OXFORD

Briefings in Bioinformatics, 2025, **26(4)**, bbaf320
https://doi.org/10.1093/bib/bbaf320
Case Study

Advancing edge-based clustering and graph embedding for biological network analysis: a case study in RASopathies

Federico García-Criado [1,‡], Pedro Seoane [1,2,3,*,‡], Elena Rojano [1,2,3], Juan AG Ranea [1,2,3,4,§], James R Perkins [1,2,3,§]

[1]Department of Molecular Biology and Biochemistry, University of Malaga, 29010 Malaga, Spain
[2]Center for Biomedical Network Research on Rare Diseases (CIBERER), Instituto de Salud Carlos III (ISCIII), 28029 Madrid, Spain
[3]Institute of Biomedical Research in Malaga and platform of nanomedicine, IBIMA Plataforma BIONAND, 29071 Malaga, Spain
[4]Spanish National Bioinformatics Institute (INB/ELIXIR-ES), Instituto de Salud Carlos III (ISCIII), 28020 Madrid, Spain

*Corresponding author. E-mail: seoanezonjic@uma.es
‡Federico García-Criado and Pedro Seoane contributed equally to this work.
§Juan AG Ranea and James R Perkins contributed equally to this work and share last authorship.

Abstract

Understanding and predicting biological processes from protein–protein interaction (PPI) networks requires accurate and efficient representations of their structure. However, many existing methods fail to capture the complex, overlapping modular structure of biological systems. To address this, we propose a network embedding strategy that improves both biological interpretability and predictive power. By transforming networks into a low-dimensional space while preserving key topological properties, embedding enables the discovery of novel functional relationships. Pre-clustering a network before embedding enhances representation quality, i.e. the ability to preserve meaningful structural and functional properties in the embedding space. However, traditional non-overlapping clustering methods can introduce bias by ignoring the overlapping nature of biological communities. We overcome this limitation by integrating the Hierarchical Link Clustering (HLC) algorithm into an embedding workflow tailored for large, weighted, undirected networks. First, we introduce two optimized HLC implementations for Python and R, both outperforming existing methods in clustering accuracy and scalability. Then, by restricting random walks to HLC-defined communities, we improve the representation of biological pathways, as shown using Reactome on the human PPI network. We also apply our full cluster embedding workflow to analyze RASopathies, a group of interrelated disorders with a diverse range of phenotypes, caused by mutations in genes from the RAS/MAPK pathway. This approach was used not only to represent known pathways, but also to identify potential novel gene candidates associated with RASopathies, including Noonan and Costello syndrome. HLC implementations are available in the CDLIB library (https://github.com/GiulioRossetti/cdlib), and at https://github.com/jimrperkins/linkcomm for Python and R, respectively.

Keywords: RASopathies; network embedding; overlap community; HLC; protein-protein interaction

Introduction

Network analysis is fundamental to bioinformatics and systems biology, where these techniques have been applied to molecular interactions (e.g. protein networks [1], metabolites [2]) and phenotypic traits such as comorbidities [3, 4]. In recent years, network embedding has emerged as a powerful addition to these approaches, allowing the transformation of nodes into low-dimensional vector spaces that preserve topological and functional relationships [5]. These embeddings facilitate tasks such as module detection, gene prioritization, and disease classification by revealing hidden patterns and improving the integration with machine learning models [6].

Methods for embedding include matrix factorization methods such as Laplacian eigenmap [7], Singular Value Decomposition [8], and network kernels [9, 10]. In addition, powerful and pioneering methods for network embedding based on deep-learning have emerged in recent years. These include random walk-based methods like DeepWalk [11] and Node2vec [5], which have been applied in diverse tasks, including link prediction [12, 13], node classification [6, 14], community detection [15–17] and single-cell analysis [18, 19].

In these random walk-based approaches, custom random walks are integrated into a word2vec framework to generate embeddings [5, 11, 20]. Various algorithms have refined this scheme, such as node2vec, which balances breadth-first and depth-first sampling to enhance representation quality [5]. More recently, these methods have been improved by incorporating classical clustering techniques, making them community-aware and showing promising results in embedding biological networks [21–24].

Choosing an appropriate clustering method is crucial when integrating classical network techniques with network

Received: March 19, 2025. **Revised:** May 20, 2025. **Accepted:** June 16, 2025

embeddings and other deep learning approaches. Most methods use non-overlapping clustering, assigning each node to a single community. However, in biological networks, genes and proteins often belong to multiple pathways or functional modules [25], making overlapping clustering methods essential [26]. To address these challenges, we propose a novel embedding pipeline that integrates overlapping edge-based clustering with random walk–based node embedding. Specifically, we incorporate Hierarchical Link Clustering (HLC) [25, 27] to identify overlapping communities, and we constrain random walks to remain within these communities. This strategy enhances the functional coherence of the embeddings and improves biological pathway representation.

Recent work has explored alternative strategies to improve network representations, preserving properties such as structural roles [28], high-order node proximity [29], and integrating global structural information [30], each capturing complementary aspects of network complexity. In contrast, our approach prioritizes the preservation of overlapping community structure, a defining characteristic of biological networks, by incorporating an edge similarity–based clustering step i.e. computationally efficient and scalable to large, weighted networks. This allows the integration of biologically meaningful community information into the embedding process without incurring substantial computational overhead.

Despite these advantages, current HLC implementations face limitations. The widely used R package linkcomm [31] supports weighted networks but struggles with large datasets, while the Python package CDLIB [32] offers only an unweighted implementation, limiting its applicability to weighted biological networks.

Network embedding can be applied to analyze genes related to diseases, such as RASopathies. These developmental disorders are caused by germline mutations in genes of the RAS/MAPK signaling pathway [33], involved in multiple processes including angiogenesis [34] and cancer [35]. Despite significant advances in genetic testing, the full mechanisms driving these conditions and the relationships among the various disorders remain unclear. Further investigation into how these genes interact and regulate the pathway may provide crucial insights into their overlapping yet distinct clinical features, as well as providing potential novel drug targets.

Applying graph embedding and overlapping clustering to study RASopathies highlights its potential to uncover crucial insights into gene interactions and pathway regulation. This not only helps explain their overlapping yet distinct clinical features but also suggests novel drug targets. Given the molecular links between RASopathies and cancer, our approach may further support the repurposing of cancer therapies for these disorders, offering new avenues for targeted treatment.

