account for the high variability in estimates of $\sigma^2$ when $n$ is small. The distribution of $Q$ under the null model is then approximated as a weighted sum of $\chi^2$ variables using Davies method [27].

## Graph laplacian

A network or graph, $G = \{V, E\}$, is a mathematical representation of any interconnected structure through a set $V$ of $p$ nodes (or vertices) and a set $E$ of edges, where the elements of $E$ are pairs $\{u, v\}$ of distinct vertices, $u, v \in V$. When applied to omic pathways, nodes represent individual metabolites, genes, microbes, etc. within the pathway and edges represent direct interactions/reactions between them.

Two important features of any graph are its adjacency matrix, $\boldsymbol{A}$, and degree matrix, $\boldsymbol{D}$. $\boldsymbol{A}$ is a $p \times p$ matrix that is non-zero when an edge exists between two vertices. $\boldsymbol{D}$ is a $p \times p$ diagonal matrix with $\boldsymbol{D}_{[i,i]}$ representing the number of nodes connected to node $i$. For this work, we represent pathways using undirected unweighted graphs, i.e., there is no ordering to the vertices defining an edge and $\{u, v\} = \{v, u\} \in E$. This means $\boldsymbol{A}$ will be a symmetric matrix with all entries either 1 or 0. Using these features, we can calculate a graph's *Laplacian* $\boldsymbol{L} := \boldsymbol{D} - \boldsymbol{A}$ and its *normalized Laplacian* $\tilde{\boldsymbol{L}} := \boldsymbol{D}^{-1/2} \boldsymbol{L} \boldsymbol{D}^{-1/2} = \boldsymbol{I} - \boldsymbol{D}^{-1/2} \boldsymbol{A} \boldsymbol{D}^{-1/2}$, where $\boldsymbol{I}$ is a $p \times p$ identity matrix.

Both $\boldsymbol{L}$ and $\tilde{\boldsymbol{L}}$ can be regarded as linear operators of functions $\boldsymbol{f}: V \rightarrow \mathbb{R}$ that induce a semi-norm $\|\boldsymbol{f}\|_L = \langle \boldsymbol{f}, \boldsymbol{L}\boldsymbol{f} \rangle = \boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f}$. This semi-norm can be interpreted as a measure of "smoothness" or how much $\boldsymbol{f}$ varies over its domain. Standardizing $\boldsymbol{L}$ by the number of connections per node to obtain $\tilde{\boldsymbol{L}}$ is a common approach in graph theory since $\tilde{\boldsymbol{L}}$ has several well-known and desirable properties. In particular, $\tilde{\boldsymbol{L}}$ is symmetric and positive semidefinite, and its eigenvalues, $\lambda_i$, are bounded such that they satisfy $0 \le \lambda_i \le 2$ for $i \in 1, 2, \ldots p$. Another interesting feature of a graph's normalized Laplacian, $\tilde{\boldsymbol{L}}$, is that the number of disjoint pieces within a graph is captured by the number of $\tilde{\boldsymbol{L}}$s eigen values equal to 0 [54].

## Graph regularization

A key component of PaIRKAT is the ability to handle missing and incorrect information from the graph. Pathway databases may not be complete, and untargeted data generating techniques may not be able capture all components within a pathway. This leaves some pathways with low connectivity and others with completely disconnected nodes. This can lead to a decrease in our power to detect associations between phenotypes and metabolomic pathways. One proposed solution is to simply manipulate the adjacency matrix by adding a small constant to all entries[17,18], i.e. working with a modified adjacency matrix $\tilde{\boldsymbol{A}} = \boldsymbol{A} + tee'$, where $t$ is a non-negative tuning parameter and $e$ is a vector of 1s. This yields a full rank matrix as desired, but we know that the subspace spanned by $\tilde{\boldsymbol{A}}$ is not the correct subspace on which our graph lies.

A more elegant solution can be drawn from Smola and Kondor's work on regularization of graphs[25] in which they draw on parallels between the standard Laplacian operator $\left(\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \cdots + \frac{\partial^2}{\partial x_m^2}\right)$ and the graph Laplacian to design regularization kernels for graphs. Rapaport, et al.[26] took a similar approach to graph smoothing, though this work was done in the context of classification not hypothesis testing. These ideas can be generalized further to represent any metric on a space. That is, for any two observations $i$ and $j$, the inner product can be expressed as $\langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle_{\boldsymbol{M}} = \boldsymbol{z}_i^T \boldsymbol{M} \boldsymbol{z}_j$, where $\boldsymbol{M}$ defines the metric on the vector space based on $\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_{\boldsymbol{M}}$. Purdom[55] presents this argument in the context of a "generalized" principal component analysis using a general metric $\boldsymbol{M}$. This can be seen as an application of a linear kernel on any metric space, whereas we apply the Gaussian kernels for hypothesis testing and, like Rapaport, focus on graph Laplacians for our metric.

For this work, we apply a regularization function to obtain a *regularized* normalized Laplacian: $r(\tilde{\boldsymbol{L}}) \equiv \tilde{\boldsymbol{L}}_R$. Regularizations of the Laplacian can be seen as regularizations of the eigenvalues of $\tilde{\boldsymbol{L}}$, $r(\lambda)$. There are many possible choices for $r$; the only requirement is that $r^{-1}(\lambda) > 0$ for

$\lambda \in [0, 2]$ to ensure $r(\tilde{\boldsymbol{L}}) \geqslant 0$. In classical Fourier analysis the size of $\lambda_i \in [0, 2]$ is directly proportional to the frequency of component $i$ within Fourier space, which translates to the degree of noise within the system. This intuition tells us to limit $r^{-1}(\lambda)$ to monotonically increasing functions in order to impose higher penalties to more uneven portions of the graph while preserving the lower frequency components, which we assume translate to the prevalent biological signals. Smola and Kondor recommend further limiting choices of $r$ to functions expressible by power series such as a *diffusion* kernel, $r(\tilde{\boldsymbol{L}}) = e^{-\tau/2\tilde{L}}$. See [56] for complete details on the derivation of different regularization functions.

