frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 05 October 2021
doi: 10.3389/fmolb.2021.666705

Performance Assessment of the Network Reconstruction Approaches on Various Interactomes

M. Kaan Arici[1,2] *and Nurcan Tuncbag*[3,4]*

[1]Graduate School of Informatics, Middle East Technical University, Ankara, Turkey, [2]Foot and Mouth Diseases Institute, Ministry of Agriculture and Forestry, Ankara, Turkey, [3]Chemical and Biological Engineering, College of Engineering, Koc University, Istanbul, Turkey, [4]School of Medicine, Koc University, Istanbul, Turkey

Beyond the list of molecules, there is a necessity to collectively consider multiple sets of omic data and to reconstruct the connections between the molecules. Especially, pathway reconstruction is crucial to understanding disease biology because abnormal cellular signaling may be pathological. The main challenge is how to integrate the data together in an accurate way. In this study, we aim to comparatively analyze the performance of a set of network reconstruction algorithms on multiple reference interactomes. We first explored several human protein interactomes, including PathwayCommons, OmniPath, HIPPIE, iRefWeb, STRING, and ConsensusPathDB. The comparison is based on the coverage of each interactome in terms of cancer driver proteins, structural information of protein interactions, and the bias toward well-studied proteins. We next used these interactomes to evaluate the performance of network reconstruction algorithms including all-pair shortest path, heat diffusion with flux, personalized PageRank with flux, and prize-collecting Steiner forest (PCSF) approaches. Each approach has its own merits and weaknesses. Among them, PCSF had the most balanced performance in terms of precision and recall scores when 28 pathways from NetPath were reconstructed using the listed algorithms. Additionally, the reference interactome affects the performance of the network reconstruction approaches. The coverage and disease- or tissue-specificity of each interactome may vary, which may result in differences in the reconstructed networks.

Keywords: protein-protein interactions, interactome, network reconstruction, heat diffusion, personalized PageRank, prize-collecting Steiner forest, pathway reconstruction

OPEN ACCESS

Edited by:
Arzucan Ozgur,
Boğaziçi University, Turkey

Reviewed by:
Hilal Kazan,
Antalya Bilim University, Turkey
Cesim Erten,
Antalya Bilim University, Turkey
Eileen Marie Hanna,
Lebanese American University,
Lebanon

****Correspondence:***
Nurcan Tuncbag
ntuncbag@ku.edu.tr

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *10 February 2021*
Accepted: *14 July 2021*
Published: *05 October 2021*

Citation:
Arici MK and Tuncbag N (2021) Performance Assessment of the Network Reconstruction Approaches on Various Interactomes. Front. Mol. Biosci. 8:666705. doi: 10.3389/fmolb.2021.666705

INTRODUCTION

Computational approaches improve our understanding about the mechanisms of perturbations, effects of drugs, and functions of genes in the biological system by interpreting multiple "omic" data and reducing their complexity (Liu et al., 2020; Paananen and Fortino, 2020). Integrative network analysis approaches are used to interpret the complex interactions between "omic" entities as a whole beyond the list of molecules. The impact of an alteration in any omic entity, for example, upregulated

Abbreviations: APSP, all-pairs shortest paths; CDGs, cancer driver genes; FPR, false positive rate; HD, heat diffusion; HDF, heat kernel diffusion with flux; MCC, Matthew's correlation coefficient; MI, MINT inspired; PCA, principal component analysis; PCSF, prize-collecting Steiner forest; PPR, personalized PageRank; PRF, personalized PageRank with flux or downregulated genes or mutated or phosphorylated proteins, may not be local; rather, it diffuses to the distant sites of the interactome.

Many pathway databases cataloged the molecular interactions. Each database explains interactions *via* different approaches. KEGG (Kanehisa et al., 2017) provides annotated pathways, while Reactome (Jassal et al., 2020) gives detailed information on components and the reactions. Additionally, integrated interactomes such as HIPPIE, ConsensusPathDB, and STRING combine multiple resources to come up with a weighted interactome. There are several scoring schemas to measure the reliability of interactions such as MI-score and IntScore. These methods combine different weights including the number of publications, detection method, or network topology (Turinsky et al., 2011; Kamburov et al., 2012; Kamburov et al., 2013; Alanis-Lobato et al., 2017; Szklarczyk et al., 2019). Although the combination of multiple resources improves the quality of the interactomes, it still does not completely solve the bias toward well-studied proteins or the artifacts from high-throughput experiments (Žitnik et al., 2013; Caraus et al., 2015; Skinniderid et al., 2018; Vitali et al., 2018). Besides the false positives, interactomes are not complete and have false negatives which are the undetected interactions. To complete the missing parts in the interactome and to detect spurious interactions, several prediction approaches have been employed using network topology (Alkan and Erten, 2017), link prediction, protein structures (Singh et al., 2006; Tuncbag et al., 2012; Mosca et al., 2014; Segura et al., 2015; Yerneni et al., 2018; Ietswaart et al., 2021), or additional data such as gene expression (Cannistraci et al., 2013; Lei and Ruan, 2013; Hulovatyy et al., 2014; Szklarczyk et al., 2021). For example, Interactome3D uses the structural knowledge in PDB and homology-based prediction to construct a highly accurate interactome (Mosca et al., 2013). The main limitation of proteome-wide structural interactome construction is the number of structurally resolved protein complexes.

Network reconstruction approaches aim to transform the list of seed genes/proteins into their interactome-wide impact based on the topological proximity. Steiner trees/forests, statistical models, and network propagation with random walk or heat diffusion systems have been frequently used in omics data integration with the molecular interactions (Leiserson et al., 2015; Cowen et al., 2017; SeahSen et al., 2017) or identifying disease-associated pathways, subnetworks, or modules (Paull et al., 2013; Kim et al., 2015; Silverbush et al., 2019). These approaches construct context-specific subnetworks under a certain condition such as disease association or for revealing the impact of an external stimulus such as drug treatment or pathogen infection (Braunstein et al., 2019; Tabei et al., 2019). Recently, DriveWays (Baali et al., 2020), MEXCOwalk (Ahmed et al., 2020), iCell (Malod-Dognin et al., 2019), ModulOmics (Silverbush et al., 2019), and Omics Integrator (Tuncbag et al., 2016b) predicted the cancer driver modules. MEXCOwalk implements a random walk on the reference interactome by using mutation frequencies and their mutual exclusivity for the identification of the cancer driver modules. ModulOmics uses protein–protein, regulatory, and gene co-expression networks together with mutual exclusivity of mutations to identify highly functional driver modules. Omics Integrator solves the prize-collecting Steiner forest problem to construct optimal subnetworks from the single- or multi-omic datasets. Omics Integrator was applied to several conditions from cancer driver network construction (Dincer et al., 2019) and to viral infection modules in the host organisms (Sychev et al., 2017). iCell uses the matrix factorization to integrate multi-omics datasets with tissue-specific interactomes. In this study, we compared the performance of four network reconstruction approaches, all-pairs shortest path (APSP), personalized PageRank with flux (PRF), heat diffusion with flux (HDF), and prize-collecting Steiner forest (PCSF), on six different interactomes. A conceptual representation of these methods is illustrated in **Figure 1**. We did not consider the methods in this comparison that modify the underlying interactome or reconstruct regulatory networks using gene expression, such as ARACNe (Lachmann et al., 2016), GENIE (Fontaine et al., 2011), and INFERELATOR (Madar et al., 2009). APSP merges the shortest paths between pairs of nodes in the seed list. HDF implements the heat diffusion process by transferring the initial heat of the seed list to their neighbors. PRF applies a random walk to find the nodes most relevant to the seed list. We calculate the edge flux in both HDF and PRF based on the resulting node weights. PCSF finds an optimal forest that connects the seeds either directly or by adding intermediate nodes. We evaluated the performance of these algorithms on a gold standard dataset containing 32 curated pathways in the NetPath database using different metrics such as precision, recall, and MCC values. The performance of each network reconstruction approach is highly dependent on the reference interactomes. Additionally, each method has its own strengths and limitations. We found that the interactomes have some critical differences that can significantly affect the performance of network reconstruction approaches, such as their edge weight distributions, the bias toward some well-studied proteins, their coverage of disease-associated proteins, and the structurally resolved interactions. APSP has the highest recall and the lowest precision, while PRF, HDF, and PCSF have more balanced and comparable performance in precision and recall. Among them, PCSF performed the best in terms of the F1 score, which represents the balance between the precision and the recall. Overall, our study presents an extensive comparison of the selected network reconstruction approaches and shows the impact of the input interactome in their performance. This comparison presenting the strong and weak aspects of the interactomes and reconstruction approaches has the potential to be beneficial to the field.

METHODS

Reference Interactomes

We used PathwayCommons v12 (Rodchenkov et al., 2019), iRefWeb v13 (Turinsky et al., 2011), HIPPIE v2.2 (Alanis-Lobato et al., 2017), ConsensusPathDB (Kamburov et al., 2013), STRING (Szklarczyk et al., 2019, 2021), and OmniPath

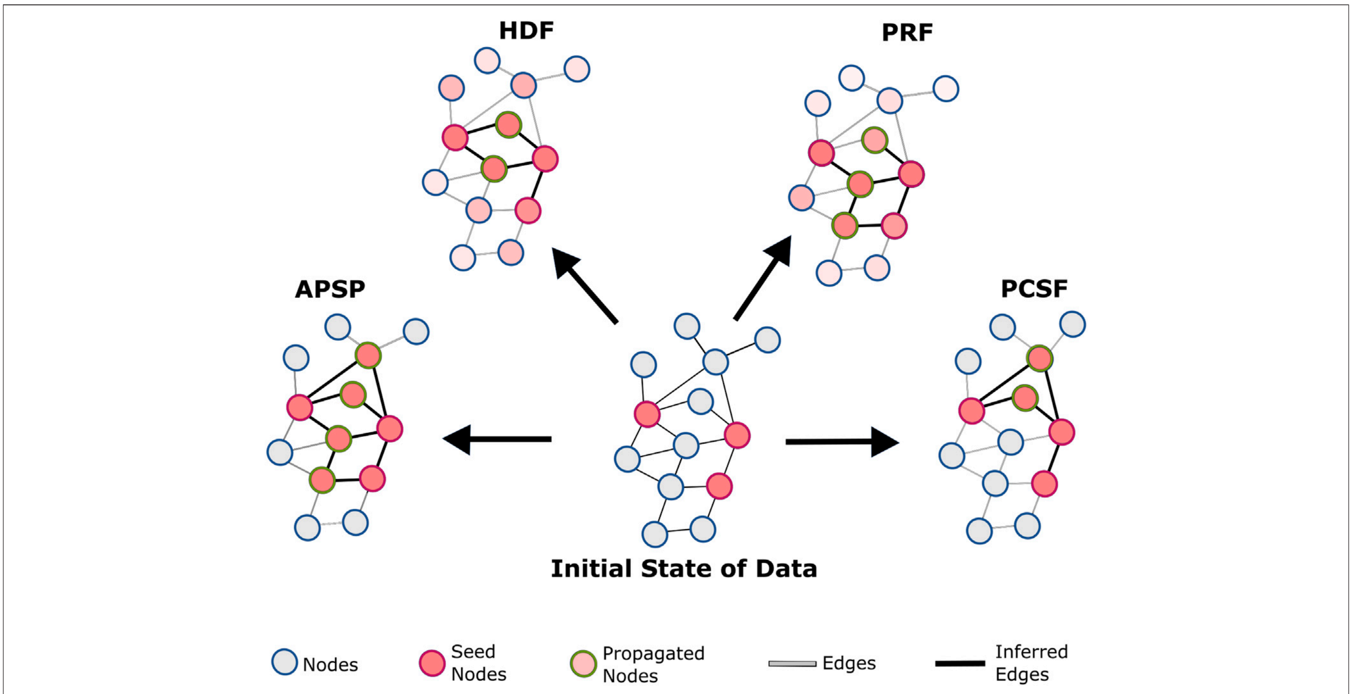

FIGURE 1 | Conceptual representation of reconstruction algorithms: all-pair shortest paths (APSP), personalized PageRank with flux (PRF), heat diffusion with flux (HDF), and prize-collecting Steiner forest (PCSF). In the APSP, the subnetwork is reconstructed with the union of all shortest paths between seed nodes. HDF diffuses the heat that initially belongs to seed nodes. After limited steps of transfer, the heat of the nodes is used for flux score calculations for edges. PRF uses a personalized PageRank algorithm to find the probability of nodes after randomly walking in the reference interactome and calculates flux scores. PCSF finds the optimum forest to link seed nodes either directly or through intermediate nodes. The union of optimum forests reconstructs subnetworks.

(Ceccarelli et al., 2020) for the interactome comparison and the assessment of subnetwork inference approaches. We mapped the names of proteins in interactomes (nodes) to their reviewed Uniprot identifiers (The UniProt Consortium, 2019). The statistics of the interactomes are listed in **Table 1**. Some interactomes have confidence scores, which represent how real an interaction is. PathwayCommons and OmniPath do not have confidence scores. iRefWeb uses the MI-scoring scheme, which considers multiple parameters including experimental detection methods. HIPPIE v2.2 and ConsensusPathDB (Release 34) have confidence scores on edges calculated based on their own scheme (Kamburov et al., 2012; Alanis-Lobato et al., 2017). We filtered the STRING interactome by recalculating confidence scores considering only the experiment and database scores (von Mering et al., 2005).