The main contributions of this work are as follows:

- We present a new node embedding framework which integrates community information obtained from HLC.
- We optimize the HLC algorithm for large-scale, weighted biological networks, providing efficient Python and R implementations.
- We demonstrate that the new embedding enhances the representation of known biological pathways in the human interactome, outperforming multiple baseline node embedding methods.
- We apply the method to RASopathies, providing novel representation of known genes and revealing novel candidate genes associated with the disorders.

Materials and methods

Implementation of HLC

We developed new implementations for analyzing weighted undirected networks for both R and Python. They extend original methods, which also apply to unweighted networks.

In R, the original linkcomm clustering method (LKCM) has significant memory limitations when analyzing large networks. We addressed this issue by utilizing sparse matrices, resulting in the implementation referred to as linkcomm sparse (LKCMsp).

We modified the linkcomm implementation to calculate Tanimoto coefficients following formulations proposed by Ahn et al. [25]. In our approach, the Tanimoto coefficient, $S(e_{ik}, e_{jk})$, is computed as shown in Equation (1).

$$S(e_{ik}, e_{jk}) = \frac{a_i \cdot a_j}{a_i \cdot a_i + a_j \cdot a_j - a_i \cdot a_j}. \quad (1)$$

where e_{ik} represents the edge between nodes i and k. a_i or a_j define the i and j rows of the adjacency matrix, known as adjacency vectors, respectively. In the original implementation [25], each ith position of a_i assigns the average degree of the ith node, whereas in linkcomm this value is set to 1. We refer to these formulations as the linkcomm and the Ahn formulation, respectively.

In Python, CDLIB includes a weighted HLC implementation derived from a repository related to the Ahn formulation; however, the authors have not fully adapted it to their library interface. In this study, we addressed this gap. The HLC algorithm is optimized for low-density networks and represents adjacency using a dictionary data structure.

We propose an alternative, HLC NumPy (HLCnp), that uses NumPy array representation and SciPy operations to vectorize the Tanimoto coefficient calculation, which is developed within the CDLIB library. This strategy is particularly efficient for dense networks in terms of execution time, although it may not optimize memory usage.

Moreover, using SciPy operations in hierarchical clustering allows for linkage methods beyond single linkage. An intermediate filtering step can also be added to eliminate spurious edge similarities, i.e. those with values below a given threshold.

All different symbols and acronyms used in this study are presented in Table 1.

Benchmarking of HLC: measuring technical performance and accuracy

We used eight different weighted networks, ranging from 100 to 12 000 nodes, with the number of edges varying from 2000 to 472 000 (Supplementary Fig. S1). Most of these networks are biological in nature and are also commonly used in bioinformatics, including protein–protein interaction (PPI) and semantic similarity networks, as described in the Supplementary Section.

For accuracy, we applied the metrics proposed in the Ahn study [25], specifically community and overlap coverage. Community coverage quantifies the total number of nodes in the network that belong to at least one non-trivial cluster (at least three nodes) while overlap coverage is the average number of non-trivial communities to which a node belongs.

For technical performance, we measured the percentage of CPU usage, total runtime, and maximum memory usage using the GNU/Linux time command. Each run was conducted on a Lenovo SR645 node equipped with 156 cores and 512 GB of physical memory.

Table 1. Glossary of terms and definitions used in the study

Term	Definition
Formulation	
Ahn	Tanimoto formula proposed by Ahn and collaborators [25] to compute the edge similarity.
Linkcomm	Version of computation by Kalinka and collaborators [31] to compute edge similarity.
HLC implementations	
HLC	Original python CDLIB implementation
HLCnp	Python CDLIB implementation based in numpy array operations
LKCM	Original linkcomm HLC implementation
LKCMsp	Linkcomm HLC implementation with sparse matrices operations
Networks	
huST700	Interactome network of proteins in *Homo sapiens* with a cutoff of 700
huST900	Interactome network of proteins in *Homo sapiens* with a cutoff of 900
ecST700	Interactome network of proteins in *Escherichia coli* with a cutoff of 700
ecST900	Interactome network of proteins in *Escherichia coli* with a cutoff of 900
coORPHA	Comorbidity network with data extracted from ORPHANET
coOMIM	Comorbidity network with data extracted from OMIM
coMERGED	Comorbidity network with data merged from ORPHANET and OMIM
Wikipedia	Interaction network from Wikipedia

For the execution of the different algorithms, the following parameters were used: HLC with default parameter values; HLCnp with *hcmethod* = *single* and *min_edges* = 2; LKCMsp with *hcmethod* = *single*, *diagnorm* = *TRUE*, *minedges* = 3; and LKCM with *edglim* = 10^8 and *removetrivial* = *TRUE*. All algorithms were configured to remove clusters with fewer than two edges, except for HLC, which required the removal of trivial clusters in subsequent steps to achieve comparable results. For LKCM, the *edglim* was set to a higher value than the number of edges in any network used in the study (10^8) to minimize the number of disk write operations.

Network embedding using implemented HLC

Our network embedding system builds upon previous approaches, such as node2vec [5], which consists of an initial generation of random walks followed by the application of a skip-gram model to generate node embeddings. For the first step, we build on the work of Zhang and collaborators with the CRARE algorithm [21], modifying their approach to develop a custom random walk model that incorporates information about overlapping communities. In CRARE, walks are generated using a random walk process that allows jumps not only between neighboring nodes but also between nodes belonging to the same community or sharing the same structural properties, such as nodes with the same degree.

This implementation has been incorporated into and made available through NetAnalyzer [10], a Python library developed by our group for advanced network analysis and embedding.

