Distance metrics for genome-scale metabolic models

Andrea Cabbia (a.cabbia@tue.nl)
Technische Universiteit Eindhoven https://orcid.org/0000-0003-4275-1454
Peter A.J. Hilbers
Technische Universiteit Eindhoven
Natal A.W. van Riel
Technische Universiteit Eindhoven

Research article

Keywords: Genome-scale metabolic models; Machine Learning; Distance metrics

Posted Date: September 5th, 2019

DOI: https://doi.org/10.21203/rs.2.10792/v1

RESEARCH

Distance metrics for genome-scale metabolic models

Andrea Cabbia[1*], Peter A.J. Hilbers[1] and Natal A.W. van Riel[1,2]

Abstract

Background: Due to algorithmic advancements and to the availability of experimental datasets, large collections of genome-scale metabolic models (GSMM) can nowadays be generated automatically. Nevertheless, few tools are available to efficiently analyze such large sets of models, for example to study the link between genetic and metabolic heterogeneity. Machine Learning (ML) algorithms use the distance between data points to find patterns in large datasets. A method to determine distance between genome-scale metabolic models was thus necessary to apply ML to large model sets. We address this issue considering three levels of model representation, and defining for each of them a different distance metric: Jaccard metric for metabolic reconstructions, graph kernels for network graph topology and cosine similarity between flux distributions for constraint-based models. We employed two benchmark datasets, each containing hundreds of metabolic models, to compare the different metrics: the first is composed of 100 human genome-scale models developed from proteomics data of four different cancer tissues, while the second contains more than 800 models of bacterial species inhabiting the human gut and was developed from metagenomic data.

Results: Metrics based on the overlap of reactions content (Jaccard) and on network similarity (graph kernels) achieve remarkably similar performances in clustering and classification tasks. Phylogenetic trees built on these two metrics have the same distance from a reference taxonomy, even if the trees themselves are different from each other. Mantel test shows high correlation between distance matrices built with Jaccard and network similarity metrics.

Conclusions: We expand the concept of distance between metabolic models, highlighting new properties of the Jaccard metric such as its correlation with network similarity and function. We show how distance metrics enable the application of machine learning algorithms to genome-scale metabolic models, enabling efficient pattern recognition in large model sets.

Keywords: Genome-scale metabolic models; Machine Learning; Distance metrics

Introduction

Phenotypic variation in responses to the same stimuli, such as pharmaceutical treatments and diet, arises from genetic and epigenetic variation, different microbiome composition and function, and from lifestyle factors such as physical activity [1]. Genome-scale metabolic models represent the metabolic network of an organism or tissue, and simulate the response to particular environmental conditions such as nutrient availability. These models can be parametrized with multiple types of molecular data, such as transcriptomics and proteomic, to produce context-specific models. This feature makes them attractive platforms for the integration of multi-omics datasets, and for the study of phenotypic heterogeneity, for example to study the complex genetics of metabolic diseases and conditions such as development of frailty with age. [2]. Genome-scale metabolic models are knowledge-driven: knowledge of physico-chemical laws is employed to build mechanistic models that are able to explain or approximate experimental data. Automated model generation algorithms [3], [4], [5] have made possible the construction of large libraries of genome-scale metabolic models, such as patient-derived models, [6], and human gut microbial communities [7], [8]. Despite this increase in the number of GSMM developed and published, very few methods are available to describe and study heterogeneity across the large number of different models included in such libraries [9], [10]. Machine Learning (ML) algorithms can automatically

*Correspondence: a.cabbia@tue.nl
[1]Computational Biology, Eindhoven University of Technology, Groene Loper, Eindhoven, NL
Full list of author information is available at the end of the article

find patterns in large amounts of data, without any previous knowledge about the system. These methods scale well to large datasets, but it is difficult to introduce any a priori knowledge about biological systems in this kind of algorithms [11]. ML algorithms can be broadly divided in supervised and unsupervised: unsupervised algorithms, often referred to as clustering algorithms, automatically detect underlying patterns and similarities in large amounts of unlabelled data, while supervised methods, also known as classification algorithms, can learn known patterns from a set of annotated training data, and use this information to create a predictive model able to match new unseen data to a known profile. Many classification algorithms, such as K-Nearest Neighbor (kNN, Support Vector Machine (SVM), and clustering methods such as Hierarchical clustering (HC) and K-means, rely on measures of distance to discover similarities between different data points. [12],[13],[14]. We see large potential in the combination of genome-scale metabolic models with ML to address the question of biological heterogeneity. The definition of methods to measure distance between metabolic models is needed as a fundamental step towards the integration of these two computational approaches, enabling the application of machine learning algorithms to find patterns in large sets of metabolic models. The Jaccard similarity metric, which is defined as the size of the intersection divided by the size of the union of two sets, has been used as measure metabolic similarity between GSMMs in several papers [15], [16], [7]. This similarity metric can be applied to metabolic reconstructions to compute a similarity score between the sets of reactions or metabolites of two different models, but it does not take into account higher level features of the models such as the topology of the metabolic network or the constraints of a constraint-based model. For this reason, we hypothesize that metrics based on different properties of the model could identify different patterns. To test such hypothesis, we identified three different representations of a genome-scale metabolic model, respectively as list of reactions, as graph topology and as flux constraints, and defined distance metrics based on each of these features. The distance metrics were then compared in a series of machine learning and phylogenetic analysis applications.