Interactome Comparison Metrics

We compared the reference interactomes at both the node and edge levels using different metrics, namely, the overlap coefficient, correlation of edge confidence scores, inclusion of disease-associated proteins, and overlap with the pathway edges. The overlap coefficient is a similarity measure for two given datasets, S_1 and S_2, which can be node sets or edge sets of graphs or information coming from a database. The overlap coefficient was calculated using **Eq. 1** for pairwise comparison of interactomes and coverage of varied knowledge (Simpson, 1966; Kuzmin et al., 2016) as follows:

$$\mathbf{overlap}\,(S_1, S_2) = \frac{|S_1 \cap S_2|}{\min\,(|S_1|, |S_2|)} \tag{1}$$

We compared each pair of interactomes, $G(V_G, E_G, c(e_G))$ and $H(V_H, E_H, c(e_H))$, where V is the node set and E is the edge set, and $0 \leq c(e) \leq 1$, where $c(e)$ is the confidence score of an edge. The node-level similarity of the given interactomes were calculated using the overlap coefficient by applying **Eq.1** where V_G is used as S_1 and V_H is S_2. Likewise, the edge level overlap coefficient in each pair of interactomes is determined using **Eq. 1** where E_G and E_H are assigned as S_1 and S_2, respectively.

Next, we explored the structurally known protein-protein interactions in each reference interactome using the overlap coefficient. Interactome INSIDER has 4,150 interactions from PDB and 2,901 interactions from Interactome3D (Meyer et al., 2018). The edge level overlap coefficient between each reference interactome (G) and each structural interactome (H) is calculated using **Eq. 1**.

The interactomes and network reconstruction methods are frequently used for revealing cancer driver modules. We downloaded the 568 cancer driver genes (CDGs) from intOGen (Martínez-Jiménez et al., 2020). The overlap coefficient between CDGs (S_1) and proteins in each reference interactome (S_2) is calculated using **Eq. 1**. Additionally, the number of publications about each CDG and the degree centrality of the CDGs are analyzed to find out the bias of the interactomes toward well-studied or cancer-associated proteins.

TABLE 1 | Reference interactomes and their statistics.

Interactome	Number of proteins	Number of interactions	Confidence score
iRefWeb v13.0	11,295	80,351	Yes
PathwayCommons v12	18,536	1,126,072	No
HIPPIE v2.2	15,984	369,584	Yes
ConsensusPathDB	17,269	359,201	Yes
STRING v11	8,992	229,306	Yes
OmniPath	6,549	35,684	No

The overlap of each reference interactome (**G**) with the known interactions in 171 pathways in KEGG (**H**) is calculated using **Eq. 1** (Kanehisa et al., 2017). Modeling a small-sized network is a challenging task because a small number of molecular interactions limit the overall dynamic range of the signals (Tkačik et al., 2009; Azpeitia et al., 2020). Therefore, we discarded KEGG pathways having less than 30 edges from the interactome evaluations.

Among the selected interactomes, iRefWeb, HIPPIE, ConsensusPathDB, and STRING have edge confidence scores that are calculated with different scoring approaches. We applied an all-pair comparison of the given interactomes (***G, H***) with the Pearson correlation analysis on the confidence scores in the intersection of edge sets in interactome pairs ($E_G \cap E_H$)

Biological networks follow the scale-free power law distribution, $P(k) = k^{-\gamma}$, where k is the degree of a node, and γ is the power coefficient (Barabási and Albert, 1995; Alm and Mack, 2016). To linearize the representation of both degree distribution and publication distribution, the logarithm of distribution was used as $log(P(k)) = -\gamma log(k)$. We collected the number of publications about each protein from UniProt. The correlation between the degree and the number of publications of the nodes was evaluated using the Pearson correlation test on a log scale.

Network Reconstruction Methods

We used four reconstruction approaches, the shortest path, heat diffusion, PageRank, and PCSF. Selected interactomes are separately employed as the reference network, $G(V, E, c(e))$, where V is the node set, E is the undirected edge set, and $c(e)$ is the weight of an edge. These networks are weighted by confidence scores in the interactions, $0 \leq c(e) \leq 1$. Network reconstruction algorithms infer the subnetwork, $R(V_R, E_R)$, where $V_R \subseteq V$ and $E_R \subseteq E$, by connecting the seed node set, $V_I \subseteq V$. The given node set is weighted with uniform $1/|V_I|$ where $|V_I|$ is the number of seed nodes, while the remaining node set is weighted as 0, so that $w(v)$ can be defined for reconstruction algorithms.

All-Pairs Shortest Paths

We found out all shortest paths between each pair of nodes, u and $v \in V_I$, $u \neq v$. When there are multiple shortest paths between u and v, we included all of them. Finally, we merged all shortest paths to obtain the final subnetwork. We did not put any edge weight–based filtering or path length threshold.

Personalized PageRank

The PageRank algorithm was normally designed for propagation in directed graphs. Personalized PageRank (PPR) is adapted to undirected graphs by converting each edge into both directed edges. The PageRank score of each node, $p(v)$, in the reference interactome, G, represents the probability of being at the node at a certain time step (t) that is calculated using the following iterative formula:

$$p_{t+1}(y) = \frac{1-\lambda}{N} + \lambda \sum_{x_i \rightarrow y} \frac{p_t(x_i)}{\deg(x_i)} \tag{2}$$

where **Eq. 2** includes the probability of node $y \in V$ that is calculated using the damping factor (λ) defining the probability of walking from neighbor nodes (x_i) to y, and N is the number of nodes (Page et al., 1998; Langville and Meyer, 2005). Initial probabilities of nodes were taken from $w(v)$. We iterated **Eq.2** 100 times by default to obtain $p(v)$.

Heat Diffusion

In the heat diffusion (HD), seed nodes having uniform heats prioritize their related nodes *via* heat transfer, which is formulated as follows:

$$p(v) = p_0\left(I + \frac{-\alpha}{N}\mathrm{L}\right)^N \tag{3}$$

In **Eq. 3**, $L = I - W$, where I represents an identity matrix and $W = D^{-1}A$ in which D and A are defined as the diagonal degree matrix and the adjacency matrix, respectively. p_0 is the initial heat vector in which nodes were weighted from $w(v)$. N and α are, respectively, the number of iterations and the heat diffusion rate. $N = 3$ is set as the default (Nitsch et al., 2010). At the end of heat diffusion, nodes have the diffused heat $p(v)$ as the weight.

Edge Selection Over Flux Scores

Personalized PageRank with flux (PRF) and heat kernel diffusion with flux (HDF) are calculated over $deg(v)$, which is defined as the number of interactions in G, and node scores $0 \leq p(v) \leq 1$, which come from PPR or HD. In our study, unlike TieDie and HotNet with heat diffusion algorithms and flux on a random walk with restart, the threshold value is employed to eliminate uncritical nodes (Vandin et al., 2011; Creighton et al., 2013; Rubel and Ritz, 2020). The related nodes with $p(v_i) \geq 1/n$ where n is the number of nodes in the interactome are considered for subnetwork reconstruction. We calculated the directional flux scores $f_{u\rightarrow t}$ using **Eq. 4** where $u, t \in V$, $p(u)$ is the score that comes from PPR or HD, and $deg(u)$ is the number of neighbors of node u. Likewise, we calculated $f_{t\rightarrow u}$ using **Eq. 5**. We determined the final flux of the edge as the minimum of $f_{u\rightarrow t}$ and $f_{t\rightarrow u}$ (**Eq. 6**).

$$f_{u\rightarrow t}(u,t) = \frac{p(u)\times c(e)}{deg(u)} \tag{4}$$

$$f_{t\rightarrow u}(t,u) = \frac{p(t)\times c(e)}{deg(t)} \tag{5}$$

$$f(e) = min\left(f_{u\rightarrow t}(u,t), f_{t\rightarrow u}(t,u)\right) \tag{6}$$

Edges are ranked from the highest flux score to the lowest by taking the negative logarithm of the flux. A total flux (F) is calculated among the related nodes as follows:

$$F = \sum f(e) \tag{7}$$

$0 \leq \tau \leq 1$, where τ is a flux threshold value that is the selection percentage of F. Edges are selected by summing flux scores from the highest to the lowest until the targeted flux amount, $\tau x F$. The edges having low flux scores are excluded from reconstructed subnetworks (Rubel and Ritz, 2020).

Prize-Collecting Steiner Forest

We used the PCSF algorithm implemented in Omics Integrator2. The seed nodes, $v_i \in V_I$, are weighted uniformly, and the edge costs are calculated using the cost function implemented in Omics Integrator 2 which combines the edge confidence score, $c(e)$, and a penalty calculated from node degrees scaled with the γ parameter. If the reference interactome does not have confidence scores, $c(e) = 1$ is uniformly defined. PCSF also penalizes the nodes based on their degrees (Tuncbag et al., 2016a). The new version, Omics Integrator 2, penalizes the edges based on the degrees of the node pair. The following function finds an optimum forest, $F(V, E)$, by minimizing the objective function (Tuncbag et al., 2013):

$$f'(F) = \sum \beta.p(v) + \sum cost(e) + \omega.\kappa \tag{8}$$

In **Eq. 8**, κ is the number of connected components, β controls the relative weight of the node prizes, and ω controls the cost of adding an additional tree to the solution network.

PCSF provides an optimum forest for each parameter set and an augmented forest which includes all the edges in the interactome that are present between the nodes in the optimal forest. We obtained the final reconstructed networks with the intersection of the optimal augmented forests that were generated using multiple parameter sets.

Performance Analysis

NetPath is the curated human signaling pathway database that is composed of immune signaling pathways and cancer signaling pathways. In this study, 32 pathways in NetPath were used as a plausible dataset (Kandasamy et al., 2010). Since the computational cost of reconstruction was expensive for all pathways in NetPath with all parameter sets, first, optimum parameter sets were determined before performance analysis.

Parameter Tuning

Parameters of reconstruction algorithms were separately optimized for each reference interactome. Thus, Wnt, TCR, TNFα, and TGFβ pathways on NetPath were used for parameter selection. Nodes in each pathway were independently shuffled and split into five-fold. Each fold was, respectively, removed from the complete pathway node list, and network reconstruction was executed with the remaining folds. Parameters of reconstruction algorithms were separately tuned for each reference interactome to maximize the F1 score (**Eq. 12**). In the APSP, all identified shortest paths among seed node sets were inserted into a reconstructed pathway without any parameter tuning, so we do not adjust any parameter. We tuned the parameters in the given interval in **Table 2** for PRF, HDF, and PCSF and for each reference interactome. Parameter sets of PRF and HDF were tuned in a two-dimensional grid *via* the mean of parameters that pooled the 10 highest F1 scores (**Supplementary Figures 1, 2**). In the PCSF, the union of all parameters that achieve the best coverage of the seed nodes, V_I, for each pathway was used as optimum parameter sets.

The Calculation of Performance Scores

After tuning the parameters on four pathways, the remaining 28 pathways in NetPath, listed in **Supplementary Table 1**, were used for performance evaluation with five-fold cross-validation. We evaluated each reconstruction algorithm separately on each reference interactome by calculating the F1 score, Matthew's correlation coefficient (MCC), recall and precision values, and false positive rate (FPR) in **Eqs 9–13** as follows:

$$recall(TP, TN) = \frac{|TP|}{(|TP| + |FN|)} \tag{9}$$

$$precision(TP, FN) = \frac{|TP|}{(|TP| + |FP|)} \tag{10}$$

$$FPR(TP, FN) = \frac{|FP|}{(|FP| + |TN|)} \tag{11}$$

$$F1_{score} = \frac{2 \times precision\ x\ recall}{precision + recall} \tag{12}$$

$$MCC(TP, TN, FP, FN) = \frac{(|TP| \times |TN|) - (|FP| \times |FN|)}{\sqrt{(|TP| + |FP|)(|TP| + |FN|)(|TN| + |FP|)(|TN| + |FN|)}} \tag{13}$$

Seed nodes were not counted in the performance calculation. However, all edges in the reconstructed network were used in the performance evaluation since interactions were not used in the initial input. For a given reference in interactome $G(V, E)$ and an seed node set (V_I) from a pathway $T(V_T, E_T)$, a network is reconstructed, $R(V_R, E_R)$, using the listed methods, where V_T;, V_R and $V_I \subseteq V$, and E_T and $E_R \subseteq E$. Node-level true positives (TP_V) and edge-level true positives (TP_E) are obtained from $|V_R \cap V_T|$ and $|E_R \cap E_T|$, respectively. Node-level true negatives (TN_v) and edge-level true negatives (TN_E) are obtained from $|V \setminus (V_R \cup V_T)|$ *and* $|E \setminus (E_R \cup E_T)|$, respectively. False positives FP_V and FP_E are

TABLE 2 | Tuning ranges of parameter sets in PageRank flux (PRF), heat diffusion flux (HDF), and prize-collecting Steiner forest.

Reconstruction algorithm	Parameter	Range	Increment
PRF	Damping factor (λ)	0–1	0.05
	Flux threshold (τ)	0–1	0.05
HDF	Heat diffusion rate(α)	0–1	0.05
	Flux threshold (τ)	0–1	0.05
PCSF	Dummy edge weight (ω)	0–5	0.5
	Edge reliability (β)	0–5	0.5
	Degree penalty (γ)	0–10	0.5

equal to $|V_R \setminus V_T|$ and $|E_R \setminus E_T|$, respectively. False negatives FN_V and FN_E, are equal to $|V_T \setminus V_R|$ and $|E_T \setminus E_R|$, respectively.

