Supplementary Material for
Reconciling Signaling Pathway Databases with Network Topologies

Tobias Rubel*, Pramesh Singh*, and Anna Ritz†

Biology Department, Reed College, Portland, Oregon, USA
**Equal author contribution*
†*E-mail: aritz@reed.edu*

S1. Data Collection and Processing

We converted every pathway into an undirected graph by parsing Simple Interaction Format (SIF) files. These files were pulled directly from PathwayCommons or were converted from BioPAX format using PaxTools. We only considered interactions that involved proteins and required pathways to contain at least ten undirected edges. Many metabolic networks, for example, were ignored due to this requirement. We mapped all proteins into HGNC namespace using the HGNC mapper (`https://www.genenames.org/download/custom/`).

Two databases, KEGG and SIGNOR, capture protein families and protein complexes in their networks. For these databases, we parsed a collapsed version which includes complexes and families as nodes in the network and an expanded version that converts such entities into their constitutive proteins. Interactions that include families or proteins were expanded to add an edge for every protein member (e.g., a two-protein family connected to a three-protein complex added six undirected edges). Further, protein complexes were connected in an "all-vs-all" manner to indicate physical interaction (e.g., a three-protein complex added three undirected edges). As expected, the average number of nodes and edges is larger for the expanded versions of the KEGG and SIGNOR databases (Fig. S1). In total, we considered $1,592$ pathways in nine datasets that captured pathways from seven distinct databases.

S2. Agglomerative Clustering

We cluster the vector representations of the pathways (either 30-dimensional graphlet vectors or 12-dimensional GHuST vectors). We perform agglomerative clustering with a mean linkage criterion and a cosine distance metric. Clustering quality is quantified using adjusted mutual information (AMI), which adjusts for random chance. Given a partition X determined by the agglomerative clustering and correct labels Y (here, "pathway" or "walker"), the AMI is defined as

$$AMI(X,Y) = \frac{MI(X,Y) - E[MI(X,Y)]}{\text{avg}(H(X),H(Y)) - E[MI(X,Y)]}, \tag{1}$$

where $MI(X,Y)$ is the mutual information of the partitions, $E[MI(X,Y)]$ is the expectation of the mutual information of two partitions based on a hypergeometric model of randomness, and $H(X)$ is the entropy of X. A larger AMI indicates that the partitions are more similar, and hence X better reflects the correct labels Y. We calculate the AMI for every possible number of clusters admitted by the agglomerative clustering algorithm.

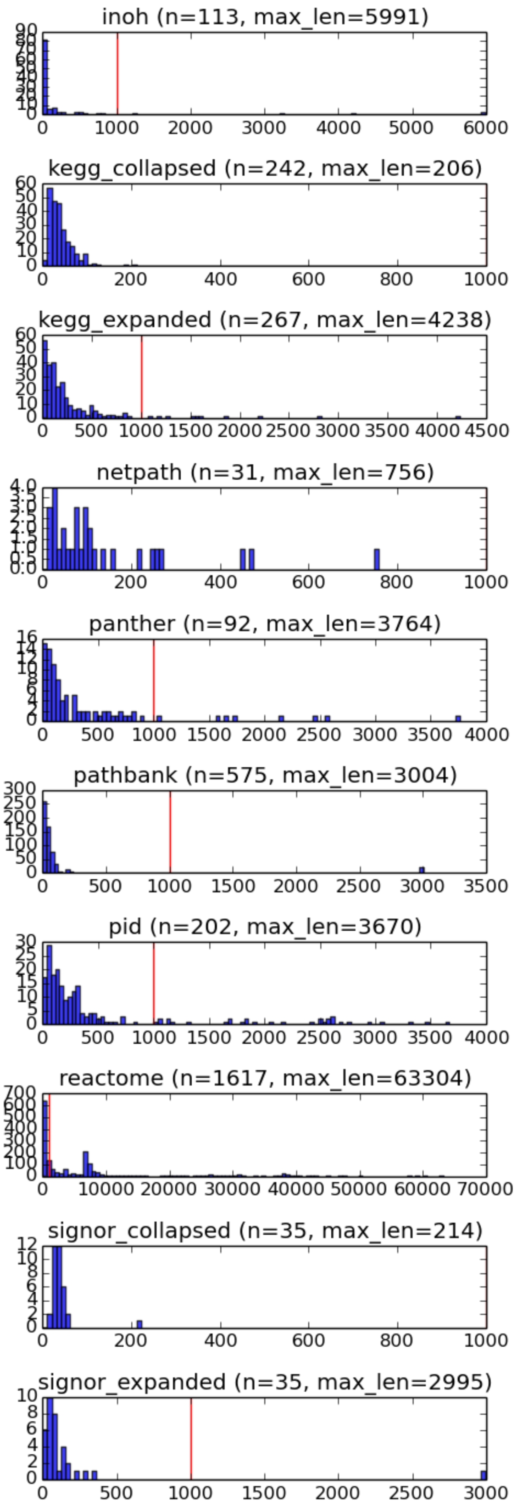

Fig. S1. Histograms of pathway sizes (by number of interactions) for all datasets considered. Vertical red bar indicates clusters with 1,000 interactions. Reactome pathways are about two orders of magnitude larger than other pathway databases. Additionally, the "-expanded" pathways are larger than the "-collapsed" pathways, as expected.