PaIRKAT implements a "linear" regularization function

$$\tilde{\boldsymbol{L}}_R = (I + \tau \tilde{\boldsymbol{L}})^{-1}, \tag{5)}$$

where $\tau > 0$ is a bandwidth parameter and $I$ is a $p \times p$ identity matrix. We choose this regularization for its simplicity and interpretability of $\tau$. Increasing $\tau$ linearly increases the amount of smoothing performed in $r^{-1}(\lambda) = 1 + \tau\lambda$. We can now conduct a kernel machine test while incorporating connectivity within a pathway through $\tilde{\boldsymbol{L}}_R$ into (1) as

$$g(\boldsymbol{Y}) = \mathbf{X}\boldsymbol{\beta} + h(\boldsymbol{Z}\tilde{\boldsymbol{L}}_R) + \epsilon, \tag{6)}$$

where $h$ is a kernel function applied to $\boldsymbol{Z}\tilde{\boldsymbol{L}}_R$ and the other model components are as described in (1). $\boldsymbol{Z}\tilde{\boldsymbol{L}}_R$ is changing $\boldsymbol{Z}$'s basis function to one defined by the Laplacian, with the new basis vectors representing noise dampened through the regularization function. This can be interpreted as transforming each subject's phenotype to a weighted sum of each element where the weights are the elements' proximity to each other within the pathway. This falls under the 'guilt by association' framework as nodes closer to each other will share more information and disconnected nodes will share none. The kernel-based score test can then be applied to obtain powerful tests for associations between connected or disconnected pathways and a phenotype of interest.

## Simulation study

**Simulation scenarios.** We conducted multiple simulation studies to assess whether the proposed method is robust to imperfect pathway information. We assumed 3 different "pathway knowledge" scenarios and 4 different "pathway structure" scenarios (Fig 4). Different pathway knowledge scenarios refer to different types of missing information, whereas pathway structure scenarios refer to different configurations of the "known" nodes and edges. We simulate using both the normalized Laplacian, $\tilde{\boldsymbol{L}}$, and PaIRKAT's regularized normalized Laplacian, $\tilde{\boldsymbol{L}}_R$, as well as ignoring the pathway information. For comparison, we also tested using an F-test on all principal components of $\boldsymbol{Z}$ and $\boldsymbol{Z}\tilde{\boldsymbol{L}}_R$ and the minimum Simes' adjusted p-value of univariate tests [29] on all columns of $\boldsymbol{Z}$ and $\boldsymbol{Z}\tilde{\boldsymbol{L}}_R$.

**Pathway knowledge.** We simulated three different *knowledge* scenarios to represent incomplete pathway database information and/or incomplete data collection.

1. No missing: Assuming the nodes measured (metabolites, genes, etc.) and edges connecting them are a perfect representation of the biological pathway of interest.
2. Missing edges: Assuming that some biological interactions (edges) are missing from the documented pathway. Here we generate a graph $G = \{V, E\}$ according to the Barabasi-Albert model for a "low" edge density. We then give every set $\{u, v\} \notin E$ a 5% or 15% percent chance of being added to $E$ for a "medium" or "high" edge density graph, respectively. $\boldsymbol{Z}$ and $\boldsymbol{Y}$ are then generated from the medium or high edge density graph, but $\tilde{\boldsymbol{L}}$ or $\tilde{\boldsymbol{L}}_R$ is

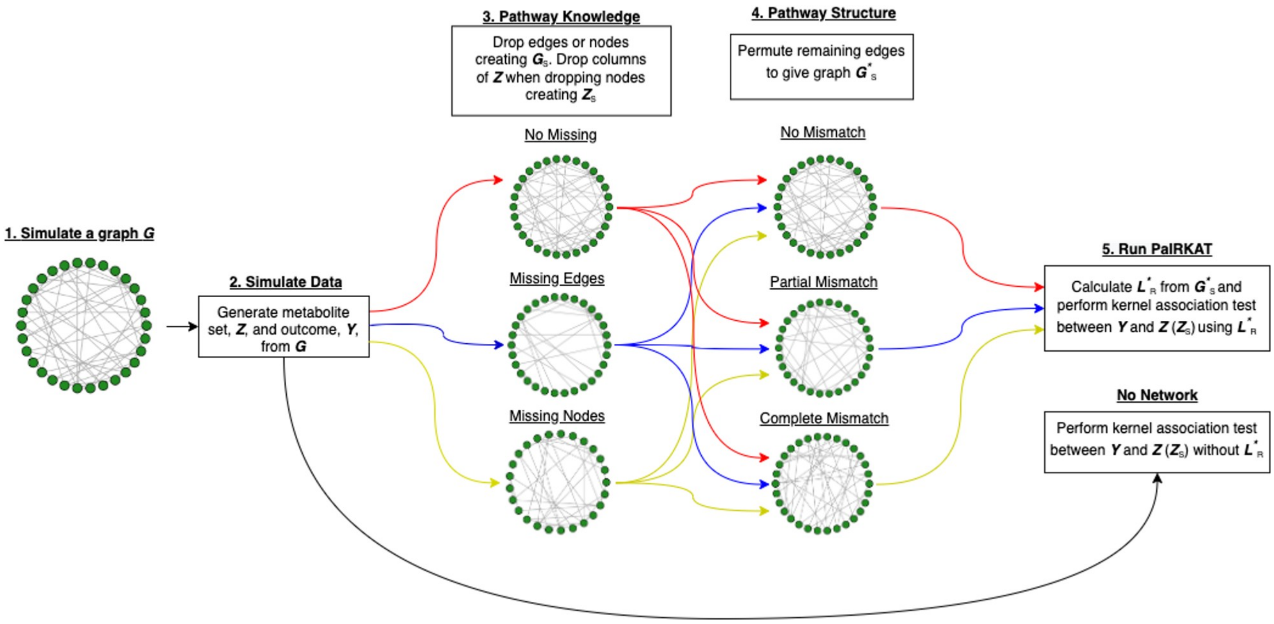
3. Pathway Knowledge
Drop edges or nodes creating $G_s$. Drop columns of $Z$ when dropping nodes creating $Z_s$
4. Pathway Structure
Permute remaining edges to give graph $G^*_s$
No Missing
No Mismatch
1. Simulate a graph G
5. Run PaIRKAT
Calculate $L^*_R$ from $G^*_s$ and perform kernel association test between $Y$ and $Z$ ($Z_s$) using $L^*_R$
2. Simulate Data
Generate metabolite set, $Z$, and outcome, $Y$, from $G$
Missing Edges
Partial Mismatch
Missing Nodes
Complete Mismatch
No Network
Perform kernel association test between $Y$ and $Z$ ($Z_s$) without $L^*_R$