In our implementation, we made the following changes: (i) we removed the jumps in the random walk between nodes with the same structural property, (ii) the method was generalized to use any clustering algorithm and adapted for overlapping clustering, and (iii) the transition probability from one node to a node in another community is weighted by the inverse of the community size. Our random walk model can then be formulated as follows:

Given n_0 as the source node and n_t as the visited node at step t, we defined a flexible random walk strategy using the following transition probability:

$$P(n_t = y \mid n_{t-1} = x) = \begin{cases} \Pi_{y,x} & \text{if } (x,y) \in E \text{ or } y \in C_x, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

where E is the edge set, C_x represents the union of all communities to which x belongs, and $\Pi_{y,x}$ is the combined probability defined in Equation (3).

$$\begin{aligned} \Pi_{y,x} &= \frac{1}{2}\left(\frac{w_{neigh}}{w_{neigh}+w_{comm}}\right) p_{neigh}(y \mid x) \\ &\quad + \frac{1}{2}\left(\frac{w_{comm}}{w_{neigh}+w_{comm}}\right) p_{comm}(y \mid x) \\ &= (1-q)p_{neigh}(y \mid x) + q p_{comm}(y \mid x), \quad q \in [0,1]. \end{aligned} \quad (3)$$

Here, w_{neigh} and w_{comm} are parameters that define the relevance of being part of the same neighbor or the same community, respectively. Unlike classical methods such as DeepWalk [11] and node2vec [5], which rely on uniform or biased random walks over direct neighbors, our model introduces a dual mechanism that considers both structural proximity (via neighbors) and functional context (via community membership). This formulation enables the embedding to capture overlapping modular structures, which are common in biological networks, while maintaining local connectivity. Additionally, by adjusting the parameter q, the walker can be tuned to favor intra-community transitions, something that traditional approaches cannot directly control. The probabilities p_{neigh} and p_{comm} are simply defined as uniform distributions (as defined in Equation (4) and Equation (5)).

$$p_{neigh}(y \mid x) = \begin{cases} \frac{1}{\text{degree}(x)} & \text{if } (x,y) \in E, \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

$$p_{comm}(y \mid x) = \begin{cases} \frac{1}{|C_x|} & \text{if } y \in C_x, \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

For each node, ten random walks with a length of 100 are generated to be used as corpus for Word2vec (vector_size=128, window=10, epochs=5, sg=1, hs=1, min_count=0), a popular skip-gram model [36]. Finally, the parameters of the model are optimized by the stochastic gradient descent [37].

Network embedding benchmark

The network embedding implementation was applied to the human PPI network from human STRING (huST) using two different score cutoffs, 700 and 900, producing the networks huST700 and huST900, respectively [1]. We used the combined score from STRING, which integrate the information from different channels in this database.

During the network embedding process, we used the communities identified by HLC in each PPI network. As a reference, we also included Louvain clustering, since it is used by the CRARE algorithm. We explored different parameter settings: baseline (random selection of edges, with parameters: $w_{neigh} = 0$, $w_{comm} = 0$); neigh (random walk between connected nodes, with

parameters: $w_{neigh} = 1, w_{comm} = 0$); HLC_comm and Louvain_comm (random selection of nodes inside the same community, with parameters: $w_{neigh} = 0, w_{comm} = 1$); and neigh-HLC_comm and neigh-Louvain_comm (random walk combining both neigh and comm strategies, with parameters: $w_{neigh} = 1, w_{comm} = 1$).

In addition to our custom embedding strategy, we incorporated a diverse set of existing network embedding methods for comparative analysis. These included *node2vec* [5], *DeepWalk* [11], *Role2Vec* [28], *NodeSketch* [29], *GloVe* [38], *GraRep* [30], *GGVec*, *NetMF* [39], *HOPE* [40] and *NMFADMM* [41]. For each of these methods, we used the default parameter settings provided by the karateclub [42] and nodevectors Python libraries.

To assess the improvement in embedding, we retrieved gene annotations from the Reactome database (https://reactome.org/download/current/Ensembl2Reactome.txt) [43]. We then evaluated the normalized distance between genes within the same Reactome pathways in the embedding space, comparing results across the different parameter settings. The normalized distances were computed starting from the euclidean distances in the embedding space is extracted. Then, all pair or nodes are z-normalized and the median of the normalized distance inside every group of nodes is obtained. Additionally, we computed the size and internal edge density for each pathway. Statistical comparisons between distributions across methods were conducted using t-tests with *P*-values adjusted using the Benjamini–Hochberg (BH) procedure.

Finally, we used the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) learning technique to visually represent the embedding spaces generated for each parameter setting. UMAP was computed using the Python umap library (https://pypi.org/project/umap-learn) with the following parameters: $n\_neighbors = 15, min\_dist = 0.1, n\_components = 2, metric =' euclidean'$.

Using the network embedding to improve our understanding of RASopathy related genes

We applied our methodology to RASopathy related genes. To obtain a list of RASopathies and associated genes, we used a semi-automated approach using the MONDO (Human Disease Ontology) and Monarch databases. The methodology was as follows:

1. Identification of Rasopathy Terms in MONDO: We began by capturing the child terms related to the term "RASopathy" (MONDO:0021060) in the MONDO ontology.
2. Manual Grouping of Disorders: After retrieving the relevant terms, we manually grouped disorders that are considered to represent the same underlying condition, based on expert curation, e.g. Noonan syndrome 1 and Noonan syndrome 2 were both considered Noonan Syndrome. Full details of the grouping are given in Supplementary Table S1.
3. Gene Association Using Monarch: For each identified RASopathy child term, we used the Monarch database (https://monarchinitiative.org/) to associate the relevant genes with the respective disorders.

We used the network embedding obtained for the human PPI networks, huST700 and huST900 to further investigate the RASopathy related genes. This was done firstly by visualizing the RASopathy genes using the UMAP representation of the network and showing each of the RASopathy related genes in different colors, corresponding to the RASopathy grouping.

Then, to perform more formal analysis, we performed prioritization analysis, to find potential new RASopathy related genes.

To prioritize genes for each RASopathy, we applied a Bayesian integration approach. Rather than ranking genes solely based on their average distance in the embedding space, we estimated the probability that each gene is associated with a causal gene of the disease. Genes with higher probabilities were given higher priority in the ranking, as described in the next Equation (6).

$$p_{\text{disease}}(gene_i) = 1 - \prod_{gene \in \text{disease}} (1 - p_{gene}(gene_i)) \quad (6)$$

where $p_{\text{disease}}(gene_i)$ represents the probability of gene i being associated with the set of genes related to the disease, and $p_{gene}(gene_i)$ denotes the probability of association between gene i and a particular gene.

To compute the probabilities of association, we trained a logistic regression model using the cosine similarity between gene embeddings as scores, pair of nodes from edges in the PPI network were considered positive labels, and the remaining cases were randomly selected as negatives labels. All the edges and the equivalent number of negative where used during the training step.

Results

Comparison of community detection across HLC implementations

To illustrate these differences, we analyzed results from smaller graphs. In Fig. 1A, we compared the linkcomm formulation (implemented in LKCM) with the Ahn formulation (used in LKCMsp, HLCnp, and HLC) across four weighted graphs. The Ahn formulation identified highly connected edge communities that were not detected by linkcomm. Consequently, in these cases, Ahn's approach outperformed linkcomm in terms of overlapping coherence and the identification of additional clusters across different network structures.