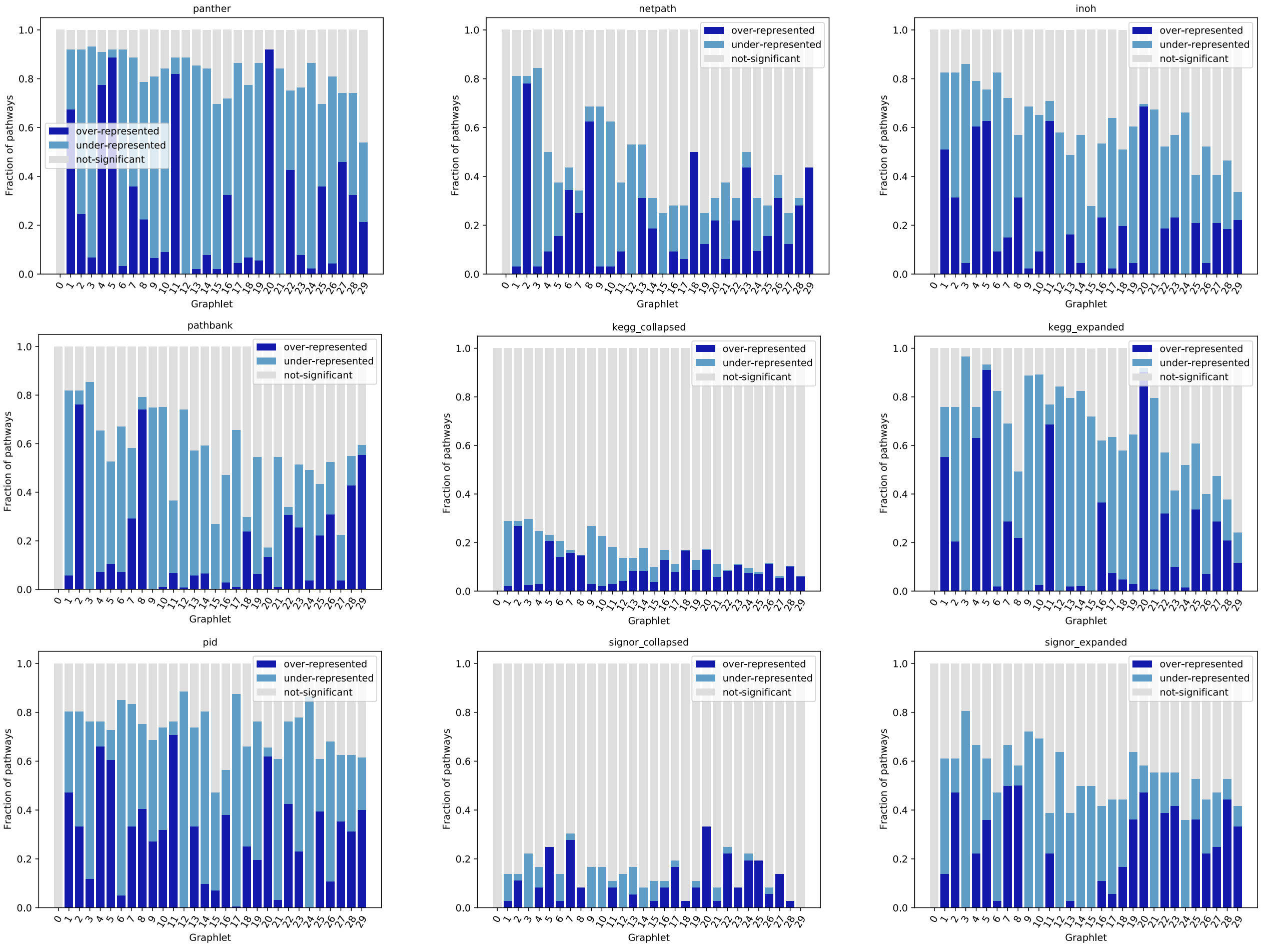

Fig. S2. Over- and under-represented dimensions of the 30 graphlet counts for all datasets. The Panther pathway database is shown in the main paper. Note that the "-collapsed" versions of KEGG and SIGNOR, which include protein complexes and families as nodes, have many fewer over- or under-represented dimensions.