We performed principal component analysis (PCA) to figure out critical scores that explain the highest variance across all pathways. We statistically assessed overall performance data including both edge- and node-based scores by individually grouping reference interactomes and reconstruction methods.

Data Availability Statement

Codes and datasets used for this study are publicly available at the online repository https://github.com/metunetlab/Interactome_Network_Reconstruction_Assessment_2021. We downloaded NetPath, http://netpath.org/browse, and PathwayCommons, https://www.pathwaycommons.org/archives/PC2/v12/PathwayCommons12.All.hgnc.txt.gz, iRefWeb, http://wodaklab.org/iRefWeb/search/index, HIPPIE, http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/download.php, STRING, https://string-db.org/cgi/download, ConsensusPathDB, http://cpdb.molgen.mpg.de/, Reference Human Proteome from UniProtDB, https://www.uniprot.org/, using the query https://www.uniprot.org/uniprot/?query=proteome:UP000005640%20reviewed:yes, INSIDER, http://interactomeinsider.yulab.org/downloads.html, and KEGG, https://www.kegg.jp/kegg/download/ and http://rest.kegg.jp/get/+'pathwayid'+'/kgml'. OmniPath and the signaling pathways in Glioblastoma (WP2261) were retrieved from WikiPathway using Cytoscape 3.8.0.

RESULTS

Systematic Evaluation of Reference Human Interactomes

Network reconstruction algorithms are highly dependent on the quality and coverage of the reference interactome. Therefore, we systematically explored the properties of iRefWeb, PathwayCommons, HIPPIE, ConsensusPathDB, OmniPath, and STRING databases. Among them, iRefWeb, HIPPIE, ConsensusPathDB, and STRING provide the measure of confidence in interactions as scores. First, we compared the pairs of interactomes to determine how similar they are in terms of their node and edge sets. PathwayCommons is the largest network in size, so it has the highest fraction of node and edge overlap compared to all other interactomes. iRefWeb, PathwayCommons, HIPPIE, and ConsensusPathDB are the most similar interactomes to each other based on the node and edge overlaps (**Figure 2A**). On the other hand, STRING and OmniPath have fewer common nodes and edges with other interactomes. We need to note that the raw data in STRING contain more than one million interactions in human interactomes, and we used only the experimental and database interactions which resulted in a relatively small-sized interactome with medium or high confidence edges. Before using the network reconstruction algorithms, obtaining the reference interactome with measurements of interaction confidence is fundamental to decreasing the impact of the false positives. Because network reconstruction algorithms leverage the edge confidence scores and the topology of the reference interactomes during the propagation or optimization, confidence scores may substantially affect the accuracy of the resulting network. Even two topologically equivalent interactomes may produce different subnetworks as a result of network reconstruction if their confidence score distributions are different from each other. In **Figure 2B**, the number of edges in each reference interactome is shown, which are categorized as low, medium, and high confidence edges based on the interaction scores. ConsensusPathDB contains predominantly high confidence interactions, while HIPPIE and iRefWeb interactions are accumulated in medium and low confidence intervals. HIPPIE and iRefWeb use MINT-inspired (MI) confidence score calculation, while ConsensusPathDB uses the IntScore tool (Braun et al., 2009; Turner et al., 2010; Kamburov et al., 2011; Kamburov et al., 2012; Turinsky et al., 2011; Schaefer et al., 2012; Alanis-Lobato et al., 2017). We recalculated the confidence scores in STRING by considering only the experiment and database scores. PathwayCommons and OmniPath do not provide confidence scores. Edge confidence scores can be computed in various ways. Different scoring schemes lead to variation in the confidence score distributions across the interactomes. As expected, the correlation of confidence scores between HIPPIE and iRefWeb is the highest (r = 0.67, $p <$ 0.01) because both use MI-Score. The correlation between confidence scores in iRefWeb and ConsensusPathDB is very low (r = 0.25, p < 0.01) (**Figure 2C**) because ConsensusPathDB uses a different scoring scheme, IntScore. While MI-Score considers homologous interactions, the detection method, and the number of publications about the interactions, IntScore includes topological properties, literature evidence, and similarities in annotation of proteins.

Confidence scores do not completely solve the bias in the interactomes despite being a powerful measurement to filter out false positives. Therefore, we additionally analyzed the interactomes based on the bias toward well-studied proteins using different features, namely, the number of publications about the proteins, coverage of the cancer driver genes, and the number of interactions having structural details. Well-studied proteins, such as TP53 and EGFR, have hundreds of high confidence interactions in the interactomes (Schaefer et al., 2015; Chen et al., 2018; Porras et al., 2020). Indeed, there is a trade-off between the interaction confidence scores of certain proteins and systematic study bias. We used the number of publications and the degree centrality of proteins in each

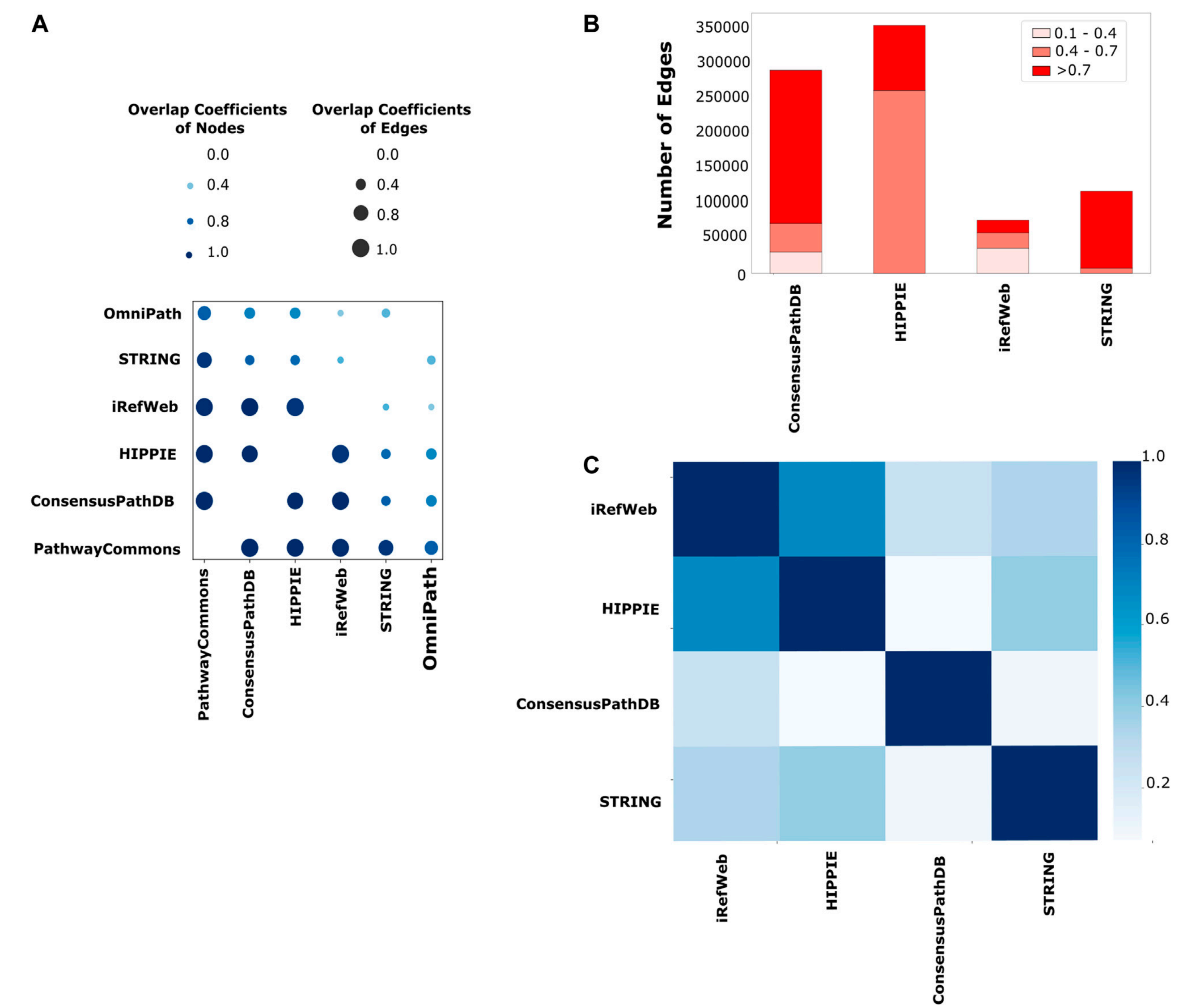

FIGURE 2 | Comparison of the reference interactomes. **(A)** Node- and edge-level commonalities between different interactomes; the node overlap score is displayed on a light-to-dark blue-color scale, while the edge similarity score is shown on a size scale where the higher similarity is represented with the larger circle. **(B)** Scores are categorized as low confidence between 0.1 and 0.4, medium confidence between 0.4 and 0.7, and high confidence between 0.7 and 1.0 for each interactome. PathwayCommons and OmniPath are not demonstrated here due to the lack of a confidence score for their edges. Edges with low confidence scores are mostly seen in iRefWeb and small portions of edges in ConsensusPathDB have low confidence scores, while low confidence edges are not seen in the filtered STRING and HIPPIE. **(C)** Different strategies on confidence score calculation are used in interactomes. Using their common edges, the correlation coefficients among their confidence scores over interactomes are demonstrated in the heatmap. The highest correlation between HIPPIE and iRefWeb is seen with the darkest blue. The same confidence score calculation strategy, MI scoring, is used in both interactomes.

reference interactome to explore if highly connected proteins are also well-studied ones.

Each analyzed interactome is a scale-free network so that their degree distributions follow the power law (**Supplementary Figure 3**) (Barabási and Albert, 1995; Vidal et al., 2011). The number of publications about proteins follows the power law distribution as their degree distribution (**Supplementary Figure 4**). Thus, the number of publications and degrees were analyzed using log-based values to find out their correlation. The number of publications and the degrees of proteins are positively correlated in all interactomes (**Figure 3A**). We observed the highest correlation in PathwayCommons (r = 0.62, p < 0.01) and HIPPIE (r = 0.61, p < 0.01), which implies the bias toward well-studied proteins in these interactomes. iRefWeb, STRING, and OmniPath have moderate correlation between the degree and the number of publications, which implies relatively less biased interactomes (**Supplementary Table 2**). We note that this comparison is performed on the whole interactome without

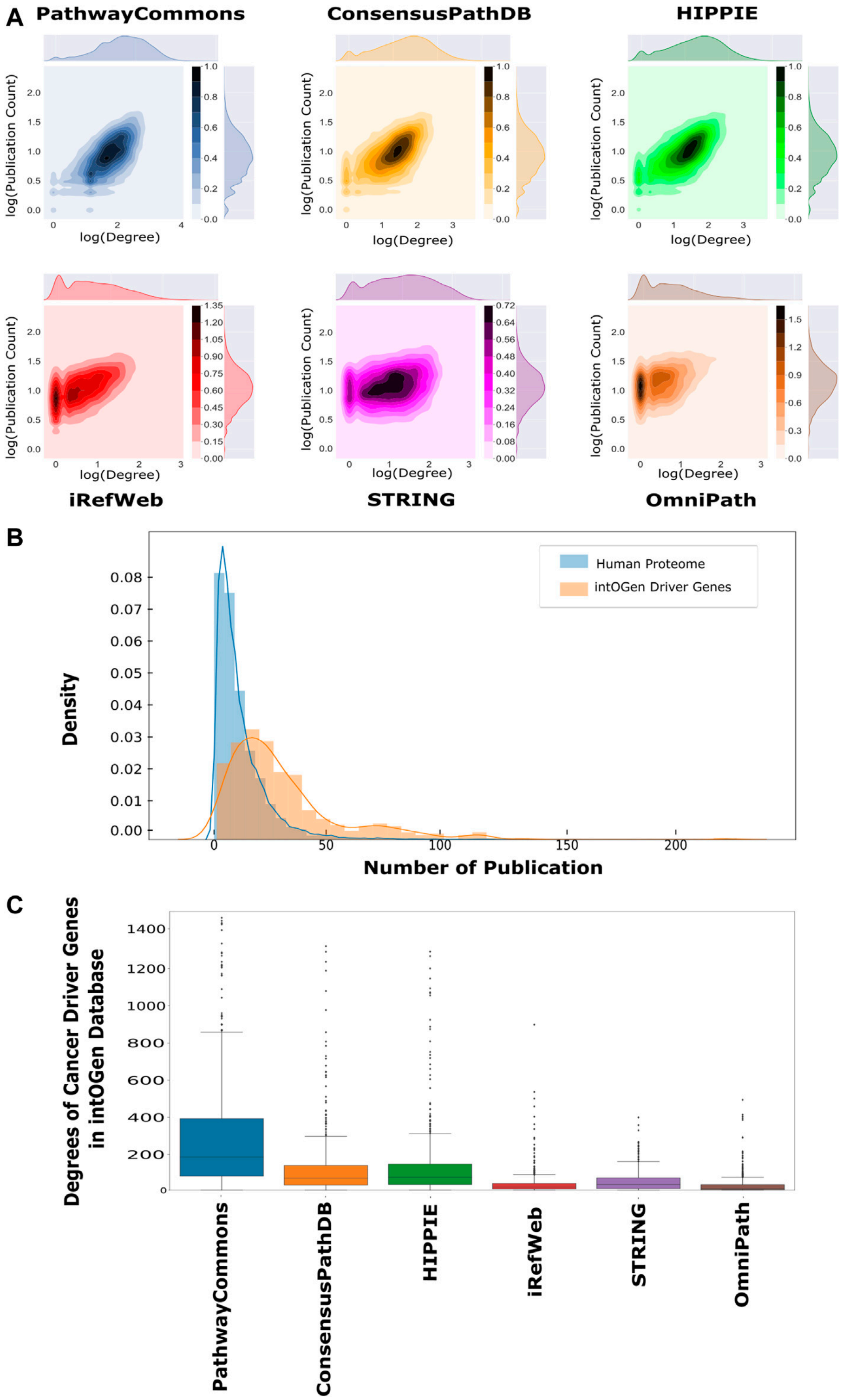

FIGURE 3 | Correlation between publication counts and degrees over interactomes. **(A)** Log–log scale joint graphs of publication distribution and degree distribution for each interactome were drawn since both follow a power-law distribution. While all interactomes have a positive correlation between the degree and publication number, PathwayCommons, HIPPIE, ConsensusPath, and iREF have well-studied hubs. On the other hand, hubs in iRefWeb and OmniPath are not composed of relatively well-studied proteins (p-values <0.001 and $r_{PathwayCommons} = 0.622$, $r_{ConsensusPathDB} = 0.556$, $r_{HIPPIE} = 0.614$, $r_{iRefWeb} = 0.508$, $r_{STRING} = 0.250$ and $r_{OmniPath} = 0.400$). **(B)** Distributions of the number of publications and cancer driver genes in the intOGen database are shown, respectively, in blue and orange. The probability of well-studied cancer driver genes (CDGs) is higher than the probability of well-studied proteins. **(C)** Driver gene degrees in the interactomes are demonstrated in the boxplot in which driver genes in PathwayCommons have more connection than other interactomes. OmniPath and iRefWeb do not have highly connected driver genes as many as ConsensusPath, HIPPIE, STRING, and PathwayCommons.