This trend is further supported by the overlap coverage metrics shown in Fig. 1B, where the linkcomm formulation consistently led to slightly lower values in terms of overlap and community coverage across the studied networks. Notably, ecST700 exhibited greater overlap coverage than ecST900, as the higher cutoff value of 900 reduced the number of connections between proteins, limiting their potential membership in multiple communities. In all three cases, overlap coverage remained above one, highlighting the necessity of clustering methods that account for overlapping communities.

Additionally, no major differences were observed in community coverage metrics between the chosen cutoff values, indicating that the proportion of nodes assigned to at least one community remained stable. However, it is worth noting that the Ahn formulation consistently achieved higher community coverage than linkcomm, despite never reaching full coverage (i.e. 1).

To assess the consistency among the three implementations of the Ahn formulation, we evaluated the coherence of the detected clusters in two widely used benchmark networks: The Karate Club and Les Misérables (Supplementary Fig. S2). In both cases, HLC, HLCnp, and LKCMsp achieved a maximum normalized mutual information value of 1 (Supplementary Fig. S2C), indicating that they identified identical clusters.

In contrast, LKCM, which deviates from the Ahn formulation, produced notably different clustering results. This divergence is

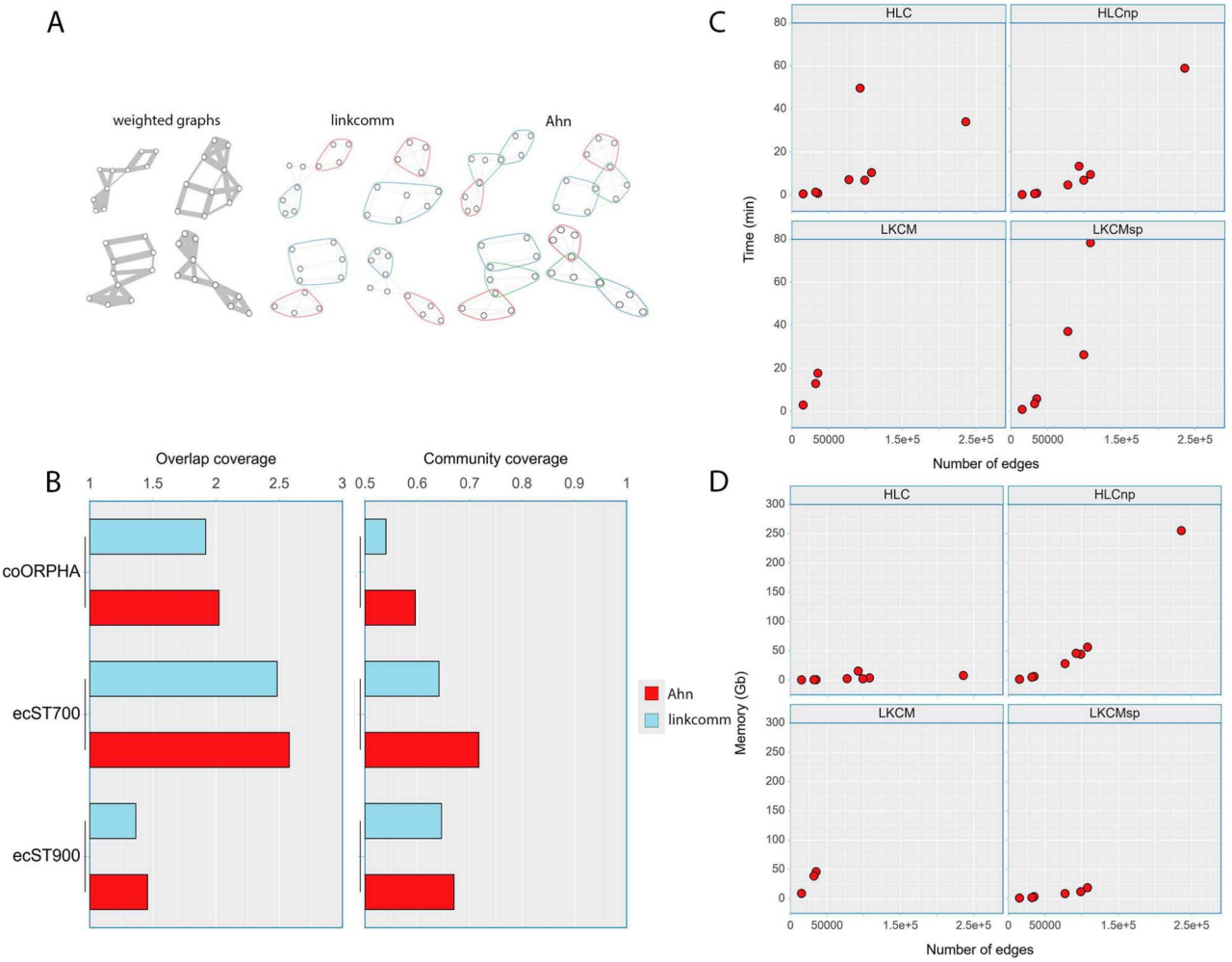

Figure 1. A) Graph representation of differences between the Ahn formulation (HLC, HLCnp and LKCMsp) and the linkcomm formulation (LKCM) using four different small networks. (B) Overlap and community coverage metrics measured for three real biological networks, calculated using the cases where LKCM could be executed. (C and D) Time and memory usage, respectively, across the different implementations.

reflected in its lower mutual information score and is visually evident in the graph plots (Supplementary Fig. S2A and B).

Technical performance and scalability of HLC implementations

With regard to technical performance, Fig. 1C and D show that both HLCnp and LKCMsp outperformed LKCM in terms of execution time (reductions of ×19.4 and ×3.3, respectively; Table 2) and memory usage (reductions of ×7.18 and ×15.21, respectively; Table 2). Notably, results could be generated for all eight networks in the cases of HLC and HLCnp, whereas LKCM only executed successfully for three networks and LKCMsp for six. Failures in execution were attributed to memory limitations (using more than 450 GB) or excessive execution time (lasting more than three days).

Although the original CDLIB implementation (HLC) demonstrated greater memory efficiency compared to our Python implementation, HLCnp (×53.4 and ×7.2, respectively; Table 2), its execution time increases significantly when handling dense networks (Supplementary Fig. S4). In contrast, our NumPy-based matrix approach exhibits superior scalability, as it uses vectorized operations to optimize performance. This is particularly evident for the Wikipedia network, where HLC requires considerably more time due to its high edge density (Fig. 1C and Supplementary Fig. S2A), as further supported by Supplementary Fig. S1.