**Fig 4. Flowchart of simulation procedure.** We (1) simulate a graph $G$, (2) generate $\boldsymbol{Z}$ and $\boldsymbol{Y}$ from $G$, (3) drop nodes or edges from $G$ to give a smaller graph $G_s$ (drop corresponding columns of $\boldsymbol{Z}$ when dropping edges to create $\boldsymbol{Z}_s$), (4) permute edges to create an improperly structured graph $G^*_s$, (5) calculate the regularized normalized Laplacian $\tilde{\boldsymbol{L}}^*_R$ from $G^*_s$, and finally (5) test for an association between $h(\boldsymbol{Z}\,\tilde{\boldsymbol{L}}^*_R)$ (or $h(\boldsymbol{Z}_s\,\tilde{\boldsymbol{L}}^*_R)$) and $\boldsymbol{Y}$ in the model $\boldsymbol{Y} = \beta_0 + h(\boldsymbol{Z}_s\,\tilde{\boldsymbol{L}}^*_R)$. For the "no network" simulations, we only use step (1), step (2) and step (5) without including $\tilde{\boldsymbol{L}}^*_R$.

https://doi.org/10.1371/journal.pcbi.1008986.g004

calculated from the original "low" edge density graph. Examples of these graphs are shown in S3 Fig.

3. Missing nodes: Assuming that some of the nodes (and hence their edges) are missing from the documented pathway. Here a graph is used to generate $\boldsymbol{Z}$ and $\boldsymbol{Y}$. Then nodes with degree below the 25th percentile have a 25% chance of being removed before calculating $\tilde{\boldsymbol{L}}$ or $\tilde{\boldsymbol{L}}_R$. The corresponding columns and rows of $\boldsymbol{Z}$ and $\rightarrow \beta$ are removed as well. Examples of these graphs are shown in S4 Fig.

**Pathway structures.** After we simulate a pathway knowledge scenario, we alter the pathway *structure* to represent incorrect edge connections within a database. Examples of structures 1, 2, and 3 are displayed in Fig 5.

1. No mismatch: No alterations to graph edges. The graph used to simulate $\boldsymbol{Z}$ and $\boldsymbol{Y}$ is the same graph used to calculate $\tilde{\boldsymbol{L}}$ or $\tilde{\boldsymbol{L}}_R$ (Fig 5, left).
2. Partial Mismatch: a graph, $G_1 = \{V_1, E_1\}$, is used to simulate $\boldsymbol{Z}$ and $\boldsymbol{Y}$. This graph's edges are permuted such that any edge $\{u, v\} \in E_1$ has a 10%, 40%, or 70% chance of being changed to some $\{u, w\} \notin E_1$; i.e., approximately 10%, 40%, or 70% of direct edges are incorrect before calculating $\tilde{\boldsymbol{L}}$ or $\tilde{\boldsymbol{L}}_R$ (Fig 5, middle).
3. Complete Mismatch: a network $G_1 = \{V_1, E_1\}$ is used to simulate $\boldsymbol{Z}$ and $\boldsymbol{Y}$. A new random graph, $G_2$, is then draw and forced to have no edges that match $G_1$, i.e., $V_1 = V_2$ but if $\{u, v\} \in E_1$ then $\{u, v\} \notin E_2$. We then calculate $\tilde{\boldsymbol{L}}$ or $\tilde{\boldsymbol{L}}_R$ from $G_2$ (Fig 5, right).