any confidence score–based filtering, except STRING. We expect that if only the high or medium confidence interactions in other interactomes would be considered, the correlations may be dramatically reduced and the bias toward well-studied proteins may be dumped. Reconstruction algorithms are also adapted to overcome this inherent bias toward the nodes and edges in interactomes. For example, heat diffusion and random walk, together with the edge flux calculation, use node degrees for normalization, while PCSF penalizes highly connected proteins (Creixell et al., 2015; Tuncbag et al., 2016a; Rubel and Ritz, 2020). In this way, false-positive edges belonging to hub nodes are excluded from the final subnetwork.

One application area of network reconstruction algorithms is the discovery of disease-associated pathways, especially in cancer, by inferring the seed proteins/genes. The resulting networks are used for patient stratification, biomarker discovery, or the analysis of drug mechanisms of action (Mo et al., 2018; Huang et al., 2019; Koh et al., 2019; Wang et al., 2021). Therefore, we searched for the coverage of the cancer driver genes (CDGs) in each interactome. CDGs provide growth advantage to the tumor cells and alter signaling pathways. Additionally, CDGs are important markers in tumor stratification, characterization, and drug development (Waks et al., 2016; Bailey et al., 2018; Zsákai et al., 2019). We obtained the list of CDGs from the intOGen database (Martínez-Jiménez et al., 2020). We found that significantly more publications are present for CDGs than for the rest of the proteomes, as shown in **Figure 3B** ($p < 0.01$). The presence of driver genes and their edges help in accurately reconstructing the driver pathways in cancer. All analyzed interactomes are highly inclusive of driver genes, especially PathwayCommons, ConsensusPathDB, and HIPPIE (**Supplementary Figure 5**). However, the degrees of CDGs in the PathwayCommons interactome are significantly higher than others (**Figure 3C**).

In terms of protein interactions, the most accurate and confident interactions can be caught by their structural identification. Structures of protein–protein complexes uncover the binding sites, domain contacts, and many more (Schmidt et al., 2014; Nero et al., 2018; Hicks et al., 2019). The only drawback is the availability of limited structural data. Despite the exponential increase in PDB with the help of the X-ray, CryoEM, and NMR techniques, the number of protein complexes can still only cover around 16% of the whole interactome (Berman et al., 2000; Mosca et al., 2013; Venko et al., 2017). Many structure-based predictive approaches are also employed to accurately identify protein–protein interactions. Therefore, we further analyzed each interactome based on the representation of structurally annotated interactions. For this purpose, we used the complexes in PDB and Interactome3D. We found that HIPPIE has the highest coverage of structurally known protein–protein interactions (**Figure 4A**). HIPPIE is followed by PathwayCommons and ConsensusPathDB. iRefWeb, OmniPath, and the filtered STRING interactome have the lowest coverages.

Another source of confident interactions is the curated pathways, despite being incomplete. Generated subnetworks are required to be biologically meaningful so that their downstream analysis can sign proper biological functions (Vidal et al., 2011; Sevimoglu and Arga, 2014). Therefore, we explored the coverage of interactomes based on the curated pathways retrieved from KEGG, which is one of the most frequently used databases for pathway annotations. We found that KEGG pathways are relatively less represented in iRefWeb, while PathwayCommons and filtered STRING highly covered them (**Figure 4B**). We need to note that some individual pathways are better covered in some interactomes although their overall coverage is relatively low (**Figure 5**). For example, the MAPK and RAS signaling pathways are better represented in OmniPath, although OmniPath has a moderate coverage of all pathways. Individual pathway coverage of each interactome is listed in **Supplementary Table**.

Performance of Network Reconstruction Algorithms

As evidenced in detail, each interactome has its own strengths and weaknesses. These properties have a direct effect on the performance of network reconstruction algorithms. Therefore, we used each interactome as the reference for each network reconstruction algorithm to monitor the variance in the performance. We used four well-established network reconstruction algorithms, the all-pair shortest paths (APSP), personalized PageRank with flux (PRF), heat diffusion with flux (HDF), and prize-collecting Steiner forest (PCSF) algorithms, to evaluate their performance on the gold standard dataset of 32 curated pathways retrieved from NetPath. Four pathways are used for parameter tuning, and the rest (28 pathways) is used for performance evaluation.

We collected both node- and edge-level performance metrics for each pair of interactomes and reconstruction methods on each pathway. We found that node-level performance is relatively more robust to different interactomes or different pathways in each approach than the edge-level performance. The largest variation is in the edge-level F1 scores, in that the balance between the recall and precision values is highly variable across pathways and interactomes (**Supplementary Figure 6**). The F1 scores ($p < 0.001$) and precision ($p < 0.001$) scores of the reconstructed pathways that are inferred from PathwayCommons are mostly lower than the scores of HIPPIE, ConsensusPathDB, OmniPath, and iRefWeb (**Figure 6A**). The second highest variation is in the edge-level MCC, used for binary classification over imbalanced data (Boughorbel et al., 2017; Magnano and Gitter, 2021). This result implies that the algorithms do not perform well with a relatively very large reference interactome because of the potential dominance of false positives over the true-positive interactions. Based on the F1 score and the precision value, we did not find a significant difference in performance when HIPPIE, ConsensusPathDB, OmniPath, or iRefWeb interactomes are used. Therefore, we continued with HIPPIE as a reference interactome for further assessments since it has the most balanced features based on the comparison in the previous part, including coverage of structurally known interactions. The comparison of edge-based performance scores showed that APSP significantly has the lowest precision values ($p <$

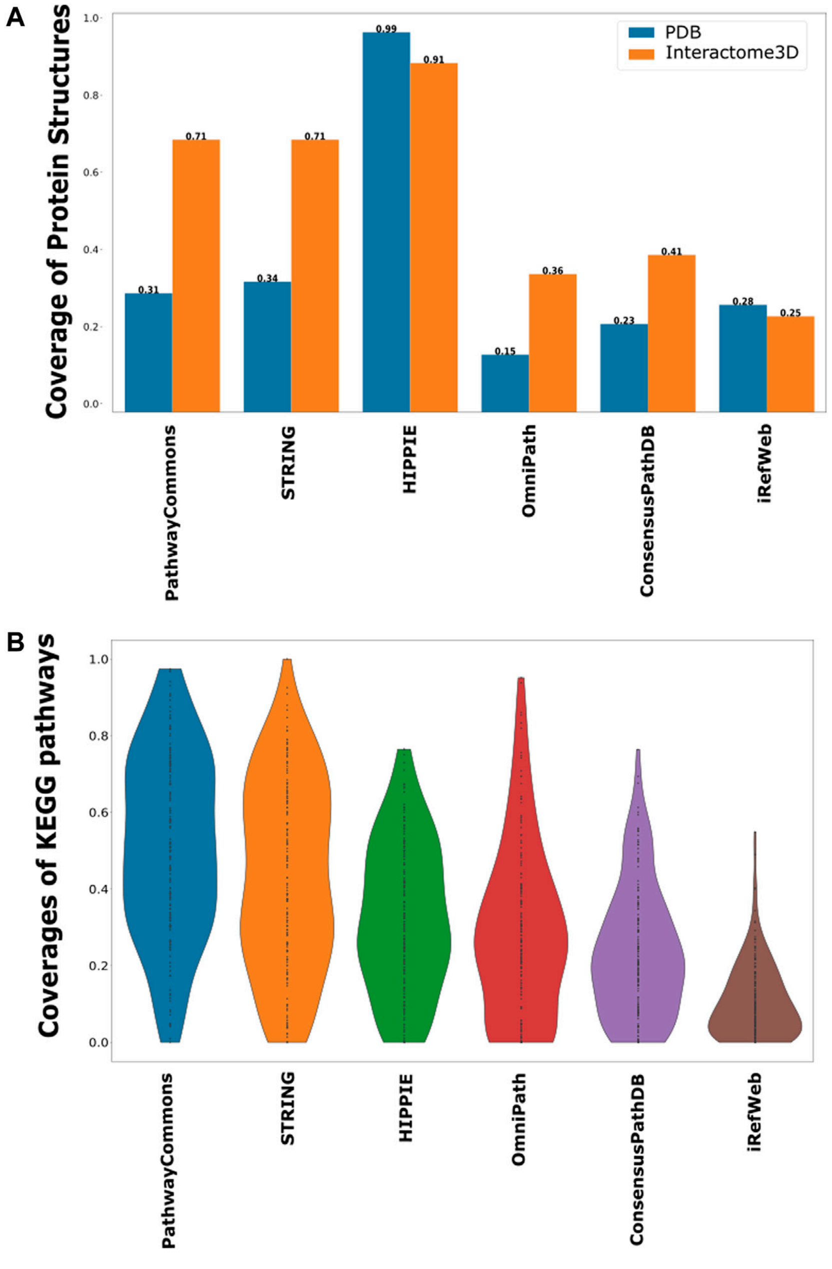

FIGURE 4 | Coverage of structurally known interactions and pathway interactions in each interactome. **(A)** Structural information is demonstrated in two groups, as known interactions in PDB in blue and predicted interactions in Interactome3D in orange. **(B)** Overlaps between the interactions in KEGG pathways and each interactome are shown as a violin-plot.

0.001) and the highest recall values among all reconstruction approaches when the performance across all pathways is evaluated. There is no significant difference in precision values between HDF, PRF, and PCSF (**Figure 6B**). The recall values of the reconstructed pathways do not significantly differ between HDF and PRF, while PCSF (p < 0.001) has significantly higher recall scores (p < 0.001) than HDF and PRF (**Figure 6C**). The trade-off between the precision and recall scores can be noticed in the results of reconstruction methods. Insertion of all shortest paths between the seed nodes in the APSP algorithm causes both the reduction in precision values and the increase in recall values. The significantly high FPR in APSP (p < 0.001) indicates that false-positive edges dominate the true-positive edges (**Supplementary Figure 7**). Therefore, F1 scores of the APSP-reconstructed pathways are significantly lower than those of other methods (p < 0.001) (**Figure 6D**). On the other hand, PCSF-reconstructed pathways have moderately high recall and precision scores and the highest F1 score by a considerable margin, optimizing the trade-off between the precision and recall values. Interestingly, the interval of recall scores in the reconstructed pathways in PCSF is not variable in a