Table 2. Mean ratio for time and memory used in all four HLC implementations with respect to LKCM, expressed as fold-change, i.e. how much faster it take to run, or how much less memory is used

Implementation	Time speedup	Memory reduction
LKCM	1×	1×
LKCMsp	3.312×	15.214×
HLCnp	19.404×	7.185×
HLC	12.77×	53.39×

Ratio is computed against LKCM time or memory values.

HLCnp introduces new functionalities that enhance performance by utilizing NumPy and SciPy for customizable linkage procedures (Supplementary Fig. S3) and Tanimoto matrix distance filtering (Fig. 2). The latter significantly reduces memory usage and execution time while preserving overlap and community coverage, particularly for low cutoffs such as 0.2 (Fig. 2, panels A and B).

Additionally, HLCnp supports multiple linkage methods, expanding beyond single linkage to include complete linkage, which can further improve coverage and overlap. This is

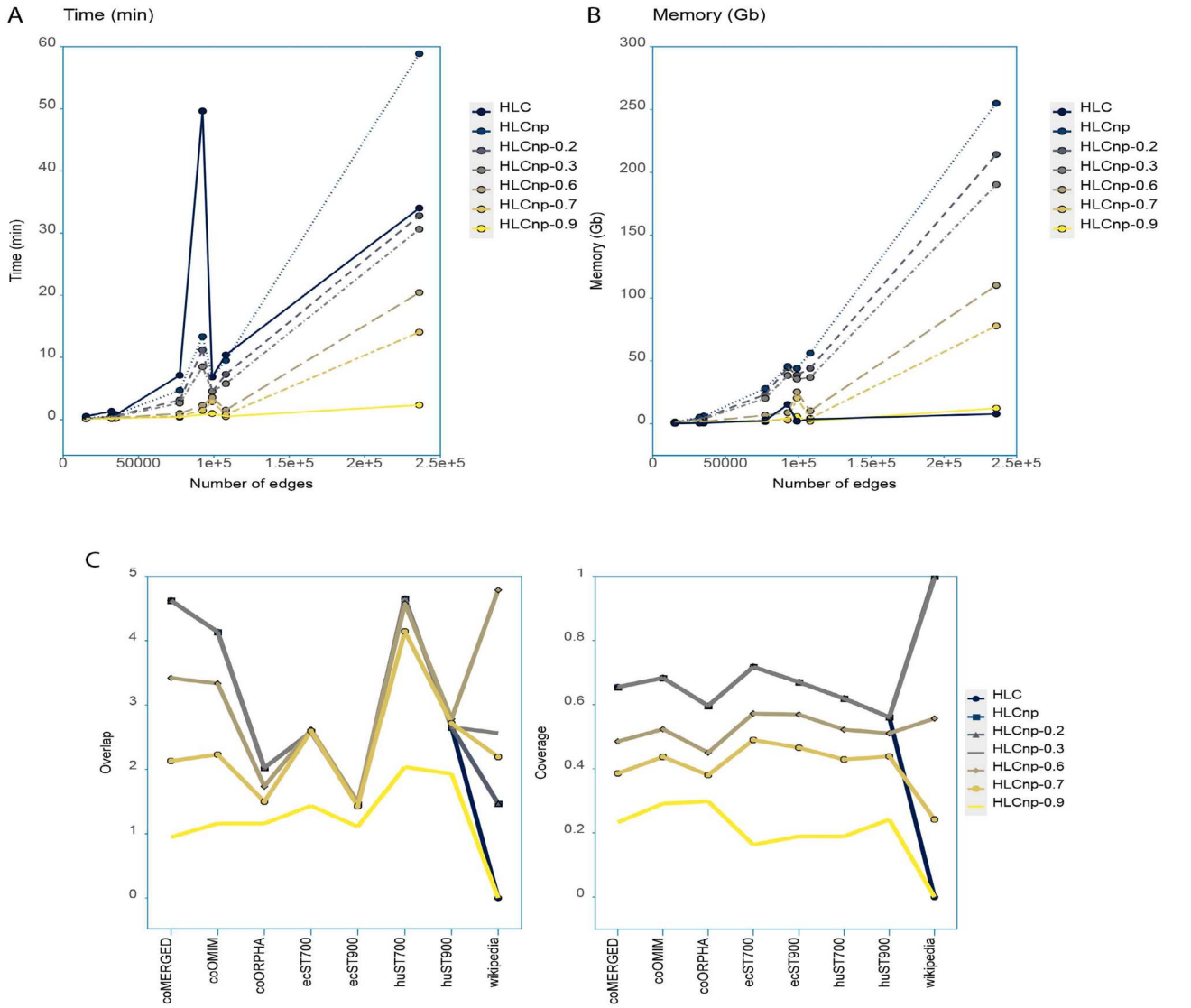

Figure 2. Application of filters to the Tanimoto matrix. Time (A), memory (B), coverage and overlap (C) are measured with respect to different cutoff filters. HLC and HLCnp were added to the analysis as controls for every dataset. The cutoff values applied to the filters are specified in the format "HLCnp-*cutoff value*".

demonstrated in Supplementary Fig. S3, where complete linkage consistently outperforms single linkage across nearly all tested networks.

Embedding analysis and evaluation of pathway cohesion in the human interactome

To evaluate the potential of different HLC algorithms in enhancing network representation learning in PPI networks, we first assessed whether pathways with high internal coherence (i.e. pathways whose members are highly connected) are better represented when HLC communities were integrated into the graph embedding procedure. As shown in Fig. 3, we observed that the distances in embedding space between genes within the same pathway decrease as the internal edge density increases. This trend is further enhanced when HLC clustering is incorporated into the graph embedding, suggesting that HLC improves the capture of pathway cohesion.

Building on these results, we evaluated the proximity of genes within coherent pathways (internal edge density > 0.75) in the embeddings generated from two human PPI networks (huST700 and huST900). We normalized the embedding distances between pathway genes as z-scores based on the overall distribution of distances in the embedding; negative z-scores indicate that pathway genes are closer together than average, whereas positive z-scores indicate larger distances.