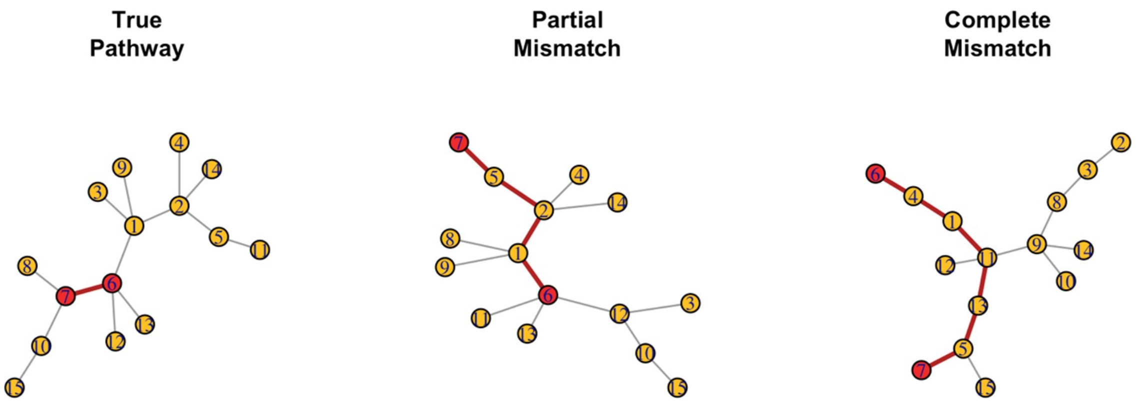

**Fig 5. Examples of the three different pathway *structures*.** Nodes 6 and 7 are highlighted in red to help display the effects of different pathway structures. (Left) The "true" pathway or graph that is used to simulated $\boldsymbol{Z}$ and $\boldsymbol{Y}$. This is the graph used for tests under a "perfect pathway structure" scenario. (Middle) A graph with approximately 40% of the edges from the "true" graph directly connecting the wrong nodes. This is used for tests under a "partial mismatch (40) structure" scenario. (Right) A graph with 0 shared edges with the "true" graph. This is the graph used for tests under a "complete mismatch structure" scenario.

https://doi.org/10.1371/journal.pcbi.1008986.g005

4. No Pathway: a graph is used to simulate $\boldsymbol{Z}$ and $\boldsymbol{Y}$. This connectivity is ignored while testing by not including $\tilde{\boldsymbol{L}}$ or $\tilde{\boldsymbol{L}}_R$ in the kernel function.

All pathway *structures* were considered under each different pathway *knowledge* scenario. The different pathway structures were imposed after simulating under different pathway knowledge assumptions. Each simulated pathway structure and knowledge combination followed 5 steps: (1) simulate a graph $G$, (2) generate $\boldsymbol{Z}$ and $\boldsymbol{Y}$ from $G$, (3) drop nodes and/or edges (based on *knowledge* assumption) from $G$ and $\boldsymbol{Z}$ to give a smaller graph and node set $G_S$ and $\boldsymbol{Z}_S$, (4) alter $G_S$ (based on *structure* assumption) to create a graph $G_s^*$ with improper edge connections, (5) calculate $\tilde{\boldsymbol{L}}^*$ or $\tilde{\boldsymbol{L}}_R^*$ from $G_s^*$, and (5) test for an association between $h(\boldsymbol{Z}_s\ \tilde{\boldsymbol{L}}_R^*)$ and $\boldsymbol{Y}$ in the model $\boldsymbol{Y} = \beta_0 + h(\boldsymbol{Z}_s\ \tilde{\boldsymbol{L}}_R^*)$. See Fig 4 for a flowchart of these simulation scenarios.

**Simulated data.** To evaluate PaIRKAT's overall testing performance and robustness to incorrect pathway information, we simulate data and tests assuming various types of misspecified pathways. All simulations were performed using R[57]. Random graphs were generated using the *igraph*[58] package according to the Barabasi-Albert model[59] with $p$ nodes representing $p$ metabolites within a pathway. The graph's adjacency matrix was converted into a positive definite precision matrix, $\boldsymbol{\Omega}$, using an approached developed by Danaher, et al.[60] and also applied by Shaddox, et al[61]. An $n$ by $p$ matrix of metabolite abundances, $\boldsymbol{Z}$, was then simulated from a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance $\Omega^{-1}$. In this way, node connectivity is captured by $\boldsymbol{\Omega}$. A continuous outcome $Y_i$ was then simulated from a normal distribution with mean 0.26+0.5 $X_1$+0.25 $X_2$+$\Sigma_j\beta_jZ_{ij}$ and variance $\sigma^2$, where $X_1$ was a binary variable, $X_2$ is a uniform random variable, $\sigma^2 = 1.3688^2$. This value for $\sigma^2$ was drawn from observed metabolomics data. The regularization parameter $\tau$ is set to 1 for all simulations. All $\beta_j$ were set to 0 to assess Type I error rates or set to 0.1 to assess power for the different pathway information scenarios described above. Each used 10,000 simulations of graphs of size $p$ = 15, 30, 45 assuming a sample size of $n$ = 160, and a testing level of $\alpha$ = 0.05 was used for all simulations.

## COPDGene data

We analyzed data collected from the COPDGene study [22], a multicenter observational study that collected genetic data as well as multiple measures of lung function to study chronic obstructive pulmonary disease (COPD). Between 2007 and 2011, 10,198 participants with and without chronic obstructive pulmonary disease (COPD) enrolled (Visit 1). A five-year follow up visit took place between 2013 and 2017 (Visit 2). Blood samples were also obtained for -omics analyses from participants who provided consent. In total, 1136 subjects (1040 non-Hispanic white, 96 African American) participated in a metabolomics ancillary study in which they provide fresh frozen plasma collected using an 8.5 mL p100 tube (Becton Dickson) at Visit 2.

## Metabolomics and data processing

P100 plasma was profiled using the Metabolon (Durham, NC, USA) Global Metabolomics platform. Briefly, untargeted liquid chromatography–tandem mass spectrometry (LC–MS/MS) was used to quantify 1392 metabolites and described in[62,63]. A data normalization step was performed to correct variation resulting from instrument inter-day tuning differences: metabolite intensities were divided by the metabolite run day median, then multiplied by the overall metabolite median. It was determined that no further normalization was necessary based on the reduction in the significance of association between the top PCs and sample run day after normalization. Subjects with aggregate metabolite median $z$-scores greater than 3.5 standard deviation from the mean ($n = 6$) of the cohort were removed. Metabolites were excluded if $>$20% of samples were missing values[64]. For the 995 remaining metabolites, missing values were imputed across metabolites with k-nearest neighbors imputation ($k = 10$) using the R package *impute*[65]. As a final step, metabolomic data was natural log transformed and standardized. Linear regression models were fit to each metabolite controlling for white blood cell count, percent eosinophil, percent lymphocytes, percent monocytes, percent neutrophils, and hemoglobin. The partial residuals were then used as the observed metabolomics data. These data are available at Metabolomics Workbench with identifier PR000907.