Methods

Benchmark Datasets

To compare and evaluate different genome-scale model distance metrics, we employed two published collections of genome-scale models as benchmark datasets. The first is a library of patient-derived genome scale models, developed from cancer patient proteomics data [6], which is available at the Biomodels database [17]: https://www.ebi.ac.uk/biomodels-main/pdgsmm. For the remainder of the paper, we will identify this dataset with the abbreviation PDGSMM. All the models included in this dataset have been generated from the same reference human metabolic model, HMR2 [18], using the same algorithm, tINIT [5], in order to minimize the technical variability (also known as batch effect) present inside each dataset [19]. To facilitate our analysis, we chose to use only a subset of this large collection of thousands of models. We selected 100 models, representing the metabolism of 4 different cancer tissues: liver, lung, pancreas and skin. Figure 1A and 1C show the content of the models in this dataset, respectively in terms of number of reaction and metaboites, for each of the four tissues. The second dataset is a collection of microbial models inhabiting the human gut, AGORA, [7], which can be accessed at the Virtual Metabolic Human (VMH) database https://www.vmh.life/#microbes/search. The latest version of the collection (1.03) includes 818 bacterial models. Figure 1B and 1D show the content of the models in this dataset.

Figure 1 Summary of the reactions and metabolite content of the models included in each benchmark dataset. (A) Reaction content of the models of the PDGSMM dataset, grouped by tissue type. (B) Reaction content of of the models in the AGORA dataset, grouped by Phylum, (C) Metabolite content of the models of the PDGSMM dataset, grouped by tissue type. (D) Metabolite content of of the models in the AGORA dataset, grouped by Phylum

Alternative representations of a genome-scale metabolic model

GSMMs are data structures that can be represented in alternative ways, depending on the type of application: as repositories of genes, metabolites and biochemical reactions experimentally found or predicted to be present in a certain tissues or organism (metabolic reconstruction); as bipartite graph structure, where reactions and metabolites are two separate sets of nodes, connected by edges, resulting in a specific topology, shown in Additional Figure 4 (network topology graph), and as constraint-based model, describing how the metabolism of a tissue or organism adapts to a specific environment or condition, i.e. subject to constraints on the metabolic fluxes through exchange and internal reactions. Additional Figure 5 shows the structure and attributes of a constraint-based metabolic model (SBML file). Such alternative representations of the same object are often mistakenly conflated with each other, leaving space for misunderstanding and interoperability issues [20]. Making this distinction ex-

plicit may improve our definition of metabolic distance, since each of these three different aspects of a GSMM (metabolic reconstruction, network topology graph and constraint-based model) encapsulates different aspects of metabolic heterogeneity. We will therefore define separate distance metrics for each of these features.

Distance and similarity metrics

Jaccard Metric

The Jaccard metric has been used extensively to measure distance between genome-scale metabolic models in several studies: [15],[16],[4]. In [7], the authors used metagenomic data to build a collection of 773 genome-scale metabolic reconstructions of bacterial species inhabiting the human gut. The variability of these reconstructions has been quantified computing the Jaccard distance between lists of reactions for each pair of reconstructions:

$$JD = 1 - \frac{|R_i \cap R_j|}{|Ri \cup Rj|}$$

where R_i is the list of reactions from reconstruction **i** and R_j is the list of reactions present in reconstruction **j**. JD = 1 implies that the two reconstructions share no reactions, while JD = 0 means that the two reconstructions have identical reaction lists.

Graph Kernels

Kernel methods are a class of machine learning algorithms [21], owing their name to the use of kernel functions, i.e. nonlinear functions that map the data into a different dimensional space, such as the Reproducing Kernel Hilbert Space (RKHS). In the RKHS, the distance of samples represents their similarity. As long as we can formulate everything in terms of kernel evaluations, we never explicitly need to compute the exact coordinates in the RKHS, but rather their distance is found by computing their inner products, an operation that is often computationally cheaper than the explicit computation of the new coordinates.

$$K(X, Y) =< X \cdot Y >$$

Kernel methods have been applied to differnt types of data: biological sequences [22], text [23], images [24], as well as graphs [25]. Graph kernels are kernel functions that compute an inner product on graphs, as a computationally efficient way to measure their similarity, retaining information about their network topology. They were developed as a way to apply machine learning algorithms to structured data, such as knowledge graphs, which are ways to describe relationships between entities (or 'ontologies') in a graph structure. This format can be read and processed by algorithms, which can then infer new non-trivial properties of the entities traversing the graph. Genome-scale metabolic models are examples of such structured objects, containing information about certain entities (i.e. metabolites, reactions, genes) and their relationships. A simple example of these functions is the random walk kernel [26], which computes random walks on two graphs simultaneously, and then quantifies their similarity as the number of common walks in the two graphs. This is equivalent to doing random walks on the direct product of the pair of graphs. Random walk kernels are efficient for small, simple graphs, but as the size and complexity of the networks increases, such as with genome-scale metabolic networks, more modern techniques are needed. The Weisfeiler-Lehman Subtree (WLS) kernel, proposed in [27], is a kernel suited for computation of similarity between large, complex graphs. This kernel computes the number of subtrees shared between two graphs, using the Weisfeiler-Lehman [28] algorithm to find an approximate solution to the problem of graph isomorphism, which is np-complete. The concept of this algorithm is to relabel the nodes with the sorted set of node labels of neighbouring nodes. This procedure is repeated for n iterations, or until the set of labels of two nodes are different. The node labels at this point will contain topological information about the neighborhood of each node. Finally, the kernel value is computed by counting common labels between two graphs. Since this procedure is performed simultaneously on all input graphs, its runtime scales only linearly with the number of edges of the graphs and the number of iterations.

Cosine Similarity

Metabolic models are particular instances of a metabolic reconstruction, subject to constraints on the value of certain fluxes. These constraints, usually derived from experimental data, restrict the space of possible flux distributions in the network. Flux Balance Analysis (FBA) [29] can be employed to find one of the possible flux distributions that satisfies a given celular objective, usually biomass or ATP production. To quantify the similarity of metabolic models, we employed FBA to find the optimal flux distribution for each of the models in the test dataset, using the biomass reaction as objective function. Cancer is in fact known to rewire cellular metabolism to maintain elevated rates of cellular growth and division. [30]. Therefore we deemed acceptable the utilization of the biomass reaction as cellular objective also in human models. We computed

the similarity between pairs of models as the cosine of the angle spanned by the two flux vectors resulting from FBA. Since this metric does not take into account the length of the two vectors, but just their angle, this provides a normalized measure of the orientation of the flux vectors. It is defined as the dot product of two numeric vectors, divided by the product of the vector lengths:

$$cos(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||}$$

Output values close to 1 indicate high similarity (i.e. the vectors point to the same direction).