As shown in Fig. 4A and B, integrating both node neighborhood and HLC communities produced a significant improvement, leading to the shortest average distance across the parameter settings tested. We further compared the HLC-based method against 11 baseline approaches (Supplementary Fig. S5), and in all cases, our method outperformed the alternatives. A t-test with BH correction confirmed the statistical significance of these results, with all adjusted *P*-values falling below.01. In contrast, when this combination was performed with Louvain clustering, the resulting embedding exhibited the poorest distance representation. This is likely because the Louvain algorithm, which produces a small number of large, non-overlapping communities (see Supplementary Fig. S6) can negatively impact the embedding process by distorting the natural structure of the network. This can occur when clustering is too coarse to reflect the finer-grained relationships,

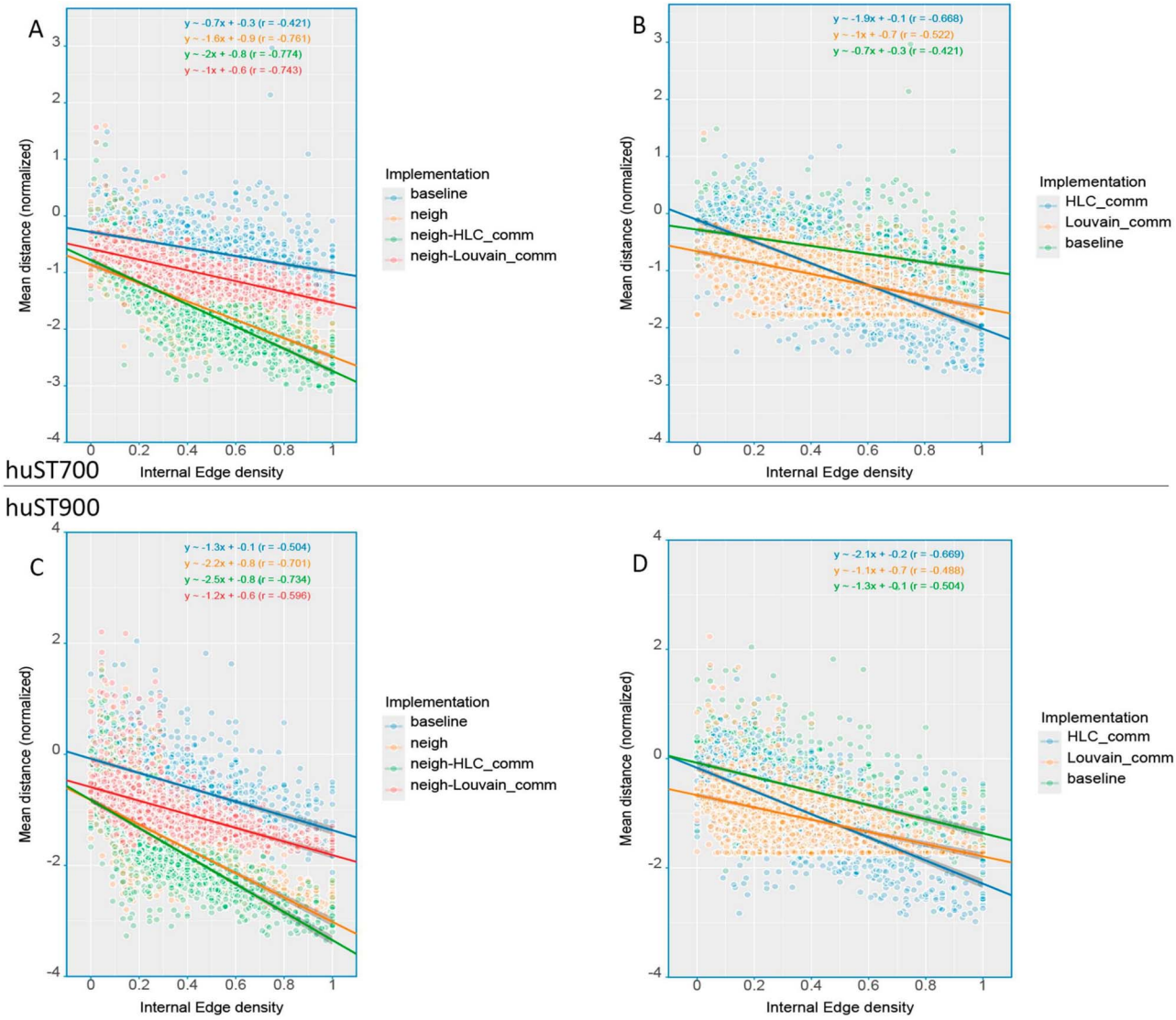

Figure 3. Relationship between embedding distance and internal edge density for huST700 (A and B) and huST900 (C and D) using various embedding strategies: baseline (random edge selection), neigh (random walk between connected nodes), HLC_comm (random selection within the same HLC community), Louvain_comm (random selection within the same Louvain community), and neigh-HLC_comm or neigh-Louvain_comm (random walk combining both neighborhood and community strategies). Linear regression is shown with the associated confidence region.

reducing the ability of embedding to capture meaningful local distinctions in the network. Additionally, since random walks play a key role in learning node representations, the rigid community structure imposed by Louvain may bias these walks by forcing them to remain within clusters longer than they naturally would.

We further explored the impact of the clustering methods on the embedding space of the huST networks (Supplementary Figs S7 and S8). Analysis of the top 20 clusters from each method revealed that Louvain clusters cover almost the entire network, while HLC clusters represent a smaller portion (Supplementary Fig. S7B vs. A for huST700 and Supplementary Fig. S8B vs. A for huST900). This difference greatly impacts the embedding: as illustrated in Supplementary Figs 7D and 8D, Louvain clustering groups nodes by community membership in a way that disrupts the original network structure (see Supplementary Figs S7B and S8B), whereas the HLC method better preserves this structure (see Supplementary Figs S7C and S8C).

Finally, we examined the top ten largest Reactome pathways (internal edge density > 0.75) to illustrate how our analysis represents biological pathways. Figures 4C and D show that using HLC clusters results in pathway members being closer together in the embedding space compared to those generated using Louvain clusters. Further UMAP analyses of the top five and ten largest pathways are shown in Figs 5 and 6, respectively. We see clear differences between the HLC and Louvain methods, with the former tending to show much clearer separation between pathways.