Four hundred and thirty six of these metabolites had an id in the KEGG database of human pathways, which was accessed using the *keggLink* function from the *KEGGREST* package[66]. These 436 metabolites appear in 161 KEGG pathways, and 28 of these 161 KEGG pathways contained 10 or more metabolites. Edges in a pathway's graph were defined by connections within a pathway from the KEGG reaction database. Note that our filtered dataset did not contain every metabolite within the 28 KEGG pathways selected, and therefore some of the analyzed pathways have less that 10 metabolites.

## Clinical variables

We focus on two COPD phenotypes: (1) percent emphysema and (2) the ratio of post-bronchodilator forced expiratory volume at one second divided by forced vital capacity ($FEV_1$/FVC). Emphysema, a measure of erosion of the distal airspaces, has been linked with the clinical severity of COPD[67]. It is an imaging-based phenotype defined as the 15th percentile lung voxel density in Hounsfield units adjusted for total lung capacity from quantitative CT imaging analyses. $FEV_1$/FVC is a measure of airflow obstruction. To normalize $FEV_1$/FVC, we use the following log ratio transformation, $\log\left(\frac{FEV_1/FVC}{1-FEV_1/FVC}\right)$. After removing incomplete cases we were left with 1,113 complete cases for the $FEV_1$/FVC analysis and 1,065 complete cases for the percent emphysema analysis.

## Analysis

We compared results from tests that included pathway connectivity via $\tilde{\boldsymbol{L}}$, $\tilde{\boldsymbol{L}}_R$, and tests that ignored pathway connectivity for the 28 pathways that had measurements on at least 10 of the metabolites in the pathway. P-values were calculated from a score test as described Section 2 with $\tau = 1$ for PaIRKAT tests. P-values from each method were indistinguishable from one another for both data sets with over 1,000 observations. However, many data sets may not be that large. To demonstrate the differences in performance, 100 random subsets of sizes 100, 200, 300, 400, and 500 were taken from both the log $FEV_1$/FVC ratio and the percent emphysema data sets. All three methods were used to test for associations between phenotype and metabolites within a pathway. The 100 p-values were then averaged to measure the performance of each method. All null models included subject age, sex, BMI, smoking status (current, former, never), pack-years of smoking, and the clinical center as covariates.

## Supporting information

**S1 Fig. Associations between metabolite subsets and log FEV1/FVC ratio.** Average p-values from kernel regressing tests that do not include pathway information (No Laplacian, red circles), include pathway information through a normalized Laplacian ($\tilde{\boldsymbol{L}}$, green triangles), and include pathway information through a regularized normalized Laplacian ($\tilde{\boldsymbol{L}}_R = (I + \tau\tilde{\boldsymbol{L}})^{-1}$, blue squares) are displayed. P-values were averaged over 100 random subsets of size 100, 200, 300, 400, and 500 from the COPDGene dataset. $\tau$ was set to 1 for all tests that used $\tilde{\boldsymbol{L}}_R$.
(TIF)

**S2 Fig. Associations between metabolite subsets and percent emphysema.** Average p-values from kernel regressing tests that do not include pathway information (No Laplacian, red circles), include pathway information through a normalized Laplacian ($\tilde{\boldsymbol{L}}$, green triangles), and include pathway information through a regularized normalized Laplacian ($\tilde{\boldsymbol{L}}_R = (I + \tau\tilde{\boldsymbol{L}})^{-1}$, blue squares) are displayed. P-values were averaged over 100 random subsets of size 100, 200, 300, 400, and 500 from the COPDGene dataset. $\tau$ was set to 1 for all tests that used $\tilde{\boldsymbol{L}}_R$.
(TIF)

**S3 Fig. Examples graphs with high, medium, and low edge densities.** Low density graphs were generated according the Barabasi-Albert model for graph simulation. Medium- and high-density graphs were generated by giving each unconnected node either a 5% or 15% chance of becoming connected, respectively.
(TIF)

**S4 Fig. Example of a graph with missing nodes.** Graphs were generated according to the Barabasi-Albert model. Then any node with degree below the 25th percentile of degrees within the graph had a 25% chance of being dropped.
(TIF)

**S1 Table. Type 1 error rates using complete pathway.** Error rates were calculated from score tests on 1000 simulated data sets. All simulations used graphs with 15, 30, or 45 nodes. No nodes or edges were dropped for these simulations. Pathway information was included in kernel score test through the normalized Laplacian $\tilde{\boldsymbol{L}}$.
(XLSX)

**S2 Table. Type 1 error rates using pathways with 5% missing edges.** Error rates were calculated from score tests on 1000 simulated data sets using graphs with 15, 30, or 45 nodes. The graph used to simulate $\boldsymbol{Z}$ and $\boldsymbol{Y}$ was of medium edge density, while the graph used to test was of low density. The low-density graphs are drawn from the Barabasi-Albert model with edge density 0.13, 0.07, and 0.04 for graphs with 15, 30, and 45 nodes, respectively. Medium edge density graphs are created by giving any 2 nodes without a direct edge between them a 5%

chance of becoming directly connected. This creates graphs with an average edge density of 0.18, 0.11, and 0.09 for graphs with 15, 30, and 45 nodes, respectively. Pathway information was included in kernel score test through the normalized Laplacian $\tilde{\boldsymbol{L}}$.
(XLSX)

**S3 Table. Type 1 error rates using pathways with 15% missing edges.** Error rates were calculated from score tests on 1000 simulated data sets using graphs with 15, 30, or 45 nodes. The graph used to simulate $\boldsymbol{Z}$ and $\boldsymbol{Y}$ was of high edge density, while the graph used to test was of low density. The low density graphs are drawn from the Barabasi-Albert model with edge density 0.13, 0.07, and 0.04 for graphs with 15, 30, and 45 nodes, respectively. High edge density graphs are created by giving any 2 nodes without a direct edge between them a 15% chance of becoming directly connected. This creates graphs with an average edge density of 0.26, 0.21, and 0.19 for graphs with 15, 30, and 45 nodes, respectively. Pathway information was included in kernel score test through the normalized Laplacian $\tilde{\boldsymbol{L}}$.
(XLSX)