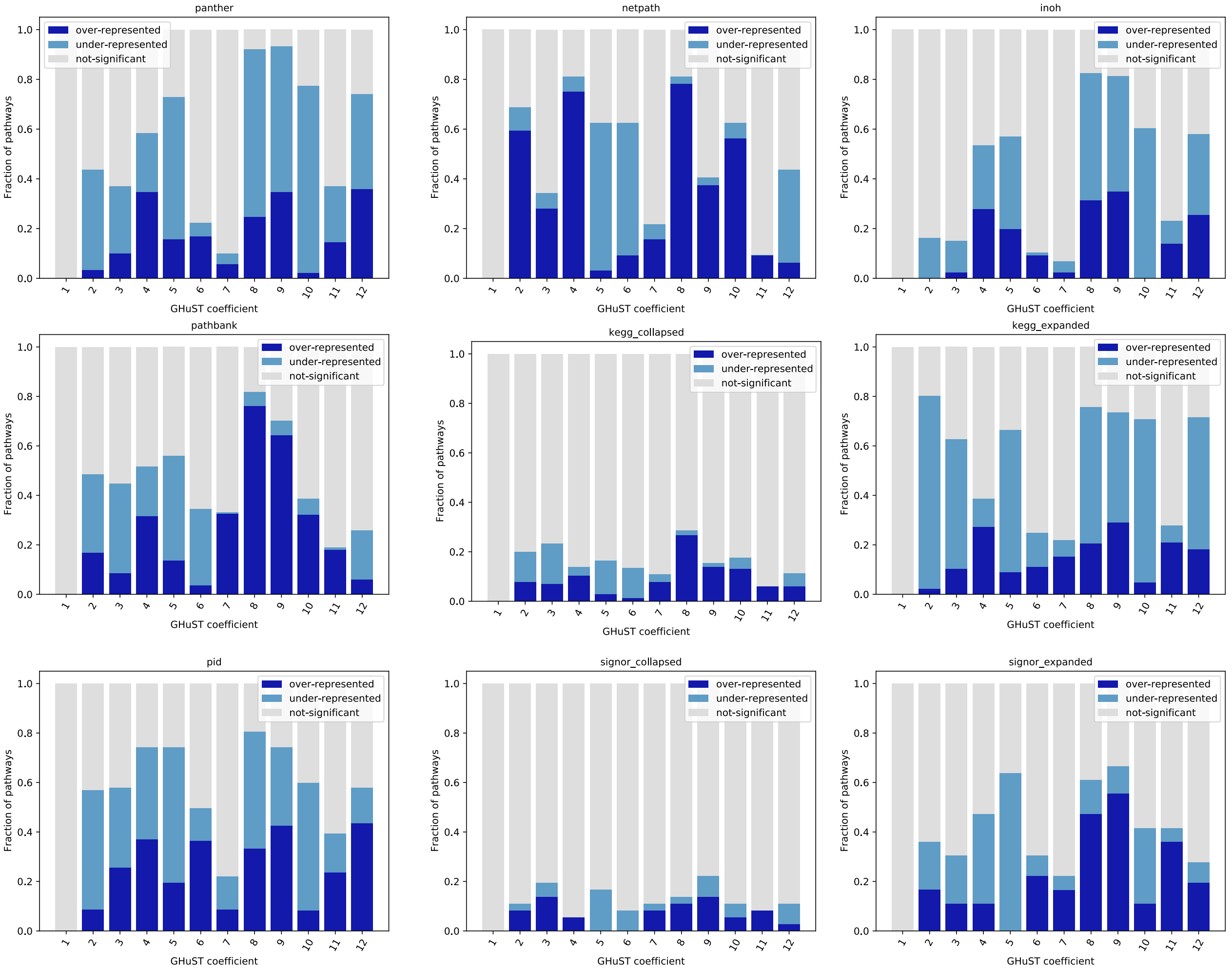

Fig. S3. Over- and under-represented dimensions of the 30 GHuST (ρ) coefficients for all datasets. The Panther pathway database is shown in the main paper. Note that the “-collapsed” versions of KEGG and SIGNOR, which include protein complexes and families as nodes, have many fewer over- or under-represented dimensions.