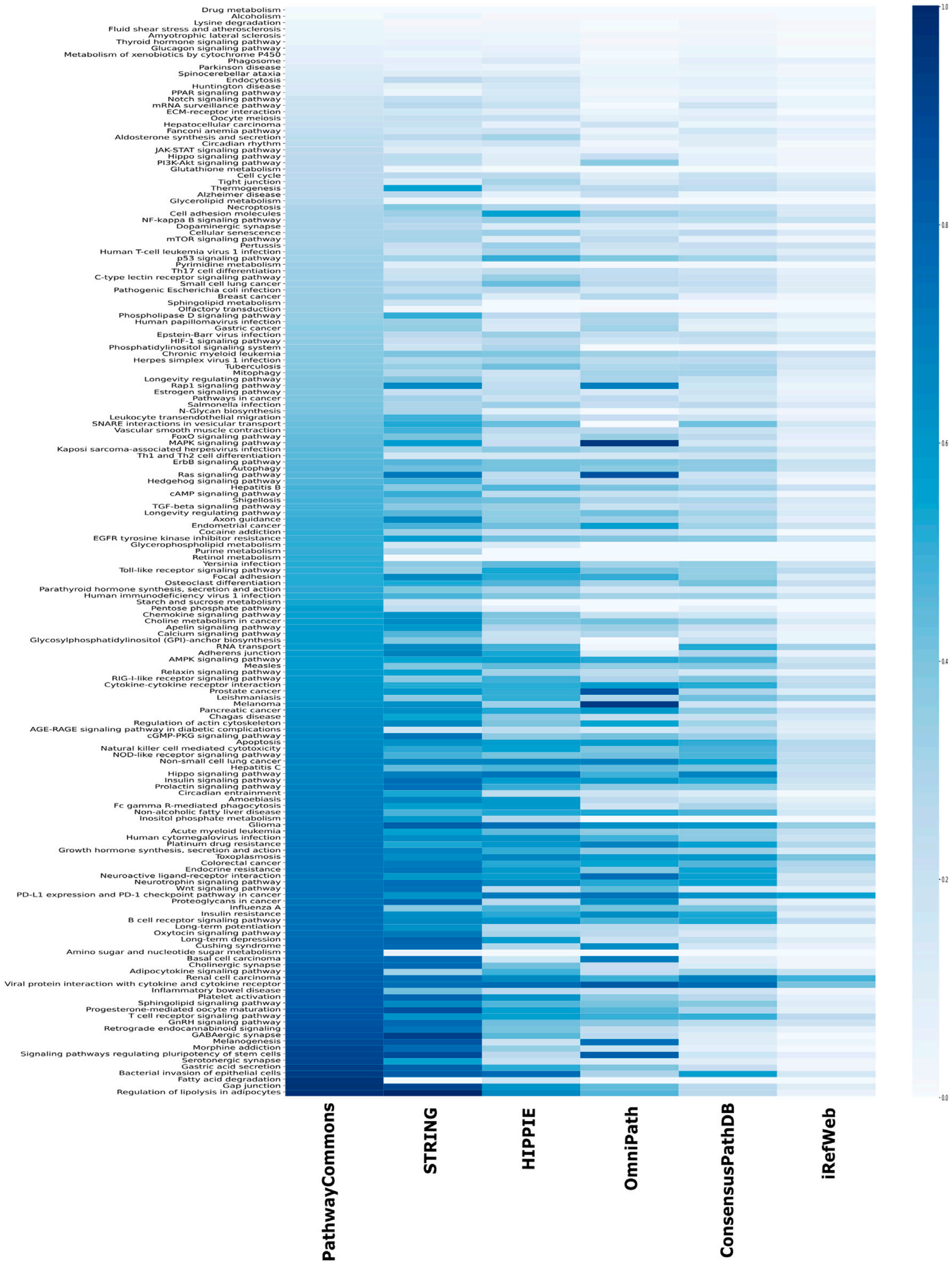

FIGURE 5 | Overlap between each pathway in KEGG and each interactome. The sorted coverage of interactomes for 171 KEGG pathways is shown as a heatmap.

wide interval as in other methods; rather, it fluctuates around 0.65. The PCSF approach gives an optimum forest as an output together with an augmented forest which includes all the edges in the interactome that are present between the nodes in the optimal forest. We obtained the final network of PCSF by taking the intersection of augmented forests from multiple parameters. In

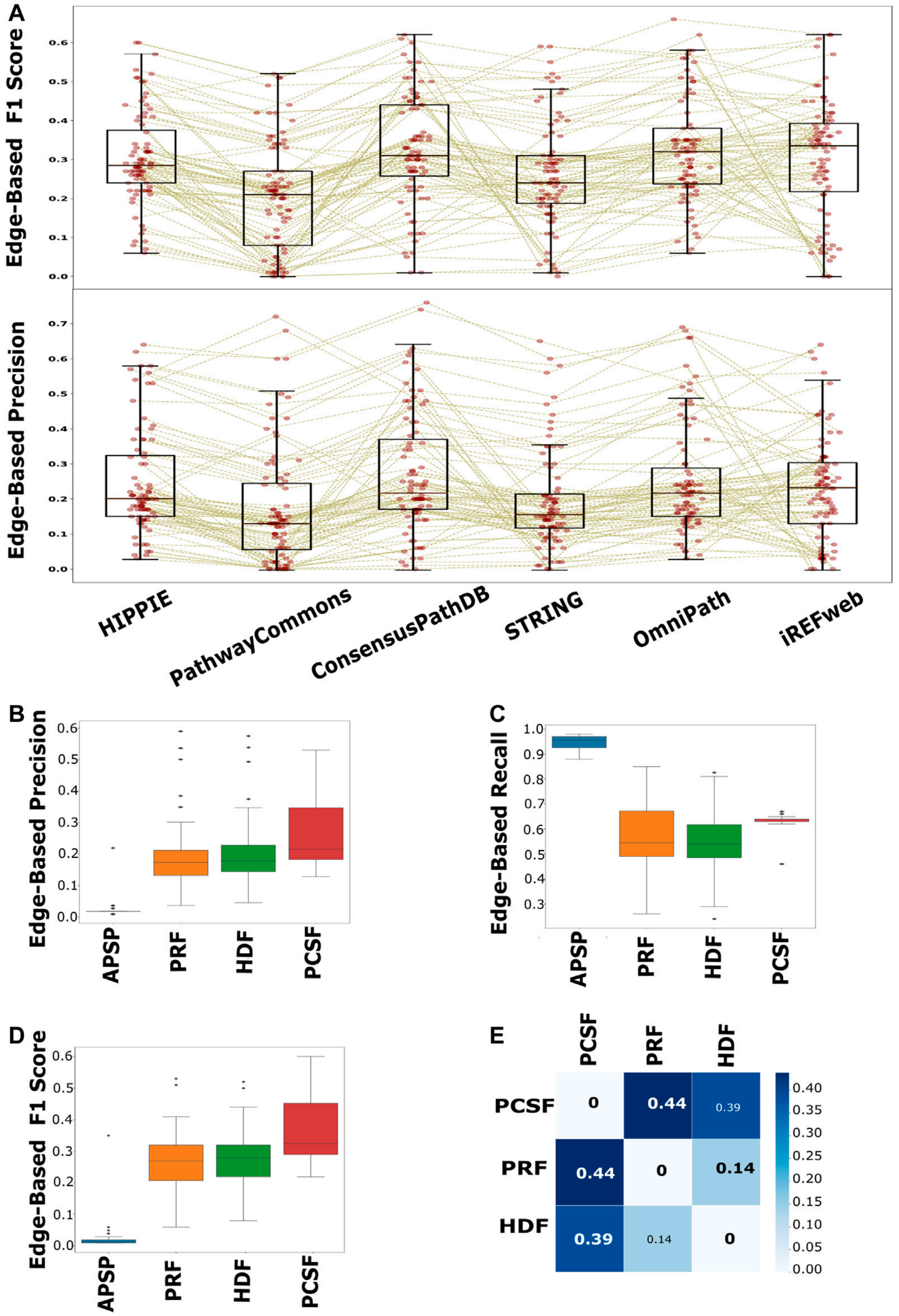

FIGURE 6 | Performance evaluation of each interactome and method in pathway reconstruction. **(A)** Boxplot of edge-based precision and F1 scores, for each interactome, shows that PathwayCommons and STRING are significantly lower than HIPPIE, ConsensusPathDB, OmniPath, and iRefWeb, while there is not any distinct difference among HIPPIE, ConsensusPathDB, OmniPath, and iRefWeb. The performance values for each reconstructed network is represented with red points in the boxplots. Brown lines connect the performance scores of the same pathway across the interactomes. **(B)** Edge-based precision, **(C)** edge-based recall, and **(D)** edge-based F1 scores are separately demonstrated for reconstruction algorithms. **(E)** HDF, PRF, and PCSF were compared in terms of the reconstructed pathways. The heatmap shows that the reconstructed pathways by PCSF are different, having %44 and %39 different edges, respectively, than the ones reconstructed by PRF and HDF.

this way, adding an edge to the final network was made very stringent. We computed the Jaccard similarity matrix among HDF, PRF, and PCSF to demonstrate the variation on the edge-level performance in the reconstructed pathways (**Figure 6E**; Ricotta et al., 2016). PCSF penalizes highly connected nodes, which reduces the dominance of well-studied or highly connected nodes in the reconstructed networks. In this way, important but low-degree nodes are also successfully included in the reconstructed pathways. As a result, PCSF has balanced precision and recall values, and its reconstructed pathways have the highest dissimilarity compared to the reconstructed pathways from other methods. Overall, the performance of the algorithms is highly affected by the parameter selection along with the used background interactome. To illustrate the reconstructed networks intuitively and to distinguish their commonalities and differences for each algorithm, we selected two case studies; one is selected from the NetPath database and the other is selected from WikiPathways.

Case Studies: Reconstruction of the Notch Pathway and Glioblastoma Disease Pathway

Our first case study is the Notch signaling pathway to intuitively illustrate the performance of each approach. The Notch signaling pathway plays a critical role in cell fate determination by regulating differentiation, apoptosis, proliferation, and morphogenesis. Its signaling cascades are associated with many human cancers (Sjölund et al., 2005; Bazzoni and Bentivegna, 2019; Guo et al., 2019). The APSP method recovers many true-positive edges, but it also introduces many false positives in the Notch pathway (**Supplementary Figure 7**). Therefore, only PRF, HDF, and PCSF results inferred from a set of seeds selected from the Notch pathway are illustrated in **Figure 7**. Notch receptors are single-pass transmembrane proteins, receiving signals from transmembrane ligands such as JAG1, JAG2, DLL1, and DLL4. The given protein list includes Notch receptors and CNTN1, JAG2, and DLL4. All reconstruction algorithms successfully identified JAG1 and the interaction between Notch receptors and their ligands except for DLL. True-positive nodes having a low degree in the reference interactome were caught better by PCSF than by PRF and HDF. Additionally, PCSF accurately included nodes such as CNTN1, WDR12, LEF1, RBX1, SIN3A, and many other true positives in the final reconstructed network. Although PCSF performs well in recovering low-degree nodes, it could not include some other nodes such as AKT1, SKP1, SPEN, and TCF3 in the pathway. PCSF successfully found the interactions between Furin–Notch receptors that regulate the Notch pathway in cancer progression where Furin, a low-degree ligand, generates biologically active heterodimer receptors (Qiu et al., 2015). On the other hand, PCSF fails to construct the interactions including low-degree nodes such as JAK2 and WDR12. HDF and PRF mostly reveal the interaction between high-degree nodes such as MAML1 and Notch receptors since the heat diffusion and the PageRank algorithm tend to give high scores to these nodes.

The Notch pathway has cross talk with other critical pathways in cancer such as the PI3K-AKT-mTOR and JAK-STAT signaling pathways (Chan et al., 2007; Hillmann and Fabbro, 2019). The cross talk is mediated by the nodes with low-degree and high betweenness centrality in the reference interactome such as PIK3R1, LCK, and JAK2. Although we could reveal these intermediate nodes that are important in cross talk between multiple pathways with PCSF, we could not achieve the same performance in the added edges. Despite correctly identifying PIK3R1 interaction with Notch1 and LCK, interactions with PIK3R2 and AKT were not found. In the JAK-STAT and Notch pathway cross talk (Rawlings et al., 2004; Liu et al., 2010), we accurately found intermediate nodes such as JAK2, HES1, and HES5, but we failed in recovering their interactions with STAT3 in the PCSF-reconstructed pathway.

Our second case study is the glioblastoma (GBM) disease pathway. Disease-related pathways are mostly composed of multiple signaling pathways. GBM is the most aggressive type of brain cancer. Multiple signaling pathways such as the PI3K/AKT/mTOR, EGFR/RAS/MAPK, P53, and RB pathways have abnormal activity in GBM tumors (Ohgaki and Kleihues, 2007). Disease-related pathways are mostly composed of multiple signaling pathways. The presence of cross talk *via* intermediate molecules is the reason why multiple pathways are related to a disease. In this regard, signaling pathways in GBM, retrieved from WikiPathways, were reconstructed by multiple algorithms using HIPPIE as the reference interactome. Multiple signaling pathways such as the PI3K/AKT/mTOR, EGFR/RAS/MAPK, P53, and RB pathways are associated with GBM. Alterations on these pathways may lead to more aggressive and invasive phenotype by disturbing DNA repair, apoptosis, and G1/S progression and enhancing cell cycle progression and cell migration (Ohgaki and Kleihues, 2007). Some nodes such as PIK3CG and CDK1NA and their interactions, mediating the cross talk between multiple pathways, were not efficiently revealed by reconstruction algorithms. CDKN1A is responsible for the inhibition of the RB signaling pathway by transducing signals coming from the PI3K/AKT/mTOR pathway. Even though the reconstructed subnetwork recovers the RB signaling pathway, all four algorithms failed in reconstructing the edges connecting two signaling pathways (**Figure 8**). Thus, these algorithms are good at revealing the mediator nodes in cross talk between pathways, but they fail in revealing the connection between them. The HDF and PRF methods ranked some nodes as important, such as APOH, FBLN5, AFP, and MMP12. Although these proteins are not present in the studied pathway, their association with GBM was previously discovered in transcriptomic or proteomic studies (Varma Polisetty et al., 2012; Kros et al., 2015; Trojan et al., 2020).