Supervised Classification

The three metrics described in the previous section have been evaluated and compared through a serie of machine learning and phylogenetic tests. For the classification test, we used K-Nearest Neighbor (K-NN) and Kernel-Support Vector Machine (Kernel-SVM) algorithms, two robust and well-studied algorithms, which rely on distance metrics. K-NN is one of the simplest classification algorithms: each test sample is classified by a majority vote of its K nearest neighbors, where K is a positive integer, typically small. The way in which distance is computed is fundamental in this algorithm, to correctly identify the neighboring training samples and achieve high accuracy. Linear-SVM models are representations of the training points in a n-dimensional space, mapped in such way that data points belonging to separate classes are divided by a linear decision boundary that is as wide as possible. Kernel-SVM is an extension of this method: the training data points are transformed to an higher-dimensional space (RKHS), allowing the classification of samples even when a linear boundary that separates the classes cannot be found in the original space. Pairwise distance matrices were used as inputs for both the classifier algorithms. The reported results are the average test accuracies after 10-fold cross validation.

Unsupervised Clustering

We evaluated the performance of the metrics in an unsupervised clustering task, measuring the agreement between the predicted label and the the ground-truth label. We used two clustering algorithms. The first is Hierarchical clustering, where each observation starts as its own cluster, and pairs of clusters are recursively merged based on their proximity.The second is Spectral clustering, which first computes the eigenvalues of the similarity matrix to perform dimensionality reduction, then clusters the samples in a lower dimensional space. Pairwise distance matrices were used as inputs for the clustering algorithms. In both cases the number of clusters was chosen to match the number of classes present in the ground truth label.

Comparison of phylogenetic trees

We further evaluated the metrics in a different application setting, creating and comparing three different phylogenetic trees for the microbial models in the AGORA dataset, one for each metric. The trees were first compared to a reference tree, built from NCBI taxonomy database, then compared to each other. We computed a similarity score between each pair of trees using Robinson-Foulds (RF) metric [31]: given two unrooted trees and a set of labels (e.g., taxa) for each node, the Robinson–Foulds metric finds the number of direct and inverse operations to convert one into the other. The number of operations defines their distance.

Correlation between distance matrices

Mantel test [32] was used to compute Pearson's correlation coefficient between two pairwise distance matrices. Statistical significance was assessed via permutation test.

Results

Three distance metrics were evaluated in different tasks: We used the SciKit-Learn [33] implementation of K-NN and Kernel-SVM classifiers and of Hierarchical and Spectral clustering. We employed the GraKeL [34] implementation of Graph Kernel algorithms and the ete3 package [35] to build and compare phylogenetic trees. All the computations were performed on a Ubuntu 18.04 machine with Xeon CPU E5-1620 @ 3.50GHz and 16GB RAM.

Classification

For the PDGSM dataset, the task was to predict the correct tissue label (skin, lung, liver, pancreas), while for the AGORA dataset, the task was to predict four different labels: Gram staining (e.g. positive or negative), the type of oxygen metabolism (e.g. aerobe, anaerobe), its taxonomy (Phylum) and type of interaction with the host (e.g. commensal, pathogen, probiotic). This information has also been retrieved from the VMH database. Figure 2 shows the average prediction accuracy for each supervised learning algorithm, when tasked to predict the label of the samples in the test set. The results shown are the averages of a 10-fold cross validation.

Figure 2 Classification results. (A) K-Nearest neighbor classification results. (B) Kernel-SVM classification results

From the results of the classification test we can immediately notice how Jaccard and Graph Kernel metrics have very similar performances in almost every task. The classification accuracy in case of Cosine similarity between flux vectors is somewhat lower but comparable to the other two metrics in the case of the microbial AGORA dataset, though significantly worse in case of the larger human models included in the PDGSM dataset. This observation suggests that the efficacy of this metric could decrease with the increase in dimensionality of the solution space of the models. A classification accuracy score > 0.9 for the taxonomy label (AGORA-Phylum) across both clustering algorithms indicates a possible correlation between model-based distance and evolutionary distance.

Clustering

In the second test we compared how different clustering algorithms perform, when given different distance matrices built from different metrics as inputs. Since the ground-truth was known, we could measure how accurate the label predicted by the clustering algorithms matched the ground truth label. Additional figures 1-3 show the result of a PCA on the AGORA dataset.

Figure 3 Clustering results. (A) Hierarchical Clustering results. (B) Spectral Clustering results

In the hierarchical clustering results (Figure 3A), flux vector similarity performs better than network similarity in many cases, and is able to recover correctly 64% of the labels in the AGORA-Gram case. Nevertheless, also in the this example, the performance of this metric degrades for the human models dataset. In the spectral clustering results (Figure 3B), the performances are higher for reactions and network similarity metrics, except once again for the AGORA-Gram case. The highest clustering accuracy is achieved with the Jaccard metric, reaching 61% of correct predictions for the AGORA-Taxonomy case.

Comparison of phylogenetic trees

We built three different phylogenetic trees for the AGORA dataset, one for each of the metrics. The trees were compared first with a reference, a tree built from the NCBI taxonomy database, and then with each other. The degree of similarity between phylogenetic trees has been expressed with Robinson-Foulds metric (RF) and in % of shared branches. Table 1 summarizes the results of the comparison. The results of the comparison between the three phylogenetic trees with the reference taxonomy tree, presented in Table 1, show how the trees built on reaction similarity (Jaccard) and network similarity (Graph Kernel) have the same distance from the reference (RF = 0.82-0.83). Their direct comparison nevertheless proves their differences: their normalized RF distance is 0.34, corresponding to an overlap of 83% in the branches of the two trees. By comparison, the tree built from flux vectors similarity (Cosine), is further away from the reference (RF = 0.92) and less similar to the other two trees (RF = 0.78-0.79), and shows less overlap with their branches (60%).