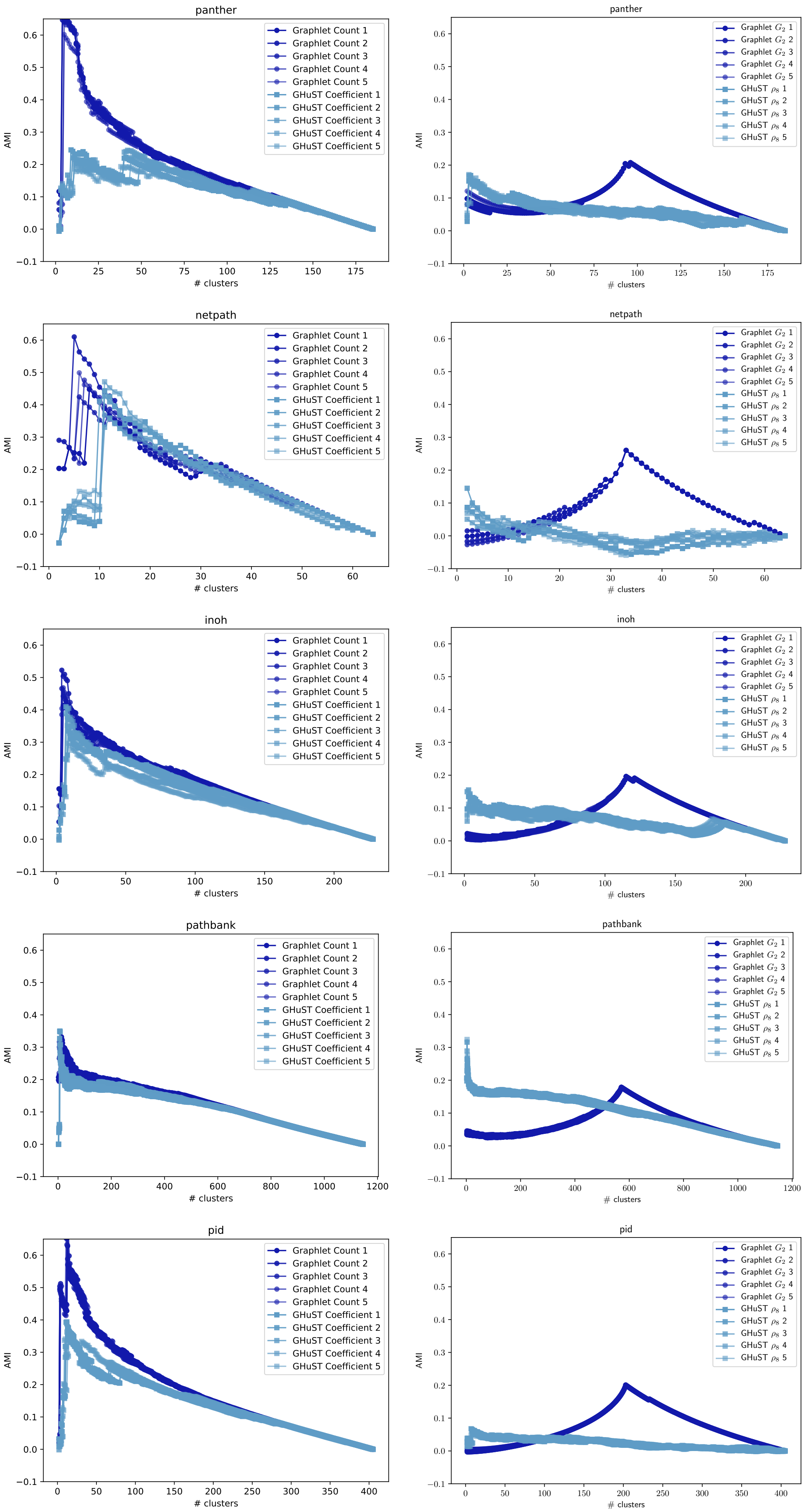

Fig. S4. AMI plots for full vectors (left) and triangles only (right) for pathway databases.

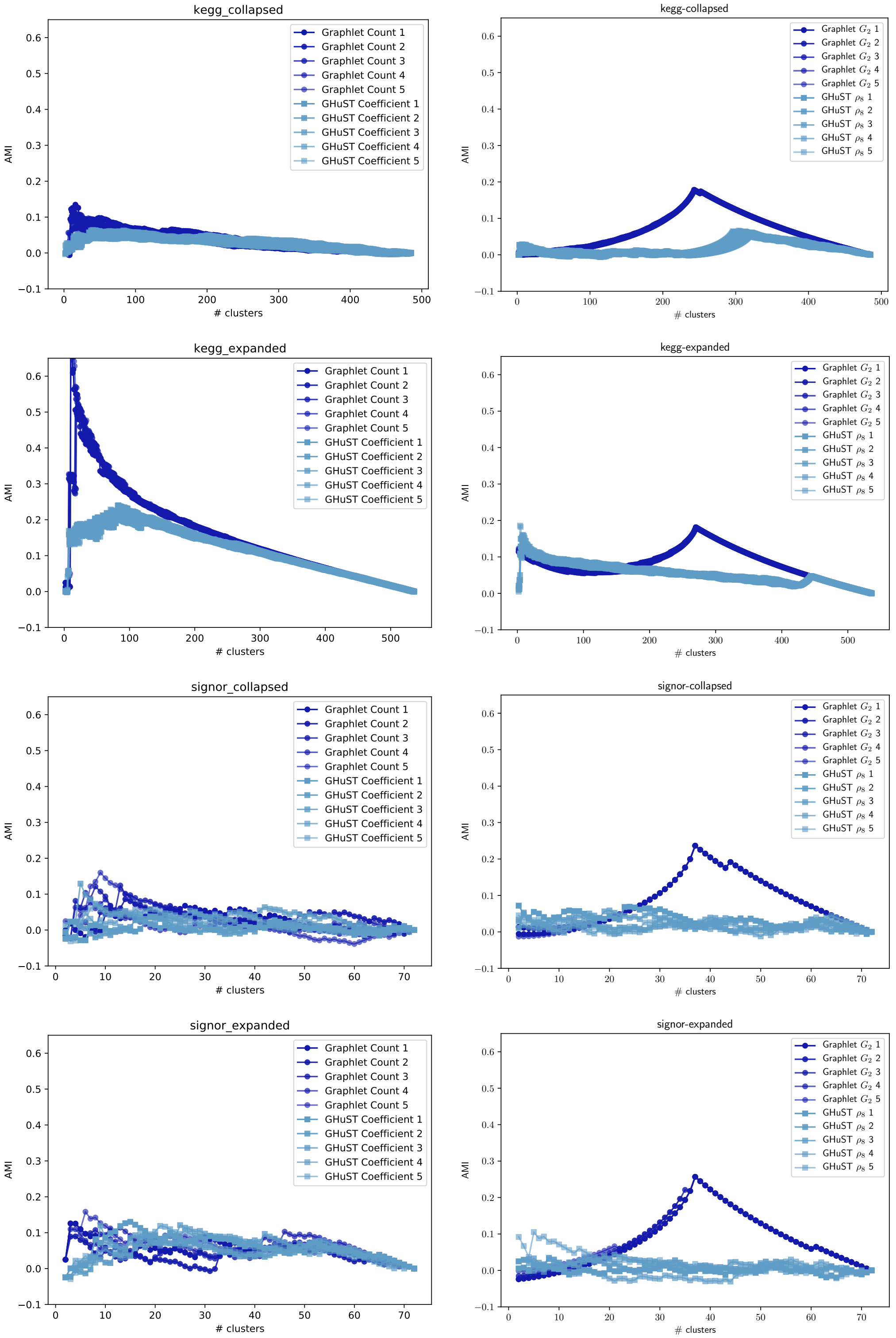

Fig. S5. AMI plots for full vectors (left) and triangles only (right) for pathway databases.