**S4 Table. Type 1 error rates using pathways with dropped nodes.** Error rates were calculated from score tests on 1000 simulated data sets using graphs 15, 30, or 45 nodes initially. The graph used to simulate $\boldsymbol{Z}$ and $\boldsymbol{Y}$ contained all nodes. Nodes with degree below the 25th percentile within a graph had a 25% chance of being dropped before testing. Pathway information was included in kernel score test through the normalized Laplacian $\tilde{\boldsymbol{L}}$.
(XLSX)

## Author Contributions

**Conceptualization:** Charlie M. Carpenter, Weiming Zhang, Katerina Kechris, Debashis Ghosh.

**Data curation:** Charlie M. Carpenter, Lucas Gillenwater, Tusharkanti Ghosh, Russell Bowler, Katerina Kechris.

**Formal analysis:** Charlie M. Carpenter.

**Funding acquisition:** Katerina Kechris, Debashis Ghosh.

**Methodology:** Charlie M. Carpenter, Weiming Zhang, Debashis Ghosh.

**Resources:** Charlie M. Carpenter.

**Software:** Charlie M. Carpenter, Cameron Severn.

**Supervision:** Katerina Kechris, Debashis Ghosh.

**Validation:** Charlie M. Carpenter.

**Visualization:** Charlie M. Carpenter.

**Writing – original draft:** Charlie M. Carpenter.

**Writing – review & editing:** Charlie M. Carpenter, Weiming Zhang, Lucas Gillenwater, Tusharkanti Ghosh, Russell Bowler, Katerina Kechris, Debashis Ghosh.

## References

1. Fiehn O. Metabolomics—the link between genotypes and phenotypes. In: Town C, editor. Functional Genomics. Dordrecht: Springer Netherlands; 2002. pp. 155–171. https://doi.org/10.1007/978-94-010-0448-0_11

2. Alonso A, Marsal S, Julià A. Analytical Methods in Untargeted Metabolomics: State of the Art in 2015. Front Bioeng Biotechnol. 2015;3. https://doi.org/10.3389/fbioe.2015.00003 PMID: 25692128

3. Suvitaival T, Rogers S, Kaski S. Stronger findings from mass spectral data through multi-peak modeling. BMC Bioinformatics. 2014; 15: 208. https://doi.org/10.1186/1471-2105-15-208 PMID: 24947013

4. Suvitaival T, Rogers S, Kaski S. Stronger findings for metabolomics through Bayesian modeling of multiple peaks and compound correlations. Bioinformatics. 2014; 30: i461–i467. https://doi.org/10.1093/bioinformatics/btu455 PMID: 25161234

5. Zhan X, Patterson AD, Ghosh D. Kernel approaches for differential expression analysis of mass spectrometry-based metabolomics data. BMC Bioinformatics. 2015;16. https://doi.org/10.1186/s12859-014-0426-7 PMID: 25591662

6. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000; 28: 27–30. https://doi.org/10.1093/nar/28.1.27 PMID: 10592173

7. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, et al. HMDB: the Human Metabolome Database. Nucleic Acids Research. 2007; 35: D521–D526. https://doi.org/10.1093/nar/gkl923 PMID: 17202168

8. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011; 39: D691–D697. https://doi.org/10.1093/nar/gkq1018 PMID: 21067998

9. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nature Methods. 2016; 13: 966–967. https://doi.org/10.1038/nmeth.4077 PMID: 27898060

10. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 2018; 46: D661–D667. https://doi.org/10.1093/nar/gkx1064 PMID: 29136241

11. Liu D, Lin X, Ghosh D. Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. Biometrics. 2007; 63: 1079–1088. https://doi.org/10.1111/j.1541-0420.2007.00799.x PMID: 18078480

12. Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics. 2008; 9: 292. https://doi.org/10.1186/1471-2105-9-292 PMID: 18577223

13. Broadaway KA, Cutler DJ, Duncan R, Moore JL, Ware EB, Jhun MA, et al. A Statistical Approach for Testing Cross-Phenotype Effects of Rare Variants. The American Journal of Human Genetics. 2016; 98: 525–540. https://doi.org/10.1016/j.ajhg.2016.01.017 PMID: 26942286

14. Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, et al. Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. The American Journal of Human Genetics. 2015; 96: 797–807. https://doi.org/10.1016/j.ajhg.2015.04.003 PMID: 25957468

15. Jensen AM, Tregellas JR, Sutton B, Xing F, Ghosh D. Kernel machine tests of association between brain networks and phenotypes. PLoS One. 2019;14. https://doi.org/10.1371/journal.pone.0199340 PMID: 30897094

16. Chaleckis R, Meister I, Zhang P, Wheelock CE. Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. Current Opinion in Biotechnology. 2019; 55: 44–50. https://doi.org/10.1016/j.copbio.2018.07.010 PMID: 30138778

17. Amini Arash A., Chen Aiyou, Bickel Peter J., Levina Elizaveta. Pseudo-Likelihood Methods for Community Detection in Large Sparse Networks. Ann Stat. 2013;41. https://doi.org/10.1214/13-Aos1138

18. Le CM, Levina E, Vershynin R. Concentration and regularization of random graphs. Random Structures & Algorithms. 2017; 51: 538–561. https://doi.org/10.1002/rsa.20713

19. Schaid DJ. Genomic Similarity and Kernel Methods II: Methods for Genomic Information. Hum Hered. 2010; 70: 132–140. https://doi.org/10.1159/000312643 PMID: 20606458