Table 1 Phylogenetic Trees comparison

Source Tree	Reference Tree	RF	%src-ref	%ref-src
Jaccard	NCBI Taxonomy	0.82	0.56	0.82
Graph Kernel	NCBI Taxonomy	0.83	0.56	0.83
Cosine similarity	NCBI Taxonomy	0.92	0.52	0.78
Jaccard	Graph Kernel	0.34	0.83	0.83
Cosine similarity	Graph Kernel	0.78	0.60	0.60
Jaccard	Cosine similarity	0.79	0.60	0.60

nRF: Normalized Robinson-Foulds Distance, **%src-ref** frequency of edges in source tree found in the reference (1.00 = 100% of branches are found), **%ref-src** frequency of edges in the reference tree found in target (1.00 = 100% of branches are found)

Correlation between distance matrices

We performed a Mantel test to check for correlations between pairs of distance matrices built from different distance metrics. The results of the Mantel test are presented in Table 2 (AGORA dataset) and Table 3 (PDGSM dataset). In both cases, distance matrices built from Jaccard and Graph Kernel metrics show very high correlation (> 0.9). This result shows how reaction similarity is correlated with the similarity in the topology of metabolic networks.

Table 2 Distance matrices correlation: AGORA dataset

Distance Matrix 1	Distance Matrix 2	Pearson's r	p-value
Jaccard Distance	Graph Kernel	0.985	0.001
Jaccard Distance	Cosine Similarity	0.371	0.001
Graph Kernel	Cosine Similarity	0.399	0.001

Table 3 Distance matrices correlation: PDGSM dataset

Distance Matrix 1	Distance Matrix 2	Pearson's r	p-value
Jaccard Distance	Graph Kernel	0.904	0.001
Jaccard Distance	Cosine Similarity	0.380	0.001
Graph Kernel	Cosine Similarity	0.386	0.001

Discussion

We employed distance metrics and ML for pattern recognition in large sets of GSMM, as a strategy for the integration of multiple omics datasets with *a priori* knowledge of the structure of metabolic networks [36]. Multiple ML algorithms, including K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), Hierarchical Clustering (HC) and Spectral Clustering (SC),

rely in fact on some measure of distance between data points, usually quantified with Euclidean metric, to identify patterns between similar samples or to match previously unseen samples to a known profile. The problem was to define how distance between metabolic models should be expressed: we addressed this issue considering three possible representations of a genome-scale metabolic model and defining for each of them a different metric: Jaccard metric for metabolic reconstructions, graph kernels for topology of metabolic network graph and cosine similarity between flux distributions for the constraint-based model. Reaction similarity can be measured easily with Jaccard or Hamming metrics [37], but measuring topology similarity between genome-scale metabolic networks graphs, containing up to thousands of nodes and edges, was an intractable problem until recently [38]. Algorithms such as the Weisfeiler-Lehman Subtree Kernel (WLS) can be applied to efficiently compute a similarity score in large sets of genome-scale metabolic networks. Cosine similarity between pairs of flux vectors, obtained from FBA while maximizing biomass production, is admittedly suboptimal for several reasons: firstly, FBA results are scarcely reproducible, mainly due to the degeneracy of stoichiometric networks (i.e. many flux distributions satisfy the same objective function) and to its sensitivity to the particular software or algorithm used to solve the linear programming (LP) problem. Additionally, FBA imposes the use of an arbitrary optimization objective, meaning that each flux distribution found with FBA is inherently biased, and not representative of the whole solution space. Moreover, a single FBA solution cannot represent the complexity of the entire solution space of a constraint-based model. Few alternative approaches are possible, such as Geometric FBA [39], random sampling of the solution space [40], Expectation Propagation [41] or by comparing sets of Elementary Flux Modes (EFM) [42]. However, their application to large datasets containing hundreds of genome-scale models, which is the focus of this article, so far remains hampered by their excessive computational requirements. Our results revealed that reaction and network similarity, measured respectively as Jaccard distance and Graph Kernel values, have comparable performances across the clustering and classification tests. The comparison between phylogenetic trees of the AGORA dataset revealed that trees built with Jaccard distance and Graph Kernel metric have the same degree of similarity with the reference tree (their normalized RF distance from reference is 0.82-0.83). Neverheless, the two metrics do not fully overlap, since the normalized RF distance between Jaccard and Graph Kernel trees is 0.34, with 84% of the branches in common. The Mantel test reported a very high correlation coefficient between distance matrices built with these two metrics (0.985 for the AGORA dataset, 0.904 for the PDGSM dataset, p-value = 0.001). This last result suggests that an underliyng common structure may exist between the composition of the metabolic network, in terms of lists of reactions and metabolites and its topology, which is responsible for its functionality. Taken together, these results highlight new properties of the Jaccard metric, which, being relatively simple and fast to compute, can also give to the researcher a remarkable amount of information also on the degree of similarity of the functionality of different metabolic networks. Despite its drawbacks, the third metric, cosine similarity between FBA flux vectors performances were remarkably close to the other two metrics, especially in the case of hierarchical clustering tests, where it sometimes outperformed the network similarity metric. An interesting observation is the degradation of its performances in the PDGSM dataset, which contains larger human models. This observation suggests that the performance of this metric may decrease with the increase in the dimensionality of the solution space of the models. Nevertheless, since model's features such as constraints are more closely related to the actual phenotype, we hypothesize that novel similarity metrics based on these properties could have a higher correlation with phenotypic heterogeneity compared to metrics based only on reaction similarity such as Jaccard distance. Follow-up studies should investigate the extent to which model-inferred heterogeneity is correlated to phenotypic heterogeneity in different contexts and organisms.