Case study: RASopathies

To further explore the biological relevance of candidate genes identified through our analysis, we conducted a focused case study examining their potential involvement in RASopathies and RASopathy-associated signaling. We selected a subset of genes for manual curation based on their presence in the prioritization lists obtained through embedding. For each gene, we assessed the strength of the evidence linking it to RASopathy phenotypes by integrating information from published literature, known signaling pathways, and documented pathogenic variants. This analysis

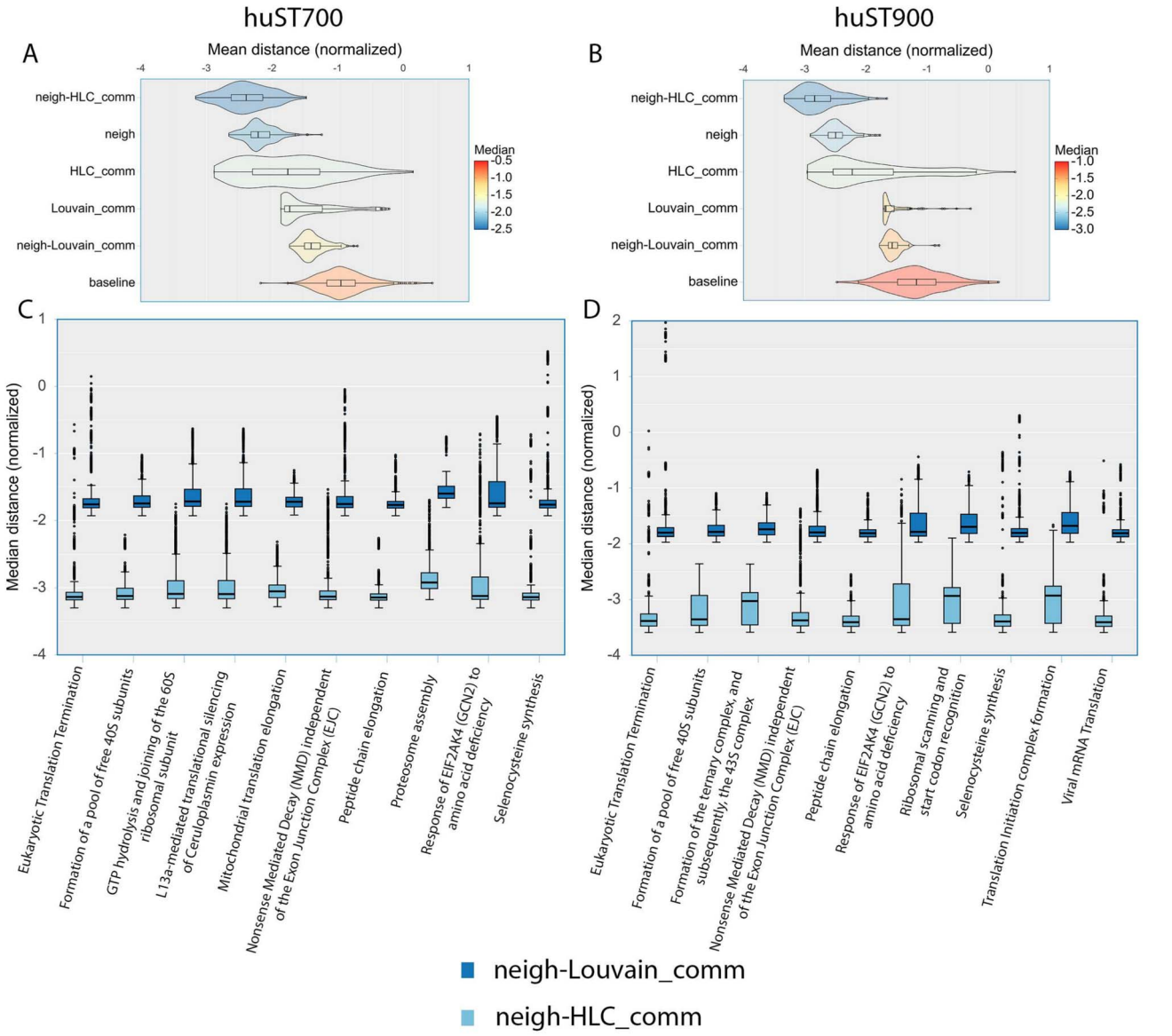

Figure 4. Embedding results using different node selection strategies: baseline (random selection of edges), neigh (random walk between connected nodes), HLC_comm (random selection of nodes within the same HLC community), Louvain_comm (random selection of nodes within the same Louvain community), and neigh-HLC_comm or neigh-Louvain_comm (random walk combining both neighborhood and community strategies). (A and B) Distributions of mean z-scores for distances between nodes within the same high-cohesion pathway (edge density > 0.75) in huST700 and huST900, respectively; (C and D) Distributions of mean z-scores for distances within the ten largest high-cohesion pathways in huST700 and huST900, highlighting neigh-HLC_comm (light blue) and neigh-Louvain_comm (dark blue).

allowed us to distinguish between genes with strong functional or genetic evidence of involvement in RASopathies, those with more tentative or indirect associations, and those with no clear evidence of relevance to the RASopathy disease spectrum.

In total, 21 RASopathy related genes, belonging to 9 different RASopathy groups were compiled through analysis of MONDO and Monarch, as shown in Table 3. The positions of these genes in the UMAP representation of the graph-embedding of the human PPI network are shown in Fig. 7A. We see that the HLC informed graph embedding approach (neigh-HLC_comm) leads to clear groupings of genes in close, but distinct areas of the UMAP representation.

For full details of the top 20 prioritized results for each of the RASopathies investigated, please see Supplementary Tables S2 and S3, for the huST700 and huST900 networks respectively.

We can also see, in Fig. 7B and C, that the top prioritized non-RASopathy related genes tend to lie close to the RASopathy related genes. Interestingly, we see a small group of prioritized genes far from the main groupings, these are all prioritized genes for Noonan syndrome-like disorder with loose anagen hair, and are close in the UMAP to the gene *PPP1CB* (Fig. 7D).

Our prioritization algorithm using all known RASopathy related genes together for the huST700 network identified several candidate genes with potential relevance to RASopathies among the top 20 genes based on the prioritization score. For instance, SHC domain proteins like SHC1 are involved in RAS-mediated signal activation [44]. In addition, MAPK7 (a lesser-known MAP kinase) and MAP3K2 also emerged among the top candidates; kinases from this family have key roles in the RAS/MAPK signaling pathway, particularly MAPK3 and MAPK1 [45].