20. Freytag S, Manitz J, Schlather M, Kneib T, Amos CI, Risch A, et al. A Network-Based Kernel Machine Test for the Identification of Risk Pathways in Genome-Wide Association Studies. Hum Hered. 2013; 76: 64–75. https://doi.org/10.1159/000357567 PMID: 24434848

21. Manica M, Cadow J, Mathis R, Rodríguez Martínez M. PIMKL: Pathway-Induced Multiple Kernel Learning. npj Systems Biology and Applications. 2019; 5: 1–8. https://doi.org/10.1038/s41540-018-0079-7 PMID: 30564456

22. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, et al. Genetic epidemiology of COPD (COPDGene) study design. COPD. 2010; 7: 32–43. https://doi.org/10.3109/15412550903499522 PMID: 20214461

**23.** Chen J, Chen W, Zhao N, Wu MC, Schaid DJ. Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. Genetic Epidemiology. 2016; 40: 5–19. https://doi.org/10.1002/gepi.21934 PMID: 26643881

**24.** Kolaczyk Eric D. Statistical ANalysis of Network Data. New York: Springer-Verlag New York; 2009. https://doi.org/10.1103/PhysRevE.79.061916 PMID: 19658533

**25.** Smola AJ, Kondor R. Kernels and Regularization on Graphs. In: Schölkopf B, Warmuth MK, editors. Learning Theory and Kernel Machines. Berlin, Heidelberg: Springer Berlin Heidelberg; 2003. pp. 144–158. https://doi.org/10.1007/978-3-540-45167-9_12

**26.** Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert J-P. Classification of microarray data using gene networks. BMC Bioinformatics. 2007; 8: 35. https://doi.org/10.1186/1471-2105-8-35 PMID: 17270037

**27.** Davies RB. The distribution of a linear combination of X2 random variables. J R Stat Soc Series C (Appl Stat). 1980; 29: 323–333.

**28.** Shen Y, Zhu J. Power analysis of principal components regression in genetic association studies*. J Zhejiang Univ Sci B. 2009; 10: 721–730. https://doi.org/10.1631/jzus.B0830866 PMID: 19816996

**29.** Simes R. J. An Improved Bonferroni Procedure for multiple tests of significance. Biometrika. 1986; 73: 751–4.

**30.** Ha SS, Kim I, Wang Y, Xuan J. Applications of Different Weighting Schemes to Improve Pathway-Based Analysis. Comp Funct Genomics. 2011;2011. https://doi.org/10.1155/2011/463645 PMID: 21687588

**31.** Kim I, Pang H, Zhao H. Bayesian semiparametric regression models for evaluating pathway effects on continuous and binary clinical outcomes. Stat Med. 2012; 31: 1633–1651. https://doi.org/10.1002/sim.4493 PMID: 22438129

**32.** Kim I, Pang H, Zhao H. Statistical properties on semiparametric regression for evaluating pathway effects. J Stat Plan Inference. 2013; 143: 745–763. https://doi.org/10.1016/j.jspi.2012.09.009 PMID: 24014933

**33.** Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007; 23: 980–987. https://doi.org/10.1093/bioinformatics/btm051 PMID: 17303618

**34.** Huang Q, Hu D, Wang X, Chen Y, Wu Y, Pan L, et al. The modification of indoor PM2.5 exposure to chronic obstructive pulmonary disease in Chinese elderly people: A meet-in-metabolite analysis. Environment International. 2018; 121: 1243–1252. https://doi.org/10.1016/j.envint.2018.10.046 PMID: 30389378

**35.** Kelly RS, Virkud Y, Giorgio R, Celedón JC, Weiss ST, Lasky-Su J. Metabolomic profiling of lung function in Costa-Rican children with asthma. Biochimica et Biophysica Acta (BBA)—Molecular Basis of Disease. 2017; 1863: 1590–1595. https://doi.org/10.1016/j.bbadis.2017.02.006 PMID: 28188833

**36.** Li X, Cheng J, Shen Y, Chen J, Wang T, Wen F, et al. Metabolomic analysis of lung cancer patients with chronic obstructive pulmonary disease using gas chromatography-mass spectrometry. Journal of Pharmaceutical and Biomedical Analysis. 2020; 190: 113524. https://doi.org/10.1016/j.jpba.2020.113524 PMID: 32795777

**37.** Cruickshank-Quinn CI, Jacobson S, Hughes G, Powell RL, Petrache I, Kechris K, et al. Metabolomics and transcriptomics pathway approach reveals outcome-specific perturbations in COPD. Sci Rep. 2018;8. https://doi.org/10.1038/s41598-017-18329-3 PMID: 29311689

**38.** Chai AB, Ammit AJ, Gelissen IC. Examining the role of ABC lipid transporters in pulmonary lipid homeostasis and inflammation. Respir Res. 2017;18. https://doi.org/10.1186/s12931-017-0503-3 PMID: 28095852

**39.** Ruzsics I, Nagy L, Keki S, Sarosi V, Illes B, Illes Z, et al. L-Arginine Pathway in COPD Patients with Acute Exacerbation: A New Potential Biomarker. COPD: Journal of Chronic Obstructive Pulmonary Disease. 2016; 13: 139–145. https://doi.org/10.3109/15412555.2015.1045973 PMID: 26514682

**40.** Scott JA, Duongh M, Young AW, Subbarao P, Gauvreau GM, Grasemann H. Asymmetric Dimethylarginine in Chronic Obstructive Pulmonary Disease (ADMA in COPD). Int J Mol Sci. 2014; 15: 6062–6071. https://doi.org/10.3390/ijms15046062 PMID: 24727374

**41.** Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9: 559. https://doi.org/10.1186/1471-2105-9-559 PMID: 19114008