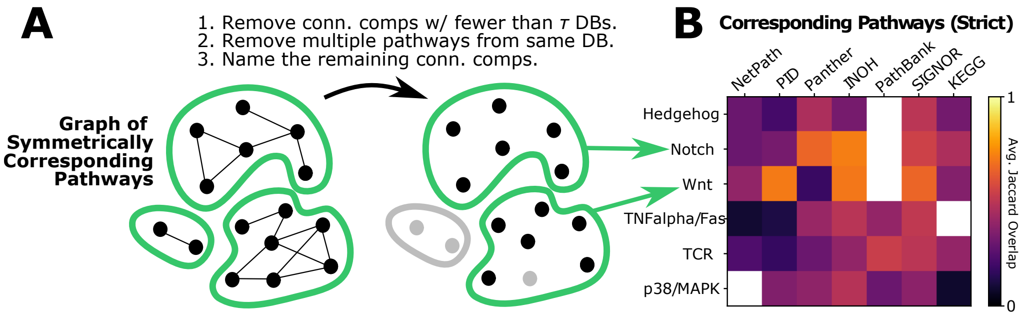
A
1. Remove conn. comps w/ fewer than τ DBs.
2. Remove multiple pathways from same DB.
3. Name the remaining conn. comps.
Graph of Symmetrically Corresponding Pathways
B
Corresponding Pathways (Strict)
NetPath
PID
Panther
INOH
PathBank
SIGNOR
KEGG
Hedgehog
Notch
Wnt
TNFalpha/Fas
TCR
p38/MAPK
1
Avg. Jaccard Overlap
0

Fig. S6. (A) Identifying corresponding pathways. We first build a graph where the nodes are pathways and two edges denote symmetric correspondence. Then, we find the connected components that contain at least τ different databases and ensure at most one pathway per database. Finally, we name each connected component based on the pathway names. (B) Each connected component is represented as a row in the matrix, which describes the average Jaccard overlap of each pathway across databases. White entries denote databases with no corresponding pathway.