Conclusion

The concept of distance between metabolic models was expanded by developing distance metrics for three levels of model representation. We highlighted new properties of the Jaccard metric such as its correlation with network similarity and function, and showed how ML can be applied to large sets of genome-scale metabolic models, enabling efficient pattern recognition in large sets of models.

List of abbreviations

GSMM - Genome-scale metabolic model
PDGSM - Patient-derived genome-scale model
ML - Machine Learning
RKHS - Reproducing Kernel Hilbert Space
KNN - K-Nearest Neighbors
K-SVM - Kernelized Support Vector Machine
HC - Hierarchical Clustering
SC - Spectral Clustering
WLS - Weisfeiler-Lehman Subtree

FBA - Flux Balance Analysis
LP - Linear Programming
EFM - Elementary Flux Mode

Declarations:

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The PDGSMM dataset is available at: https://www.ebi.ac.uk/biomodels-main/pdgsmm

The AGORA dataset is available at: https://www.vmh.life/#microbes/search

The code used to generate the results and figures is available at: https://github.com/ACabbia/GSMM-distance

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme, under the Marie Sklodowska-Curie grant agreement 675003.

Authors' contributions

AC and NvR designed the study. PH and NvR supervised the study. AC drafted the manuscript. All authors contributed to and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Not applicable

Author details
[1]Computational Biology, Eindhoven University of Technology, Groene Loper, Eindhoven, NL. [2]Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, NL.

References

1. Zeisel, S.H.: A Conceptual Framework for Studying and Investing in Precision Nutrition. Frontiers in Genetics **10**(March), 1–11 (2019). doi:10.3389/fgene.2019.00200
2. Whittaker, A.C., Delledonne, M., Finni, T., Garagnani, P., Greig, C., Kallen, V., Kokko, K., Lord, J., Maier, A.B., Meskers, C.G.M., Santos, N.C., Sipila, S., Thompson, J.L., van Riel, N.: Physical Activity and Nutrition INfluences In ageing (PANINI): consortium mission statement. Aging Clinical and Experimental Research **30**(6), 685–692 (2018). doi:10.1007/s40520-017-0823-7
3. Schultz, A., Qutub, A.A.: Reconstruction of Tissue-Specific Metabolic Networks Using CORDA. PLoS Computational Biology **12**(3), 1–33 (2016). doi:10.1371/journal.pcbi.1004808
4. Machado, D., Andrejev, S., Tramontano, M., Patil, K.R.: Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Research **46**(15), 7542–7553 (2018). doi:10.1093/nar/gky537
5. Agren, R., Mardinoglu, A., Asplund, A., Kampf, C., Uhlen, M., Nielsen, J.: Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. Molecular Systems Biology **10**(3), 1–13 (2014). doi:10.1002/msb.145122
6. Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., Sanli, K., Von Feilitzen, K., Oksvold, P., Lundberg, E., Hober, S., Nilsson, P., Mattsson, J., Schwenk, J.M., Brunnström, H., Glimelius, B., Sjöblom, T., Edqvist, P.H., Djureinovic, D., Micke, P., Lindskog, C., Mardinoglu, A., Ponten, F.: A pathology atlas of the human cancer transcriptome. Science **357**(6352) (2017). doi:10.1126/science.aan2507
7. Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D.A., Bauer, E., Noronha, A., Greenhalgh, K., Jäger, C., Baginska, J., Wilmes, P., Fleming, R.M.T., Thiele, I.: Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nature Biotechnology **35**(1), 81–89 (2017). doi:10.1038/nbt.3703
8. Garza, D.R., Van Verk, M.C., Huynen, M.A., Dutilh, B.E.: Towards predicting the environmental metabolome from metagenomics with a mechanistic model. Nature Microbiology **3**(4), 456–460 (2018). doi:10.1038/s41564-018-0124-8
9. Biggs, M.B., Papin, J.A.: Managing uncertainty in metabolic network structure and improving predictions using EnsembleFBA. PLoS Computational Biology **13**(3), 1–25 (2017). doi:10.1371/journal.pcbi.1005413
10. Damiani, C., Di Filippo, M., Pescini, D., Maspero, D., Colombo, R., Mauri, G.: PopFBA: Tackling intratumour heterogeneity with Flux Balance Analysis. Bioinformatics **33**(14), 311–318 (2017). doi:10.1093/bioinformatics/btx251
11. Imangaliyev, S., Prodan, A., Nieuwdorp, M., Groen, A.K., van Riel, N.A.W., Levin, E.: Domain intelligible models. Methods **149**, 69–73 (2018). doi:10.1016/j.ymeth.2018.06.011
12. Singh, A., Yadav, A., Rana, A.: K-means with Three different Distance Metrics. International Journal of Computer Applications **67**(10), 13–17 (2013). doi:10.5120/11430-6785
13. Gonzalez-Abril, L., Velasco, F., Ortega, J.A., Franco, L.: Support vector machines for classification of input vectors with different metrics. Computers and Mathematics with Applications **61**(9), 2874–2878 (2011). doi:10.1016/j.camwa.2011.03.071
14. Hu, L.Y., Huang, M.W., Ke, S.W., Tsai, C.F.: The distance function effect on k-nearest neighbor classification for medical datasets. SpringerPlus **5**(1) (2016). doi:10.1186/s40064-016-2941-7
15. Estévez, S.R., Nikoloski, Z.: Context-specific metabolic model extraction based on regularized least squares optimization. PLoS ONE **10**(7), 1–21 (2015). doi:10.1371/journal.pone.0131875
16. Duan, G., Christian, N., Schwachtje, J., Walther, D., Ebenhöh, O.: The metabolic interplay between plants and phytopathogens. Metabolites **3**(1), 1–23 (2013). doi:10.3390/metabo3010001
17. Le Novere, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J.L., Hucka, M.: BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Research **34**(90001), 689–691 (2006). doi:10.1093/nar/gkj092
18. Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., Uhlen, M., Nielsen, J.: Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nature Communications **5**(May 2013), 1–11 (2014). doi:10.1038/ncomms4083
19. Opdam, S., Richelle, A., Kellman, B., Li, S., Zielinski, D.C., Lewis, N.E.: A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. Cell Systems **4**(3), 318–3296 (2017). doi:10.1016/j.cels.2017.01.010
20. Babaei, P., Shoaie, S., Ji, B., Nielsen, J.: Challenges in modeling the human gut microbiome. Nature Biotechnology **36**(8), 682–686 (2018).