DISCUSSION

In this study, we comprehensively explored the properties of interactomes from seven sources and the performance of four network reconstruction algorithms on known pathways. Our comparison reveals that PathwayCommons, having the highest number of nodes and edges, has the highest coverage of nodes and edges across all interactomes, including CDGs, and known

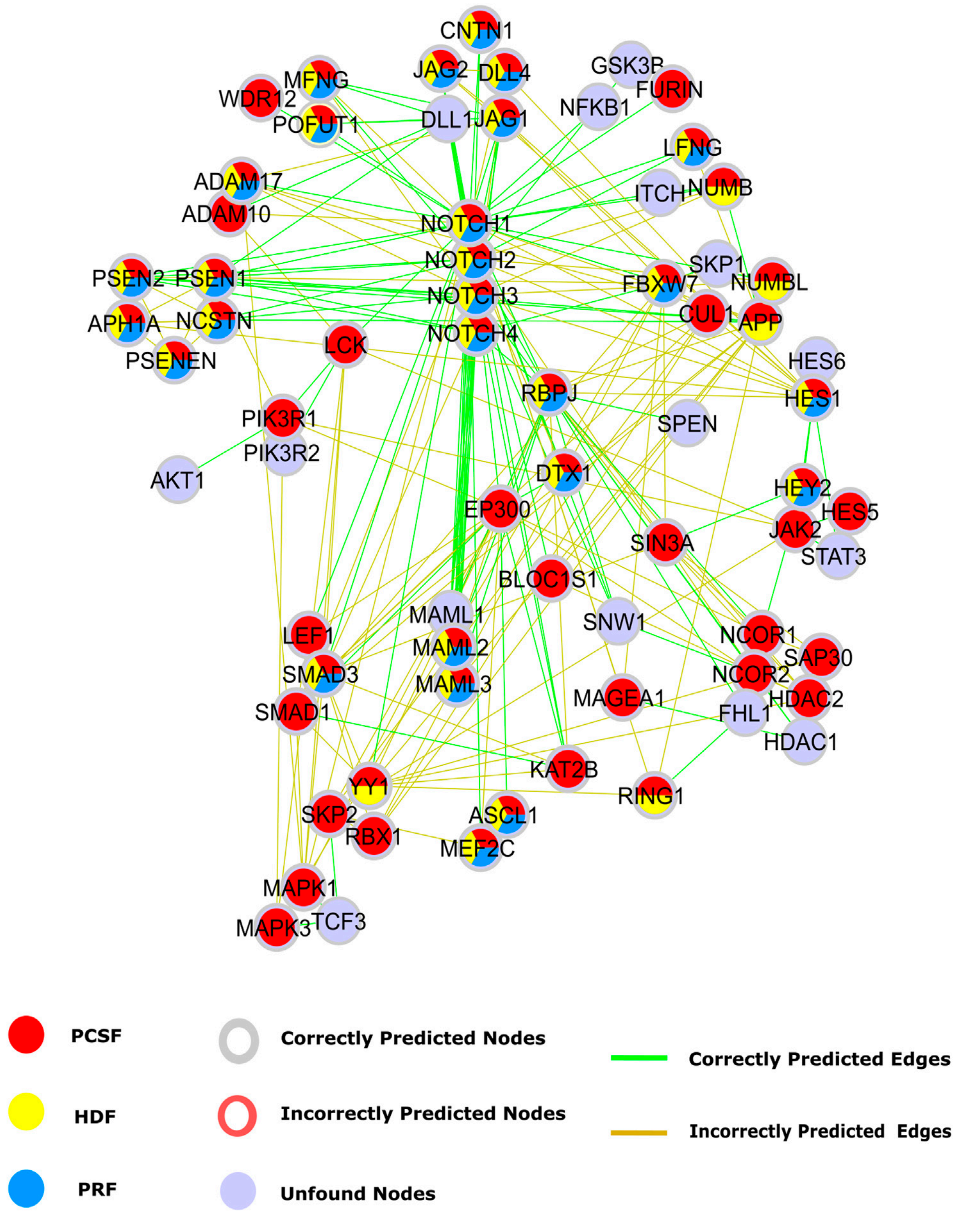

FIGURE 7 | Reconstructed Notch pathway. Nodes that are present in the pathway, but are not found by any algorithms are colored light blue. Nodes that are found by PCSF, PRF, and HDF are colored red, yellow, and cyan, respectively. Green edges are present in Notch pathway in NetPath, while incorrectly included edges by any algorithm are shown in brown.

pathways. However, precision values of the reconstruction methods are significantly lower than the others when PathwayCommons is used as the reference interactome. We did not observe a significant difference in recall values among all interactomes. The significant correlation between the degree and the number of publications of the nodes in PathwayCommons shows a bias toward well-studied proteins. Interestingly, although HIPPIE and ConsensusPathDB have a similar bias, the precision of the algorithms on these interactomes is better than that of PathwayCommons. These results imply that HIPPIE and ConsensusPathDB have a good balance in down-weighting the false positives and preserving high confidence edges.

The results of the different network reconstruction algorithms may include disjoint edges. The highest recall scores in APSP come along with the highest FPR score because the APSP algorithm adds many false-positive edges besides the true positives. Some studies, such as PathLinker (Ritz et al., 2016), use a distance threshold during shortest path calculation, a limited number of shortest paths between the source and the target, or additional data including orientation of the signal from the receptors to the transcription factors so that the false positive

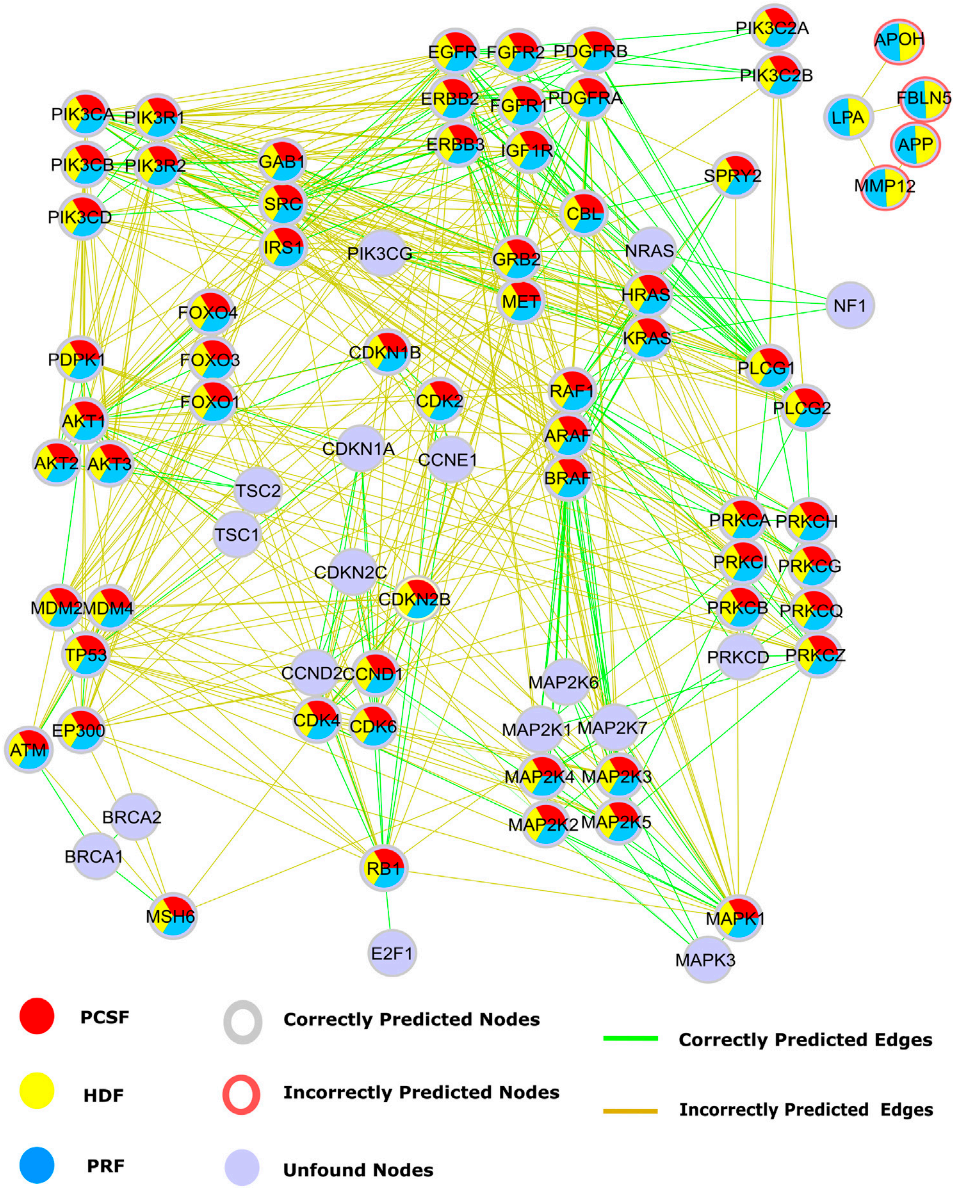

FIGURE 8 | Reconstructed signaling pathways in glioblastoma (WP2261) are predicted by PCSF are red, PRF are yellow, HDF are cyan, and nodes that are not found by any algorithm are colored light blue. Additionally, red rings represent the nodes that are incorrectly propagated. Green edges are the correct ones that are present in GBM pathway, while incorrect edges included by any algorithm are colored brown.

rate is controlled. We need to note that we did not apply any distance-based threshold, additional data, or refinement in the APSP algorithm. Thus, F1 scores and precision scores are extremely low in APSP. On the other hand, PRF, HDF, and PCSF have similar performances of false positive and true positive edges. PCSF has the highest F1 score compared to PRF and HDF. Interactomes are imbalanced datasets where true-negative edges are significantly more than true-positive edges. Naturally, the precision scores seem relatively low in the pathways formed by our algorithms since the FPR gets higher in such imbalanced datasets. The reconstructed Notch pathway shows that PCSF is better at finding weakly connected nodes. However, PCSF does not perform well in revealing the intermediate nodes and their edges achieving the cross talk between the Notch pathway and the PI3K-AKT-mTOR and JAK-STAT signaling pathways. Moreover, the intermediate nodes that links signaling pathways in GBM cannot construct completely true edges. In our study, the nodes are proteins; however, pathways may include small molecules and non-peptide nodes. Therefore, the reconstruction algorithms probably add false edges to include true terminals. The lack of some nodes in reference interactomes may be one of the reasons for the low precision scores.

Network reconstruction algorithms are highly dependent on topological properties and edge weights of the reference interactomes (Janjić and Pržulj, 2017; Liu et al., 2017). Among the evaluated approaches, the highest recall values are achieved by using the APSP algorithm together with the lowest precision values. The APSP algorithm adds many false-positive edges, besides the true positives. On the other hand, PRF, HDF, and PCSF have similar performances, while PCSF has a higher F1 score than PRF and HDF. High recall scores together with low precision scores are the result of the unbalanced data where the number of edges in the target pathway is dramatically lower than that in the rest of the interactome (Saito and Rehmsmeier, 2015). The low precision score with the moderate recall score is common among reconstruction algorithms of human signaling networks (Atias and Sharan, 2011; Ritz et al., 2016; Grimes et al., 2019). Additionally, edge-based performances of reconstruction algorithms are not as good as their node-based performance. We also observe a similar pattern of performances in our evaluation.

In a recent study, the performance of flux algorithms was shown to exceed the performance of PCSF with default parameters (Rubel and Ritz, 2020). However, the selected set of parameters significantly affects the performance of reconstruction algorithms, especially in PCSF. Automating parameter tuning that considers topological properties of reconstructed subnetworks can improve the performance (Magnano and Gitter, 2021). Therefore, in this study, we reconstructed pathways by extensively tuning the parameter set, followed by merging multiple optimal forests to reach the best performance. Parameter sets of other reconstruction algorithms were also tuned to find the optimum parameters. We can explain the overperformance of PCSF compared to other methods with detailed parameter tuning and considering multiple optimal solutions.

Several methods use topological properties of reference interactomes to predict new links and to filter out false-positive interactions (Cannistraci et al., 2013; Lei and Ruan, 2013; Hulovatyy et al., 2014; Alkan and Erten, 2017). Additionally, functional annotations, protein structures, and domain–domain interactions were also used to identify missing protein associations (Singh et al., 2006; Segura et al., 2015; Yerneni et al., 2018; Ietswaart et al., 2021). We need to note that we did not use the methods that modify the underlying interactome (Alanis-Lobato et al., 2018) and the methods that construct regulatory networks (Madar et al., 2009; Fontaine et al., 2011; Lachmann et al., 2016) in our evaluation. The performance of the APSP, HDF, PRF, and PCSF algorithms may change upon any modification or refinement of the reference interactomes. These reference interactomes are undirected graphs, but signaling pathways are intrinsically directed graphs. Indeed, the directionality of the edges can be incorporated either with the known or with the predicted ones. Orientation of the reconstructed networks can improve the mechanistic understanding of biological pathways. Therefore, using a directed reference interactome can boost the performance of each algorithm. Finally, biomolecular interactions are temporally and spatially diverse. Interactomes are incomplete sets of interactions, and the time dimension is not considered in our evaluation. Subnetwork reconstruction algorithms may be improved in the future to include biological annotations and temporal and spatial interactions of proteins.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

Conceptualization: MA and NT. Data curation: MA. Formal analysis: MA and NT. Methodology: MA and NT. Project administration: NT. Supervision: NT. Visualization: MA.

FUNDING

NT has received support from the Career Development Program of TUBITAK under the project number 117E192. MA has been financially supported with the TUBITAK-2211 fellowship.