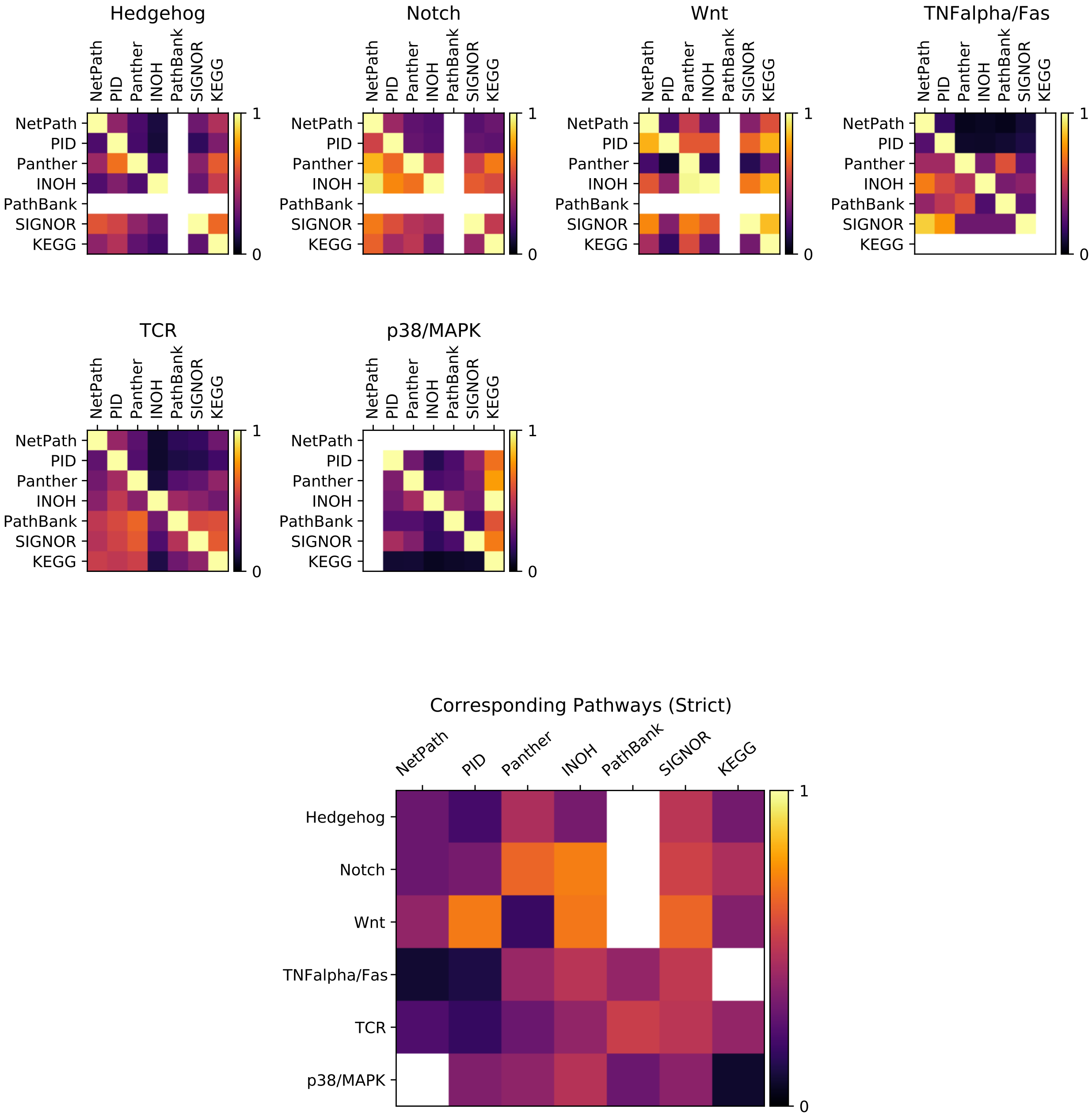

Fig. S7. Asymmetric Jaccard overlap of nodes across databases for each pathway (top) and averaged (bottom) when $\tau = 6$. White entries denote databases that do not have corresponding pathways. In the bottom figure, all non-identity and non-missing entries of the rows are averaged.

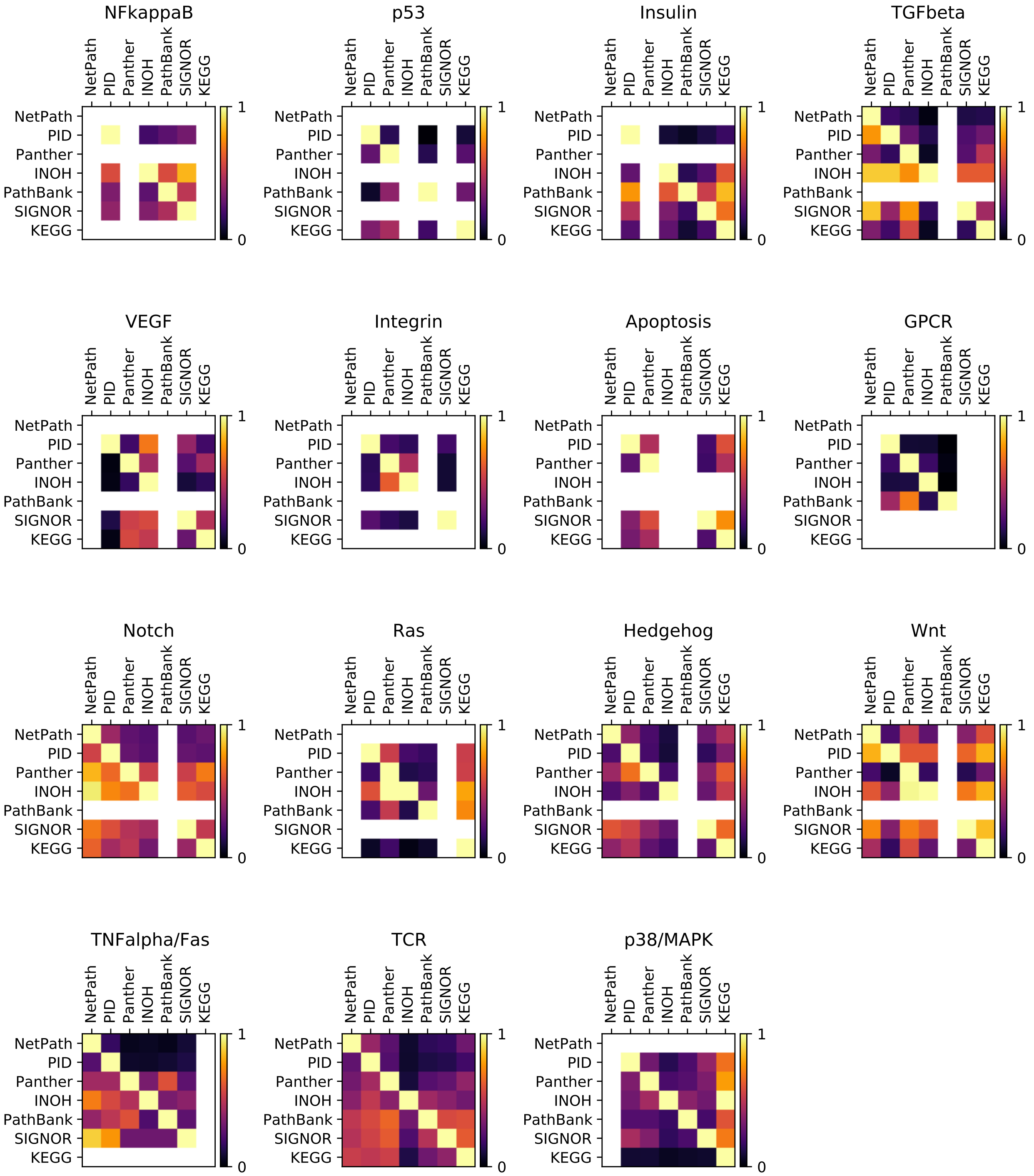

Fig. S8. Asymmetric Jaccard overlap of nodes across databases for each pathway when $\tau = 4$. White entries denote databases that do not have corresponding pathways.