doi:10.1038/nbt.4213
21. Schölkopf, B., Smola, A.: Support Vector Machines and Kernel Algorithms. MIT Press, ??? (2002). doi:10.1080/01926180701643313. http://alex.smola.org/papers/2002/SchSmo02b.pdf
22. Tsuda, K., Kin, T., Asai, K.: Marginalized kernels for biological sequences. Bioinformatics **18**(SUPPL. 1) (2002). doi:10.1093/bioinformatics/18.suppl_1.S268
23. Kudo, T., Matsumoto, Y.: Fast methods for kernel-based text analysis. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03 **1**, 24–31 (2003). doi:10.3115/1075096.1075100
24. Yang, M., Ahuja, N., Kriegman, D.: Face recognition using kernel eigenfaces. Image processing, 2000. ... **X**, 37–40 (2000). doi:10.1109/ICIP.2000.900886
25. Borgwardt, K.M., Ong, C.S., Schönauer, S., Vishwanathan, S.V.N., Smola, A.J., Kriegel, H.P.: Protein function prediction via graph kernels. Bioinformatics **21**(SUPPL. 1), 47–56 (2005). doi:10.1093/bioinformatics/bti1007
26. Gärtner, T., Flach, P., Wrobel, S.: On Graph Kernels: Hardness Results and Efficient Alternatives, pp. 129–143 (2003). doi:10.1007/978-3-540-45167-9_11. http://link.springer.com/10.1007/978-3-540-45167-9{%}5C{_}11
27. Shervashidze, N., Schweitzer, P., van Leeuwen, J.E., Mehlhorn, K., Borgwardt, K.M.: Weisfeiler-Lehman Graph Kernels. Journal of Machine Learning Research **12**, 2539–2561 (2011). doi:10.1.1.232.1510
28. Weisfeiler, B.Y., Lehman, A.A.: Reduction of a graph to a canonical form and an algebra which appears in the process. Nauchno-Technicheskaya Informatsiya, Ser. 2 **9**, 12 (1968)
29. Orth, J.D., Thiele, I., Palsson, B.Ø.: What is flux balance analysis? Nat. Biotechnol. **28**(3), 245–248 (2011). doi:10.1038/nbt.1614.What
30. Heiden, M.G.V., Lunt, S.Y., Dayton, T.L., Fiske, B.P., Israelsen, W.J., Mattaini, K.R., Vokes, N.I., Stephanopoulos, G., Cantley, L.C., Metallo, C.M., Locasale, J.W.: Metabolic pathway alterations that support: Cell Proliferation. Cold Spring Harbor Symposia on Quantitative Biology **76**, 325–334 (2011). doi:10.1101/sqb.2012.76.010900
31. Robinson, D.F., Foulds, L.R.: Comparison of Phylogenetic Trees. Mathematical Biosciences **141**, 131–141 (1981). doi:10.1109/ICT-ISPC.2017.8075303
32. Mantel, N.: The Detection of Disease Clustering and a Generalized Regression Approach. Cancer Research **27**(2 Part 1), 209–220 (1967)
33. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**(X), 2825–2830 (2011). doi:10.1007/s13398-014-0173-7.2. 1201.0490
34. Siglidis, G., Nikolentzos, G., Limnios, S., Giatsidis, C., Skianis, K., Vazirgianis, M.: GraKeL: A Graph Kernel Library in Python. ArXiv **X**, 1–5 (2018). 1806.02193
35. Huerta-Cepas, J., Serra, F., Bork, P.: ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Molecular Biology and Evolution **33**(6), 1635–1638 (2016). doi:10.1093/molbev/msw046
36. Nielsen, J.: Systems Biology of Metabolism: A Driver for Developing Personalized and Precision Medicine. Cell Metabolism **25**(3), 572–579 (2017). doi:10.1016/j.cmet.2017.02.002
37. Ghaffari, P., Mardinoglu, A., Asplund, A., Shoaie, S., Kampf, C., Uhlen, M., Nielsen, J.: Identifying anti-growth factors for human cancer cell lines through genome-scale metabolic modeling. Scientific Reports **5**, 1–10 (2015). doi:10.1038/srep08183
38. Ghosh, S., Das, N., Gonçalves, T., Quaresma, P., Kundu, M.: The journey of graph kernels through two decades. Elsevier Inc. (2018). doi:10.1016/j.cosrev.2017.11.002. https://doi.org/10.1016/j.cosrev.2017.11.002
39. Smallbone, K., Simeonidis, E.: Flux balance analysis: A geometric perspective. Journal of Theoretical Biology **258**(2), 311–315 (2009). doi:10.1016/j.jtbi.2009.01.027
40. Bordel, S., Agren, R., Nielsen, J.: Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. PLoS Computational Biology **6**(7), 16 (2010). doi:10.1371/journal.pcbi.1000859
41. Braunstein, A., Muntoni, A.P., Pagnani, A.: An analytic approximation of the feasible space of metabolic networks. Nature Communications **8**(May 2016), 1–9 (2017). doi:10.1038/ncomms14915
42. Schuster, S., Dandekar, T., Fell, D.A.: Detection of elementary flux modes in biochemical networks: A promising tool for pathway analysis and metabolic engineering. Trends in Biotechnology **17**(2), 53–60 (1999). doi:10.1016/S0167-7799(98)01290-6