**42.** Langfelder P, Cantle JP, Chatzopoulou D, Wang N, Gao F, Al-Ramahi I, et al. Integrated genomics and proteomics define huntingtin CAG length–dependent networks in mice. Nat Neurosci. 2016; 19: 623–633. https://doi.org/10.1038/nn.4256 PMID: 26900923

**43.** Shirasaki DI, Greiner ER, Al-Ramahi I, Gray M, Boontheung P, Geschwind DH, et al. Network Organization of the Huntingtin Proteomic Interactome in Mammalian Brain. Neuron. 2012; 75: 41–57. https://doi.org/10.1016/j.neuron.2012.05.024 PMID: 22794259

44. Zhang G, He P, Tan H, Budhu A, Gaedcke J, Ghadimi BM, et al. Integration of Metabolomics and Transcriptomics Revealed a Fatty Acid Network Exerting Growth Inhibitory Effects in Human Pancreatic Cancer. Clin Cancer Res. 2013; 19: 4983–4993. https://doi.org/10.1158/1078-0432.CCR-13-0209 PMID: 23918603

45. Mamdani M, Williamson V, McMichael GO, Blevins T, Aliev F, Adkins A, et al. Integrating mRNA and miRNA Weighted Gene Co-Expression Networks with eQTLs in the Nucleus Accumbens of Subjects with Alcohol Dependence. PLOS ONE. 2015; 10: e0137671. https://doi.org/10.1371/journal.pone.0137671 PMID: 26381263

46. Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis. 2004; 90: 196–212. https://doi.org/10.1016/j.jmva.2004.02.009

47. Shi WJ, Zhuang Y, Russell PH, Hobbs BD, Parker MM, Castaldi PJ, et al. Unsupervised discovery of phenotype-specific multi-omics networks. Bioinformatics. 2019; 35: 4336–4343. https://doi.org/10.1093/bioinformatics/btz226 PMID: 30957844

48. Chai LE, Loh SK, Low ST, Mohamad MS, Deris S, Zakaria Z. A review on the computational approaches for gene regulatory network construction. Computers in Biology and Medicine. 2014; 48: 55–65. https://doi.org/10.1016/j.compbiomed.2014.02.011 PMID: 24637147

49. Seoane JA, Day INM, Gaunt TR, Campbell C. A pathway-based data integration framework for prediction of disease progression. Bioinformatics. 2014; 30: 838–845. https://doi.org/10.1093/bioinformatics/btt610 PMID: 24162466

50. Larson NB, Chen J, Schaid DJ. A review of kernel methods for genetic association studies. Genetic Epidemiology. 2019; 43: 122–136. https://doi.org/10.1002/gepi.22180 PMID: 30604442

51. Karoui NE. The spectrum of kernel random matrices. Ann Statist. 2010;38. https://doi.org/10.1214/08-AOS648

52. Bernhard Schölkopf, Alexander J. Smola. Learning with Kernels. Massachusetts Institute of Technology; 2002.

53. Cristianini Nello, John Shawe-Taylor. An Introduction to Support Vector Machines. Cambridge University Press; 2000. Available: Http://www.cambridge.org

54. Chung Fan, Graham. Spectral Graph Theory. 1997.

55. Purdom E. Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree. Ann Appl Stat. 2011; 5: 2326–2358. https://doi.org/10.1214/10-AOAS402

56. Kondor RI, Lafferty J. Diffusion Kernels on Graphs and Other Discrete Input Spaces.: 8.

57. R Core Team. R: A language and environment for statistical computing. 2019. Available: https://www.R-project.org/

58. Csardi G, Nepusz T. The igraph software package for complex network research.: 9.

59. Barabási A-L, Albert R. Emergence of Scaling in Random Networks. Science. 1999; 286: 509–512. https://doi.org/10.1126/science.286.5439.509 PMID: 10521342

60. Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. J R Stat Soc Series B Stat Methodol. 2014; 76: 373–397. https://doi.org/10.1111/rssb.12033 PMID: 24817823

61. Shaddox E, Peterson CB, Stingo FC, Hanania NA, Cruickshank-Quinn C, Kechris K, et al. Bayesian inference of networks across multiple sample groups and data types. Biostatistics. 2020; 21: 561–576. https://doi.org/10.1093/biostatistics/kxy078 PMID: 30590505

62. Gillenwater LA, Pratte KA, Hobbs BD, Cho MH, Zhuang Y, Halper-Stromberg E, et al. Plasma Metabolomic Signatures of Chronic Obstructive Pulmonary Disease and the Impact of Genetic Variants on Phenotype-Driven Modules. Network and Systems Medicine. 2020; 3: 159–181. https://doi.org/10.1089/nsm.2020.0009 PMID: 33987620

63. Gillenwater LA, Kechris KJ, Pratte KA, Reisdorph N, Petrache I, Labaki WW, et al. Metabolomic Profiling Reveals Sex Specific Associations with Chronic Obstructive Pulmonary Disease and Emphysema. Metabolites. 2021;11. https://doi.org/10.3390/metabo11030161 PMID: 33799786

64. Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, et al. Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. Anal Chem. 2006; 78: 567–574. https://doi.org/10.1021/ac051495j PMID: 16408941

65. Hastie Trevor, Tibshirani Robert, Narasimhan Balasubramanian, Chu Gilbert. impute: Imputation for microarray data. Available: https://www.bioconductor.org/packages/release/bioc/html/impute.html

66. Tenenbaum D. KEGGREST: Client-side REST access to KEGG. Available: https://bioconductor.riken.jp/packages/3.0/bioc/html/KEGGREST.html

67. Li K, Gao Y, Pan Z, Jia X, Yan Y, Min X, et al. Influence of Emphysema and Air Trapping Heterogeneity on Pulmonary Function in Patients with COPD. Int J Chron Obstruct Pulmon Dis. 2019; 14: 2863–2872. https://doi.org/10.2147/COPD.S221684 PMID: 31839706