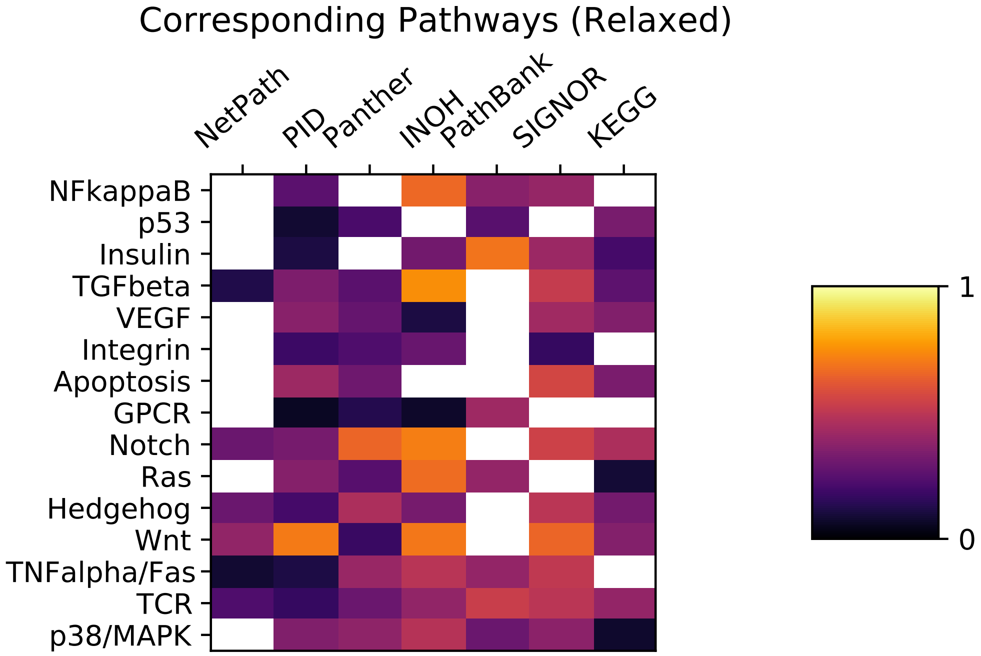

Fig. S9. Asymmetric Jaccard overlap of nodes averaged across databases for each pathway in the relaxed scenario (entries from Fig. S8). White entries denote databases that do not have corresponding pathways. All non-identity and non-missing entries of the rows are averaged.

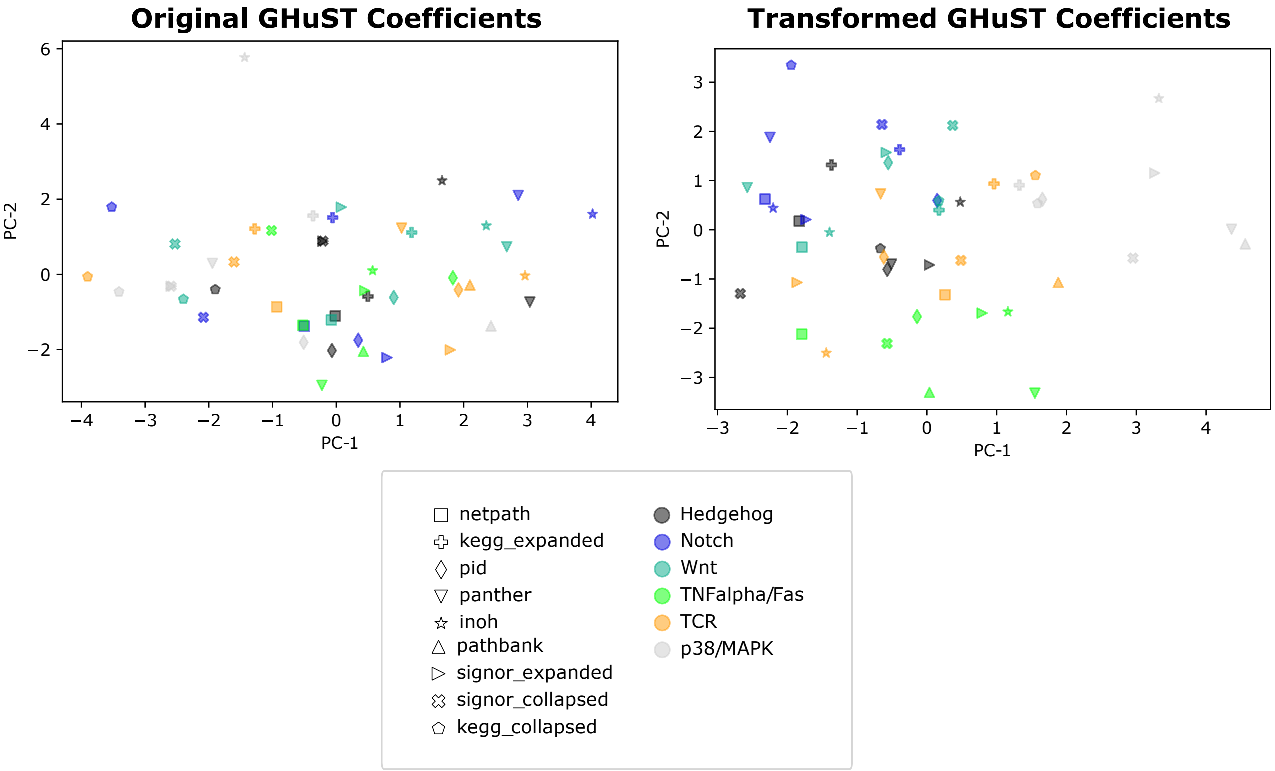

Fig. S10. Principal component analysis (PCA) of the first two components for the original GHuST coefficients (left) and the transformed coefficients (right) for all nine datasets. Datasets are denoted by marker shape, pathways are denoted by colors.