Figures

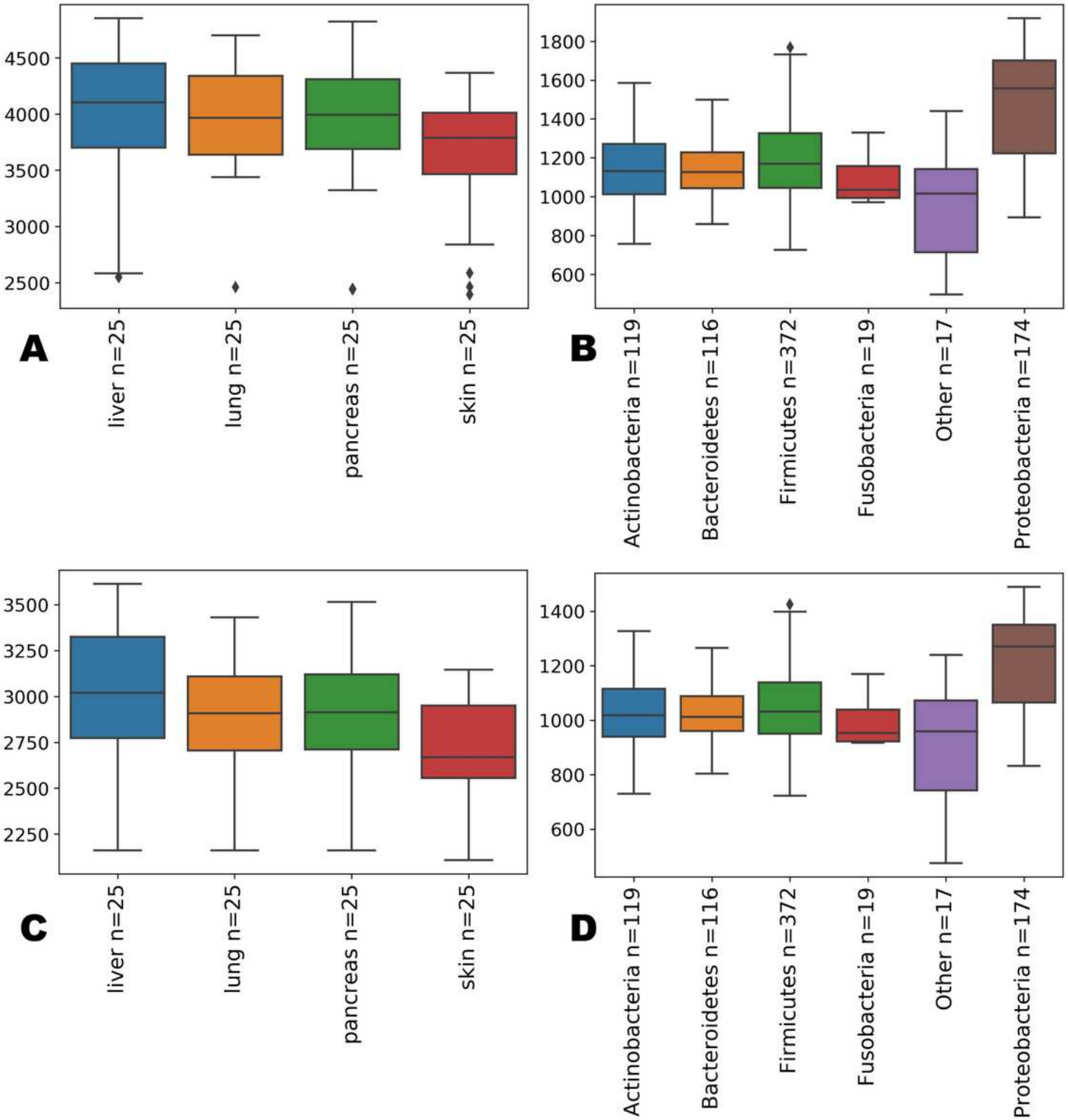

Figure 1

Summary of the reactions and metabolite content of the models included in each benchmark dataset. (A) Reaction content of the models of the PDGSMM dataset, grouped by tissue type. (B) Reaction content of of the models in the AGORA dataset, grouped by Phylum, (C) Metabolite content of the models of the

PDGSMM dataset, grouped by tissue type. (D) Metabolite content of of the models in the AGORA dataset, grouped by Phylum.