ACKNOWLEDGMENTS

NT acknowledges the support from the UNESCO-L'Oreal National for Women in Science Fellowship and the UNESCO-L'Oréal International Rising Talent Fellowship and TUBA-GEBIP.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.666705/full#supplementary-material

REFERENCES

Ahmed, R., Baali, I., Erten, C., Hoxha, E., and Kazan, H. (2020). MEXCOwalk: Mutual Exclusion and Coverage Based Random Walk to Identify Cancer Modules. *Bioinformatics* 36 (3), 872–879. doi:10.1093/bioinformatics/btz655

Alanis-Lobato, G., Andrade-Navarro, M. A., and Schaefer, M. H. (2017). HIPPIE v2.0: Enhancing Meaningfulness and Reliability of Protein-Protein Interaction Networks. *Nucleic Acids Res.* 45, D408–D414. doi:10.1093/nar/gkw985

Alanis-Lobato, G., Mier, P., and Andrade-Navarro, M. (2018). The Latent Geometry of the Human Protein Interaction Network. *Bioinformatics* 34 (16), 2826–2834. doi:10.1093/bioinformatics/bty206

Alkan, F., and Erten, C. (2017). RedNemo: Topology-Based PPI Network Reconstruction via Repeated Diffusion with Neighborhood Modifications. *Bioinformatics* 33 (4), btw655–544. doi:10.1093/bioinformatics/btw655

Alm, J. F., and Mack, K. M. L. (2016). Degree-correlation, Robustness, and Vulnerability in Finite Scale-free Networks. *Asian Res. J. Maths.* 2 (5), 1–6. http://arxiv.org/abs/1606.08768.

Atias, N., and Sharan, R. (2011). An Algorithmic Framework for Predicting Side Effects of Drugs. *J. Comput. Biol.* 18 (3), 207–218. doi:10.1089/cmb.2010.0255

Azpeitia, E., Balanzario, E. P., and Wagner, A. (2020). Signaling Pathways Have an Inherent Need for Noise to Acquire Information. *BMC Bioinformatics* 21 (1). doi:10.1186/s12859-020-03778-x

Baali, I., Erten, C., and Kazan, H. (2020). DriveWays: a Method for Identifying Possibly Overlapping Driver Pathways in Cancer. *Sci. Rep.* 10 (1), 1–14. doi:10.1038/s41598-020-78852-8

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* 173 (2), 371–e18. doi:10.1016/j.cell.2018.02.060

Barabási, A.-L., and Albert, R. (1995). Emergence of Scaling in Random Networks. *Mat. Res. Soc. Symp. Proc.* 286, 509. doi:10.1126/science.286.5439.509

Bazzoni, R., and Bentivegna, A. (2019). Role of Notch Signaling Pathway in Glioblastoma Multiforme Pathogenesis. *Cancers* 11 (3), 292. doi:10.3390/cancers11030292

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28 (1), 235–242. doi:10.1093/nar/28.1.235

Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal Classifier for Imbalanced Data Using Matthews Correlation Coefficient Metric. *PLOS ONE* 12 (6), e0177678. doi:10.1371/journal.pone.0177678

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., et al. (2009). An Experimentally Derived Confidence Score for Binary Protein-Protein Interactions. *Nat. Methods* 6 (1), 91–97. doi:10.1038/nmeth.1281

Braunstein, A., Ingrosso, A., and Muntoni, A. P. (2019). Network Reconstruction from Infection Cascades. *J. R. Soc. Interf.* 16 (151), 20180844. doi:10.1098/rsif.2018.0844

Cannistraci, C. V., Alanis-Lobato, G., and Ravasi, T. (2013). Minimum Curvilinearity to Enhance Topological Prediction of Protein Interactions by Network Embedding. *Bioinformatics* 29, i199–209. doi:10.1093/bioinformatics/btt208

Caraus, I., Alsuwailem, A. A., Nadon, R., and Makarenkov, V. (2015). Detecting and Overcoming Systematic Bias in High-Throughput Screening Technologies: a Comprehensive Review of Practical Issues and Methodological Solutions. *Brief. Bioinform.* 16 (6), 974–986. doi:10.1093/bib/bbv004

Ceccarelli, F., Turei, D., Gabor, A., and Saez-Rodriguez, J. (2020). Bringing Data from Curated Pathway Resources to Cytoscape with OmniPath. *Bioinformatics* 36 (8), 2632–2633. doi:10.1093/bioinformatics/btz968

Chan, S. M., Weng, A. P., Tibshirani, R., Aster, J. C., and Utz, P. J. (2007). Notch Signals Positively Regulate Activity of the mTOR Pathway in T-Cell Acute Lymphoblastic Leukemia. *Blood* 110 (1), 278–286. doi:10.1182/blood-2006-08-039883

Chen, Z., Oh, D., Dubey, A. K., Yao, M., Yang, B., Groves, J. T., et al. (2018). EGFR Family and Src Family Kinase Interactions: Mechanics Matters? *Curr. Opin. Cel Biol.* 51, 97–102. doi:10.1016/j.ceb.2017.12.003

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network Propagation: A Universal Amplifier of Genetic Associations. *Nat. Rev. Genet.* 18 (9), 551–562. doi:10.1038/nrg.2017.38

Creighton, C. J., Morgan, M., Gunaratne, P. H., Wheeler, D. A., Gibbs, R. A., Robertson, G., et al. (2013). Comprehensive Molecular Characterization of clear Cell Renal Cell Carcinoma. *Nature* 499 (7456), 43–49. doi:10.1038/nature12222

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., et al. (2015). Pathway and Network Analysis of Cancer Genomes. *Nat. Methods* 12 (7), 615–621. doi:10.1038/nmeth.3440

Dincer, C., Kaya, T., Keskin, O., Gursoy, A., and Tuncbag, N. (2019). 3D Spatial Organization and Network-Guided Comparison of Mutation Profiles in Glioblastoma Reveals Similarities across Patients. *Plos Comput. Biol.* 15 (9), e1006789. doi:10.1371/journal.pcbi.1006789

Fontaine, J.-F., Priller, F., Barbosa-Silva, A., and Andrade-Navarro, M. A. (2011). Génie: Literature-Based Gene Prioritization at Multi Genomic Scale. *Nucleic Acids Res.* 39 (Suppl. 2), W455–W461. doi:10.1093/nar/gkr246

Grimes, T., Potter, S. S., and Datta, S. (2019). Integrating Gene Regulatory Pathways into Differential Network Analysis of Gene Expression Data. *Sci. Rep.* 9 (1). doi:10.1038/s41598-019-41918-3

Guo, J., Li, P., Liu, X., and Li, Y. (2019). NOTCH Signaling Pathway and Non-coding RNAs in Cancer. *Pathol. Res. Pract.* 215 (11), 152620. doi:10.1016/j.prp.2019.152620

Hicks, M., Bartha, I., Di Iulio, J., Venter, J. C., and Telenti, A. (2019). Functional Characterization of 3D Protein Structures Informed by Human Genetic Diversity. *Proc. Natl. Acad. Sci. USA* 116 (18), 8960–8965. doi:10.1073/pnas.1820813116

Hillmann, P., and Fabbro, D. (2019). PI3K/mTOR Pathway Inhibition: Opportunities in Oncology and Rare Genetic Diseases. *Int. J. Mol. Sci.* 20 (22), 5792. doi:10.3390/ijms20225792

Huang, L., Brunell, D., Stephan, C., Mancuso, J., Yu, X., He, B., et al. (2019). Driver Network as a Biomarker: Systematic Integration and Network Modeling of Multi-Omics Data to Derive Driver Signaling Pathways for Drug Combination Prediction. *Bioinformatics* 35 (19), 3709–3717. doi:10.1093/bioinformatics/btz109

Hulovatyy, Y., Solava, R. W., and Milenković, T. (2014). Revealing Missing Parts of the Interactome via Link Prediction. *PLoS ONE* 9 (3), e90073. doi:10.1371/journal.pone.0090073

Ietswaart, R., Gyori, B. M., Bachman, J. A., Sorger, P. K., and Churchman, L. S. (2021). GeneWalk Identifies Relevant Gene Functions for a Biological Context Using Network Representation Learning. *Genome Biol.* 22 (1), 55. doi:10.1186/s13059-021-02264-8

Janjić, V., and Pržulj, N. (2017). The Topology of the Growing Human Interactome Data. *J. Integr. Bioinformatics* 11 (2), 27–42. doi:10.1515/jib-2014-238

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., et al. (2020). The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* 48 (D1), D498–D503. doi:10.1093/nar/gkz1031

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., and Herwig, R. (2011). ConsensusPathDB: Toward a More Complete Picture of Cell Biology. *Nucleic Acids Res.* 39 (Suppl. 1), D712–D717. doi:10.1093/nar/gkq1156

Kamburov, A., Stelzl, U., and Herwig, R. (2012). IntScore: A Web Tool for Confidence Scoring of Biological Interactions. *Nucleic Acids Res.* 40 (W1), W140–W146. doi:10.1093/nar/gks492

Kamburov, A., Stelzl, U., Lehrach, H., and Herwig, R. (2013). The ConsensusPathDB Interaction Database: 2013 Update. *Nucleic Acids Res.* 41 (D1), D793–D800. doi:10.1093/nar/gks1055

Kandasamy, K., Mohan, S., Raju, R., Keerthikumar, S., Kumar, G. S. S., Venugopal, A. K., et al. (2010). NetPath: A Public Resource of Curated Signal Transduction Pathways. *Genome Biol.* 11 (1), R3. doi:10.1186/gb-2010-11-1-r3

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res.* 45 (D1), D353–D361. doi:10.1093/nar/gkw1092

Kim, Y.-A., Cho, D.-Y., Dao, P., and Przytycka, T. M. (2015). MEMCover: Integrated Analysis of Mutual Exclusivity and Functional Network Reveals Dysregulated Pathways across Multiple Cancer Types. *Bioinformatics* 31 (12), i284–i292. doi:10.1093/bioinformatics/btv247

Koh, H. W. L., Fermin, D., Vogel, C., Choi, K. P., Ewing, R. M., and Choi, H. (2019). iOmicsPASS: Network-Based Integration of Multiomics Data for Predictive Subnetwork Discovery. *Npj Syst. Biol. Appl.* 5 (1), 22. doi:10.1038/s41540-019-0099-y

Kros, J. M., Huizer, K., Hernández-Laín, A., Marucci, G., Michotte, A., Pollo, B., et al. (2015). Evidence-based Diagnostic Algorithm for Glioma: Analysis of the Results of Pathology Panel Review and Molecular Parameters of EORTC 26951 and 26882 Trials. *J. Clin. Oncol.* 33 (17), 1943–1950. doi:10.1200/JCO.2014.59.0166

Kuzmin, K., Gaiteri, C., and Szymanski, B. K. (2016). Synergy Landscapes: A Multilayer Network for Collaboration in Biological Research. *Adv. Netw. Sci.* 9564, 205–212. doi:10.1007/978-3-319-28361-6_18

Lachmann, A., Giorgi, F. M., Lopez, G., and Califano, A. (2016). ARACNe-AP: Gene Network Reverse Engineering through Adaptive Partitioning Inference of Mutual Information. *Bioinformatics* 32 (14), 2233–2235. doi:10.1093/bioinformatics/btw216

Langville, A. N., and Meyer, C. D. (2005). A Survey of Eigenvector Methods for Web Information Retrieval. *SIAM Rev.* 47 (1), 135–161. doi:10.1137/S0036144503424786

Lei, C., and Ruan, J. (2013). A Novel Link Prediction Algorithm for Reconstructing Protein-Protein Interaction Networks by Topological Similarity. *Bioinformatics* 29 (3), 355–364. doi:10.1093/bioinformatics/bts688

Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015). Pan-cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes. *Nat. Genet.* 47 (2), 106–114. doi:10.1038/ng.3168

Liu, C., Ma, Y., Zhao, J., Nussinov, R., Zhang, Y.-C., Cheng, F., et al. (2020). Computational Network Biology: Data, Models, and Applications. *Phys. Rep.* 846, 1–66. doi:10.1016/j.physrep.2019.12.004

Liu, G., Wang, H., Chu, H., Yu, J., and Zhou, X. (2017). Functional Diversity of Topological Modules in Human Protein-Protein Interaction Networks. *Sci. Rep.* 7 (1), 16199. doi:10.1038/s41598-017-16270-z

Liu, W., Singh, S. R., and Hou, S. X. (2010). JAK-STAT Is Restrained by Notch to Control Cell Proliferation of theDrosophilaintestinal Stem Cells. *J. Cel. Biochem.* 109 (5), a–n. doi:10.1002/jcb.22482

Madar, A., Greenfield, A., Ostrer, H., Vanden-Eijnden, E., and Bonneau, R. (2009). The Inferelator 2.0: A Scalable Framework for Reconstruction of Dynamic Regulatory Network Models. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2009, 5448–5451. doi:10.1109/IEMBS.2009.5334018

Magnano, C. S., and Gitter, A. (2021). Automating Parameter Selection to Avoid Implausible Biological Pathway Models. *Npj Syst. Biol. Appl.* 7 (1), 1–12. doi:10.1038/s41540-020-00167-1

Malod-Dognin, N., Petschnigg, J., Windels, S. F. L., Povh, J., Hemingway, H., Ketteler, R., et al. (2019). Towards a Data-Integrated Cell. *Nat. Commun.* 10 (1), 805. doi:10.1038/s41467-019-08797-8

Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., et al. (2020). A Compendium of Mutational Cancer Driver Genes. *Nat. Rev. Cancer* 20, 555–572. doi:10.1038/s41568-020-0290-x

Meyer, M. J., Beltrán, J. F., Liang, S., Fragoza, R., Rumack, A., Liang, J., et al. (2018). Interactome INSIDER: a Structural Interactome Browser for Genomic Studies. *Nat. Methods* 15 (2), 107–114. doi:10.1038/nmeth.4540

Mo, Q., Shen, R., Guo, C., Vannucci, M., Chan, K. S., and Hilsenbeck, S. G. (2018). A Fully Bayesian Latent Variable Model for Integrative Clustering Analysis of Multi-type Omics Data. *Biostatistics* 19 (1), 71–86. doi:10.1093/biostatistics/kxx017

Mosca, R., Céol, A., and Aloy, P. (2013). Interactome3D: Adding Structural Details to Protein Networks. *Nat. Methods* 10 (1), 47–53. doi:10.1038/nmeth.2289

Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P. (2014). 3did: A Catalog of Domain-Based Interactions of Known Three-Dimensional Structure. *Nucl. Acids Res.* 42 (D1), D374–D379. doi:10.1093/nar/gkt887

Nero, T. L., Parker, M. W., and Morton, C. J. (2018). Protein Structure and Computational Drug Discovery. *Biochem. Soc. Trans.* 46 (5), 1367–1379. doi:10.1042/BST20180202

Nitsch, D., Gonçalves, J. P., Ojeda, F., de Moor, B., and Moreau, Y. (2010). Candidate Gene Prioritization by Network Analysis of Differential Expression Using Machine Learning Approaches. *BMC Bioinformatics* 11, 460. doi:10.1186/1471-2105-11-460

Ohgaki, H., and Kleihues, P. (2007). Genetic Pathways to Primary and Secondary Glioblastoma. *Am. J. Pathol.* 170 (5), 1445–1453. doi:10.2353/ajpath.2007.070011

Paananen, J., and Fortino, V. (2020). An Omics Perspective on Drug Target Discovery Platforms. *Brief. Bioinform.* 21 (6), 1937–1953. doi:10.1093/bib/bbz122

Page, L. B., Brin, S., Motwani, R., and Winograd, T. (1998). *The PageRank Citation Ranking: Bringing Order to the Web*.

Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D., and Stuart, J. M. (2013). Discovering Causal Pathways Linking Genomic Events to Transcriptional States Using Tied Diffusion through Interacting Events (TieDIE). *Bioinformatics* 29 (21), 2757–2764. doi:10.1093/bioinformatics/btt471

Porras, P., Barrera, E., Bridge, A., del-Toro, N., Cesareni, G., Duesbury, M., et al. (2020). Towards a Unified Open Access Dataset of Molecular Interactions. *Nat. Commun.* 11 (1), 1–12. doi:10.1038/s41467-020-19942-z

Qiu, H., Tang, X., Ma, J., Shaverdashvili, K., Zhang, K., and Bedogni, B. (2015). Notch1 Autoactivation via Transcriptional Regulation of Furin, Which Sustains Notch1 Signaling by Processing Notch1-Activating Proteases ADAM10 and Membrane Type 1 Matrix Metalloproteinase. *Mol. Cel Biol.* 35 (21), 3622–3632. doi:10.1128/mcb.00116-15

Rawlings, J. S., Rosler, K. M., and Harrison, D. A. (2004). The JAK/STAT Signaling Pathway. *J. Cel Sci.* 117 (8), 1281–1283. doi:10.1242/jcs.00963

Ricotta, C., Podani, J., and Pavoine, S. (2016). A Family of Functional Dissimilarity Measures for Presence and Absence Data. *Ecol. Evol.* 6 (15), 5383–5389. doi:10.1002/ece3.2214

Ritz, A., Poirel, C. L., Tegge, A. N., Sharp, N., Simmons, K., Powell, A., et al. (2016). Pathways on Demand: Automated Reconstruction of Human Signaling Networks. *Npj Syst. Biol. Appl.* 2 (1), 1–9. doi:10.1038/npjsba.2016.2

Rodchenkov, I., Babur, O., Luna, A., Aksoy, B. A., Wong, J. V., Fong, D., et al. (2019). Pathway Commons 2019 Update: Integration, Analysis and Exploration of Pathway Data. *Nucleic Acids Res.* 48, 489–497. doi:10.1093/nar/gkz946

Rubel, T., and Ritz, A. (2020). Augmenting Signaling Pathway Reconstructions. *Proc. 11th ACM Int. Conf. Bioinformatics Comput. Biol. Health Inform.* 10, 1–10. doi:10.1145/3388440.3412411

Saito, T., and Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative Than the ROC Plot when Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE* 10 (3), e0118432. doi:10.1371/journal.pone.0118432

Schaefer, M. H., Fontaine, J.-F., Vinayagam, A., Porras, P., Wanker, E. E., and Andrade-Navarro, M. A. (2012). Hippie: Integrating Protein Interaction Networks with experiment Based Quality Scores. *PLoS ONE* 7 (2), e31826. doi:10.1371/journal.pone.0031826

Schaefer, M. H., Serrano, L., and Andrade-Navarro, M. A. (2015). Correcting for the Study Bias Associated with Protein-Protein Interaction Measurements Reveals Differences between Protein Degree Distributions from Different Cancer Types. *Front. Genet.* 6, 260. doi:10.3389/fgene.2015.00260

Schmidt, T., Bergner, A., and Schwede, T. (2014). Modelling Three-Dimensional Protein Structures for Applications in Drug Design. *Drug Discov. Today* 19 (7), 890–897. doi:10.1016/j.drudis.2013.10.027

SeahSen, C. S., Kasim, S., Fudzee, M. F. M., Law Tze Ping, J. M., Mohamad, M. S., Saedudin, R. R., et al. (2017). An Enhanced Topologically Significant Directed Random Walk in Cancer Classification Using Gene Expression Datasets. *Saudi J. Biol. Sci.* 24 (8), 1828–1841. doi:10.1016/j.sjbs.2017.11.024

Segura, J., Sorzano, C. O. S., Cuenca-Alba, J., Aloy, P., and Carazo, J. M. (2015). Using Neighborhood Cohesiveness to Infer Interactions between Protein Domains. *Bioinformatics* 31 (15), 2545–2552. doi:10.1093/bioinformatics/btv188

Sevimoglu, T., and Arga, K. Y. (2014). The Role of Protein Interaction Networks in Systems Biomedicine. *Comput. Struct. Biotechnol. J.* 11 (18), 22–27. doi:10.1016/j.csbj.2014.08.008

Silverbush, D., Cristea, S., Yanovich-Arad, G., Geiger, T., Beerenwinkel, N., and Sharan, R. (2019). Simultaneous Integration of Multi-Omics Data Improves the Identification of Cancer Driver Modules. *Cel Syst.* 8 (5), 456–466.e5. doi:10.1016/j.cels.2019.04.005

Simpson, G. (1966). Notes on the Measurement of Faunal Resemblance. *Am. J. Sci.* 258-A, 300–311. http://earth.geology.yale.edu/~ajs/1960/ajs_258A_11.pdf/300.pdf.

Singh, R., Xu, J., and Berger, B. (2005). Struct2Net: Integrating Structure into Protein-Protein Interaction Prediction. *Pac. Symp. Biocomput* 2006, 403–414. doi:10.1142/9789812701626_0037

Sjölund, J., Manetopoulos, C., Stockhausen, M.-T., and Axelson, H. (2005). The Notch Pathway in Cancer: Differentiation Gone Awry. *Eur. J. Cancer* 41 (17), 2620–2629. doi:10.1016/j.ejca.2005.06.025

Skinnider, M. A., Stacey, R. G., Foster, L. J., and Iakoucheva, L. M. (2018). Genomic Data Integration Systematically Biases Interactome Mapping. *Plos Comput. Biol.* 14, e1006474. doi:10.1371/journal.pcbi.1006474

Sychev, Z. E., Hu, A., DiMaio, T. A., Gitter, A., Camp, N. D., Noble, W. S., et al. (2017). Integrated Systems Biology Analysis of KSHV Latent Infection Reveals Viral Induction and reliance on Peroxisome Mediated Lipid Metabolism. *Plos Pathog.* 13 (3), e1006256. doi:10.1371/journal.ppat.1006256

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-wide Experimental Datasets. *Nucleic Acids Res.* 47 (D1), D607–D613. doi:10.1093/nar/gky1131

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/measurement Sets. *Nucleic Acids Res.* 49 (D1), D605–D612. doi:10.1093/nar/gkaa1074

Tabei, Y., Kotera, M., Sawada, R., and Yamanishi, Y. (2019). Network-based Characterization of Drug-Protein Interaction Signatures with a Space-Efficient Approach. *BMC Syst. Biol.* 13 (S2), 39. doi:10.1186/s12918-019-0691-1

The UniProt Consortium (2019). UniProt: a Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* 47, D506. doi:10.1093/nar/gky1049

Tkačik, G., Walczak, A. M., and Bialek, W. (2009). Optimizing Information Flow in Small Genetic Networks. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* 80 (3), 1–18. doi:10.1103/PhysRevE.80.031920

Trojan, A., Kasprzak, H., Gutierrez, O., Penagos, P., Briceno, I., O. Siachoque, H., et al. (2020). "Neoplastic Brain, Glioblastoma, and Immunotherapy," in *Brain and Spinal Tumors - Primary and Secondary* (The Shard: IntechOpen). doi:10.5772/intechopen.84726

Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S.-S. C., Chayes, J., Borgs, C., et al. (2013). Simultaneous Reconstruction of Multiple Signaling Pathways via the Prize-Collecting Steiner forest Problem. *J. Comput. Biol.* 20 (2), 124–136. doi:10.1089/cmb.2012.0092

Tuncbag, N., Gosline, S. J. C., Kedaigle, A., Soltis, A. R., Gitter, A., and Fraenkel, E. (2016a). Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. *Plos Comput. Biol.* 12 (4), e1004879. doi:10.1371/journal.pcbi.1004879

Tuncbag, N., McCallum, S., Huang, S.-s. C., and Fraenkel, E. (2012). SteinerNet: a Web Server for Integrating 'omic' Data to Discover Hidden Components of Response Pathways. *Nucleic Acids Res.* 40 (W1), W505–W509. doi:10.1093/nar/gks445

Tuncbag, N., Milani, P., Pokorny, J. L., Johnson, H., Sio, T. T., Dalin, S., et al. (2016b). Network Modeling Identifies Patient-specific Pathways in Glioblastoma. *Sci. Rep.* 6 (1), 1–12. doi:10.1038/srep28668

Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M., and Wodak, S. J. (2011). Interaction Databases on the Same page. *Nat. Biotechnol.* 29, 391. doi:10.1038/nbt.1867

Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., et al. (2010). iRefWeb: Interactive Analysis of Consolidated Protein Interaction Data and Their Supporting Evidence. *Database (Oxford)* 2010, baq023. doi:10.1093/database/baq023

Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for Detecting Significantly Mutated Pathways in Cancer. *J. Comput. Biol.* 18 (3), 507–522. doi:10.1089/cmb.2010.0265

Varma Polisetty, R., Gautam, P., Sharma, R., Harsha, H. C., Nair, S. C., Kumar Gupta, M., et al. (2012). LC-MS/MS Analysis of Differentially Expressed Glioblastoma Membrane Proteome Reveals Altered Calcium Signalling and Other Protein Groups of Regulatory Functions Running Title-Glioblastoma Membrane Proteins. Available at: https://www.mcponline.org.

Venko, K., Roy Choudhury, A., and Novič, M. (2017). Computational Approaches for Revealing the Structure of Membrane Transporters: Case Study on Bilitranslocase. *Comput. Struct. Biotechnol. J.* 15, 232–242. doi:10.1016/j.csbj.2017.01.008

Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011). Interactome Networks and Human Disease. *Cell* 144 (6), 986–998. doi:10.1016/j.cell.2011.02.016

Vitali, F., Marini, S., Pala, D., Demartini, A., Montoli, S., Zambelli, A., et al. (2018). Patient Similarity by Joint Matrix Trifactorization to Identify Subgroups in Acute Myeloid Leukemia. *JAMIA Open* 1 (1), 75–86. doi:10.1093/jamiaopen/ooy008

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., et al. (2004). STRING: Known and Predicted Protein-Protein Associations, Integrated and Transferred across Organisms. *Nucleic Acids Res.* 33 (DATABASE ISS.), D433–D437. doi:10.1093/nar/gki005

Waks, Z., Weissbrod, O., Carmeli, B., Norel, R., Utro, F., and Goldschmidt, Y. (2016). Driver Gene Classification Reveals a Substantial Overrepresentation of Tumor Suppressors Among Very Large Chromatin-Regulating Proteins. *Sci. Rep.* 6 (1), 1–12. doi:10.1038/srep38988

Wang, Y., Yang, Y., Chen, S., and Wang, J. (2021). DeepDRK: a Deep Learning Framework for Drug Repurposing through Kernel-Based Multi-Omics Integration. *Brief. Bioinform.* 00 (August 2020), 1–10. doi:10.1093/bib/bbab048

Yerneni, S., Khan, I. K., Wei, Q., and Kihara, D. (2018). IAS: Interaction Specific GO Term Associations for Predicting Protein-Protein Interaction Networks. *Ieee/acm Trans. Comput. Biol. Bioinf.* 15 (4), 1247–1258. doi:10.1109/tcbb.2015.2476809

Žitnik, M., Janjić, V., Larminie, C., Zupan, B., and Pržulj, N. (2013). Discovering Disease-Disease Associations by Fusing Systems-Level Molecular Data. *Scientific Rep.* 3 (1), 1–9. doi:10.1038/srep03202

Zsákai, L., Sipos, A., Dobos, J., Erős, D., Szántai-Kis, C., Bánhegyi, P., et al. (2019). Targeted Drug Combination Therapy Design Based on Driver Genes. *Oncotarget* 10 (51), 5255–5266. doi:10.18632/oncotarget.26985

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.