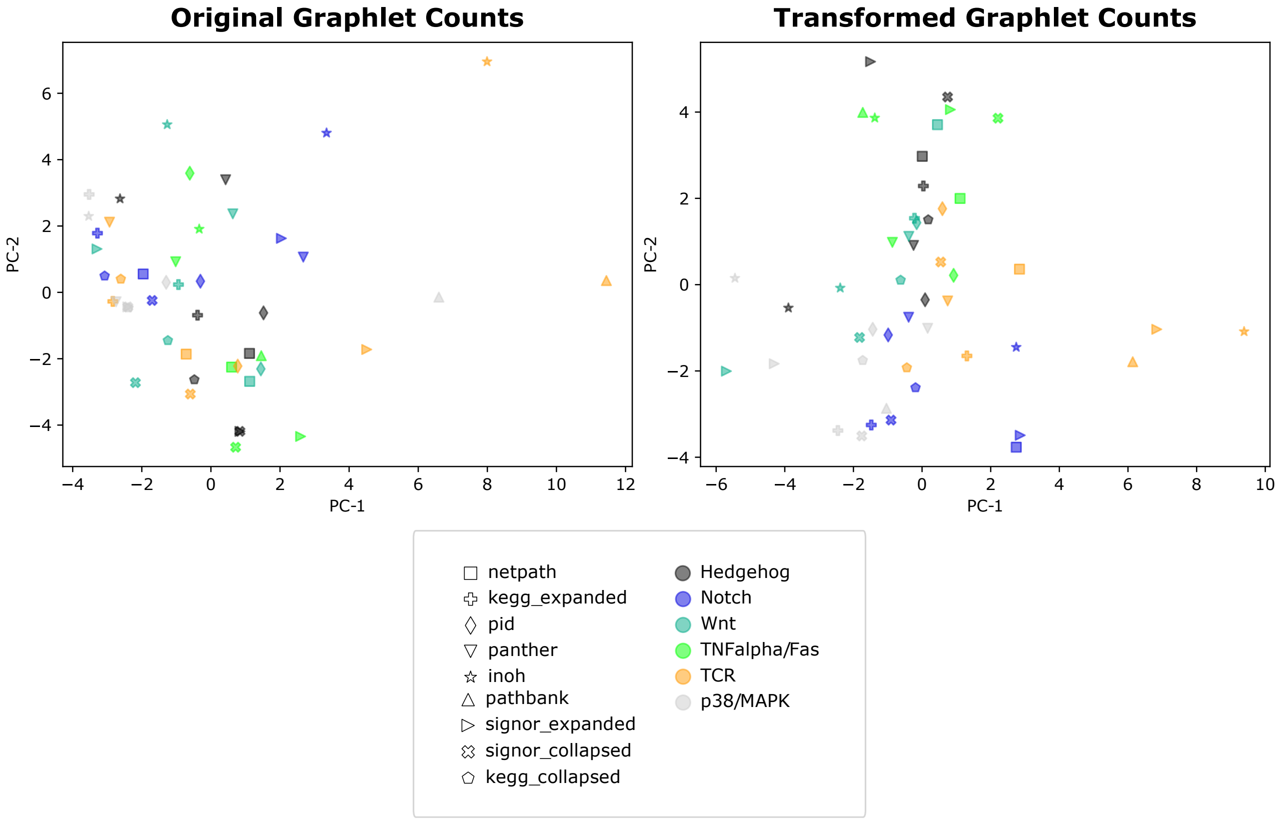

Fig. S11. Principal component analysis (PCA) of the first two components for the original graphlet counts (left) and the transformed coefficients (right) for all nine datasets. Datasets are denoted by marker shape, pathways are denoted by colors.

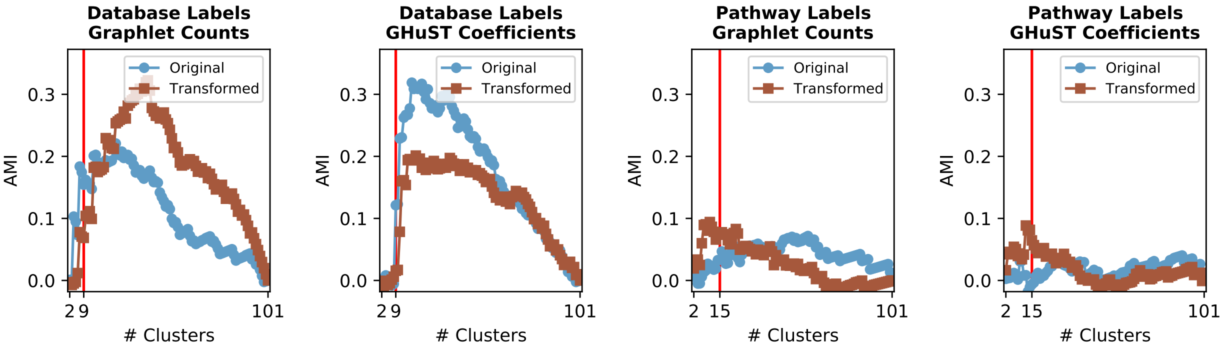

Fig. S12. AMI of graphlet counts and GHuST coefficients when clustering corresponding pathways ($\tau = 4$) using databases as ground truth labels (first two plots) or pathways as ground truth labels (last two plots). Blue lines indicate clustering by original values; brown lines indicate clustering by regression-transformed coordinates. Vertical red line indicates the correct number of clusters for each ground truth dataset.