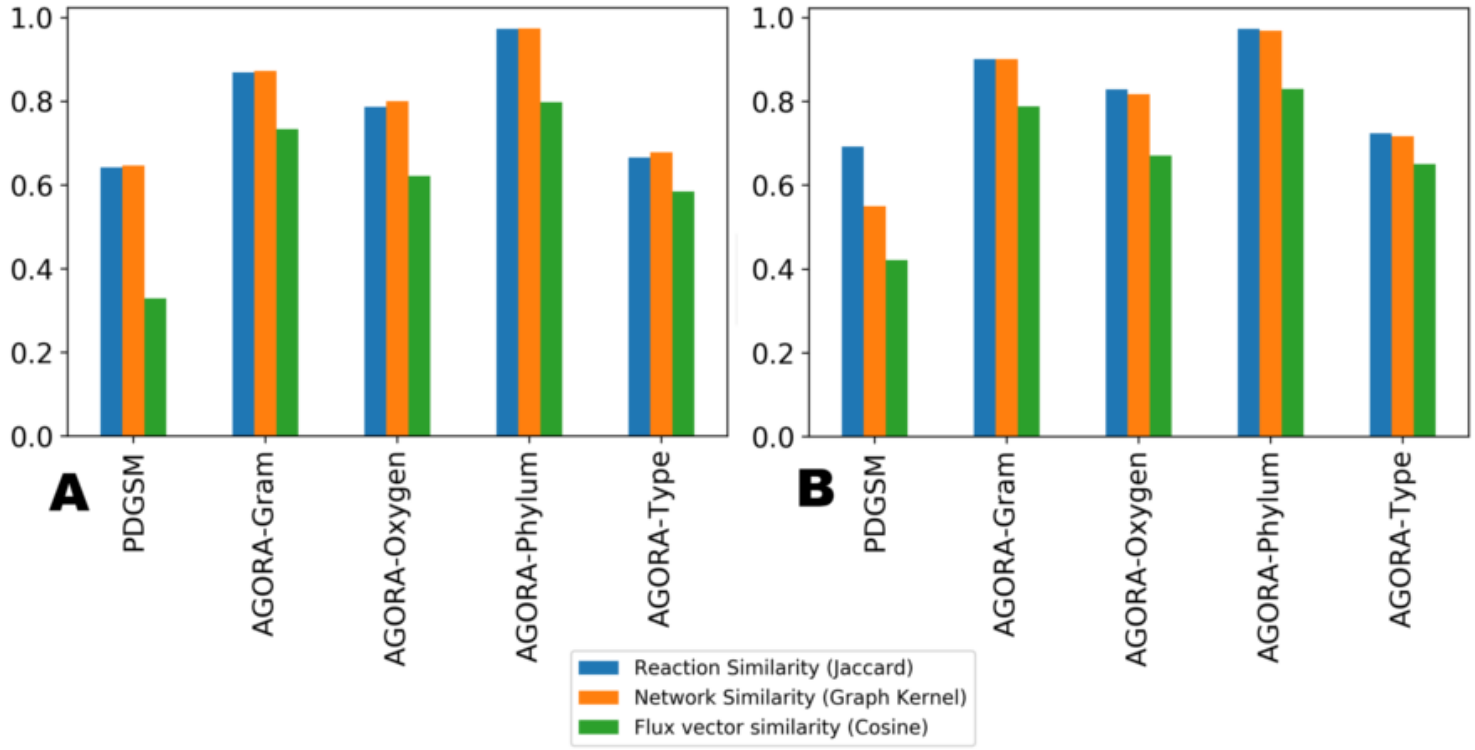

Figure 2

Classication results. (A) K-Nearest neighbor classication results. (B) Kernel-SVM classication results.

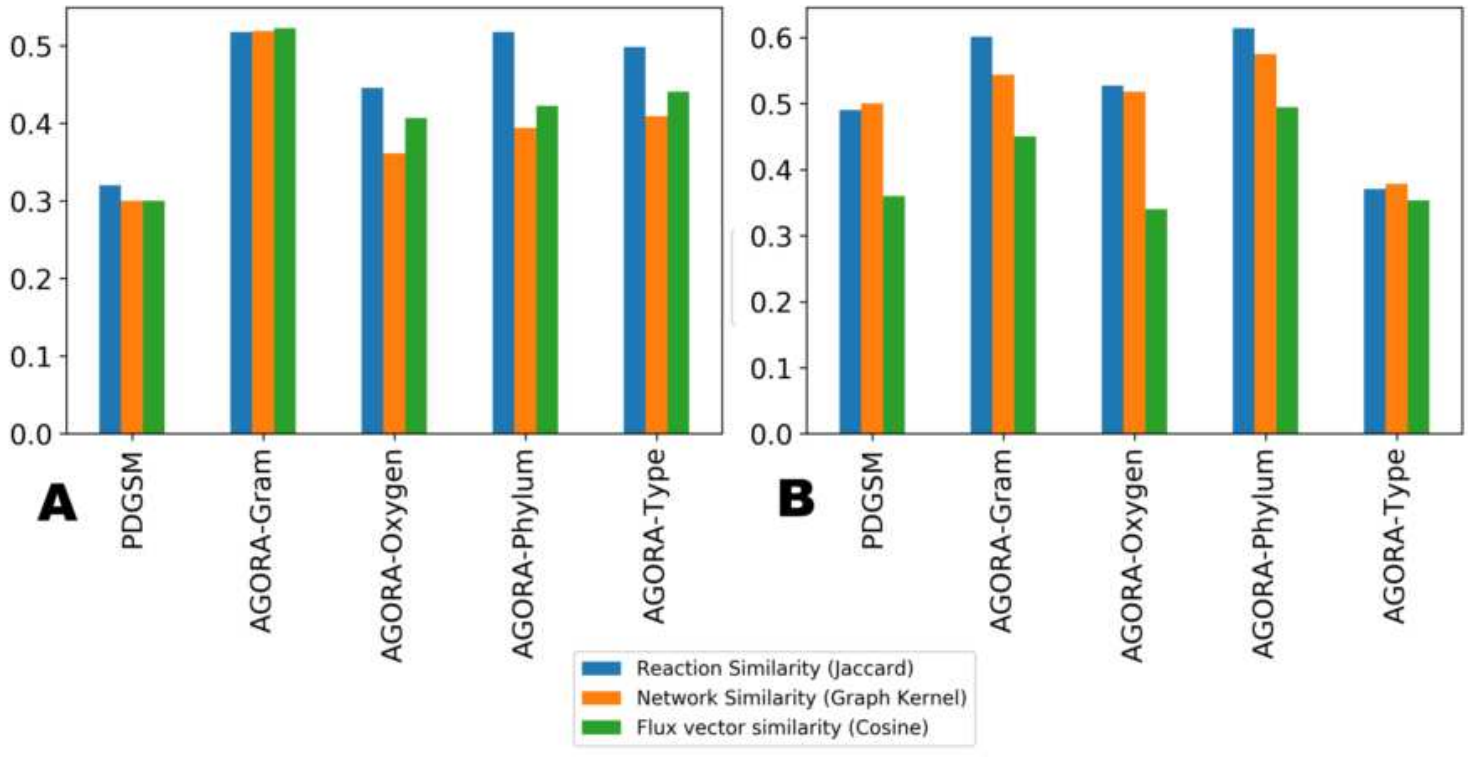

Figure 3

Clustering results. (A) Hierarchical Clustering results. (B) Spectral Clustering results.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- sup4graph.png
- sup1JDAGORAphy.png
- sup3COSAGORAphy.png
- sup5SBMLcontent2.svg.png
- sup2GKAGORAphy.png