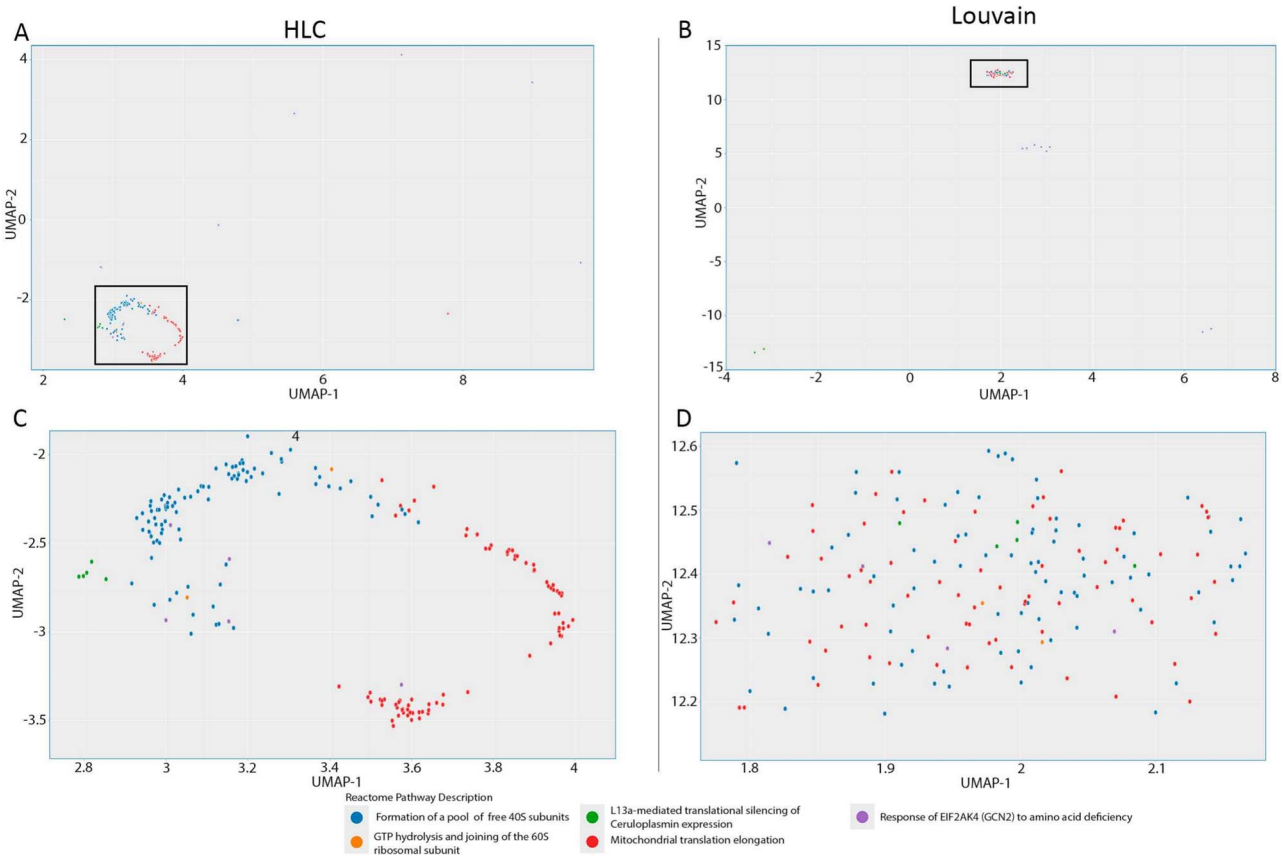

Figure 5. UMAP representation of the embeddings generated for huST700. The top five largest Reactome biological pathways (with an internal edge density > 0.75) are shown in different colors. (A and C) Embedding representation using HLC; (B and D) the representation obtained using Louvain clustering.

Table 3. RASopathies under study and their associated genes

RASopathy	Genes
CBL-related disorder	*CBL*
Costello syndrome	*HRAS*
Legius syndrome	*SPRED1*
Noonan syndrome	*BRAF*, *KRAS*, *LZTR1*, *MAP2K1*, *MAPK1*, *MRAS*, *NRAS*, *PTPN11*, *RAF1*, *RIT1*, *RRAS2*, *SOS1*, *SOS2*, *SPRED2*
Noonan syndrome with multiple lentigines	*BRAF*, *PTPN11*, *RAF1*
Noonan syndrome-like disorder with loose anagen hair	*PPP1CB*, *SHOC2*
RASopathy	*BRAF*, *CBL*, *HRAS*, *KRAS*, *LZTR1*, *MAP2K1*, *MAP2K2*, *MAPK1*, *MRAS*, *NF1*, *NRAS*, *PPP1CB*, *PTPN11*, *RAF1*, *RIT1*, *RRAS2*, *SHOC2*, *SOS1*, *SOS2*, *SPRED1*, *SPRED2*
Watson syndrome	*NF1*
cardiofaciocutaneous syndrome	*BRAF*, *KRAS*, *MAP2K1*, *MAP2K2*
neurofibromatosis type 1	*NF1*

Notably, *RAP1A* and *RAP1B*, members of the Ras subfamily implicated in cellular adhesion and morphogenesis, were also prioritized and have been shown to influence cell proliferation [46]. Rap1 has also shown possible association to Kabuki syndrome [47], a disorder increasingly recognized as part of the RASopathy spectrum due to its overlapping features and involvement of genes in the RAS/MAPK signaling pathway [48]. Another prioritized gene, *NKS1A* (also known as Odin) is a PTB-domain adaptor that regulates ER export of receptor tyrosine kinases (RTKs). It binds EphA2/ErbB2 in the ER and loads them into COPII vesicles.

The gene *ANKS1A* also acts downstream of EGFR and Eph receptors via its ankyrin and PTB domains [49]. By controlling the trafficking of RTKs that signal through RAS, ANKS1A can modulate Ras-MAPK activity. SKAP1 (Src kinase–associated phosphoprotein 1, or SKAP-55) is an adaptor in T cell signaling. It mediates RapL (NORE1B/RASSF3) membrane localization and LFA-1 activation in a process dependent on its PH domain and PI3K activity [50]. Although SKAP1 has not been studied in RASopathies, this role in PI3K-driven Rap1 signaling suggests a possible crosstalk between RAS and PI3K networks. Another highly ranked gene *RAPGEF5* (MR-GEF) is a Rap guanine exchange factor expressed during development. Mutations have been reported associated with congenital heart defects and neurological disorders, reflecting its developmental importance [51]. Because RapGEF5 activates Rap (a Ras-family GTPase) and influences Wnt/cell polarity pathways, it may interface with Ras/MAPK signaling in embryogenesis.

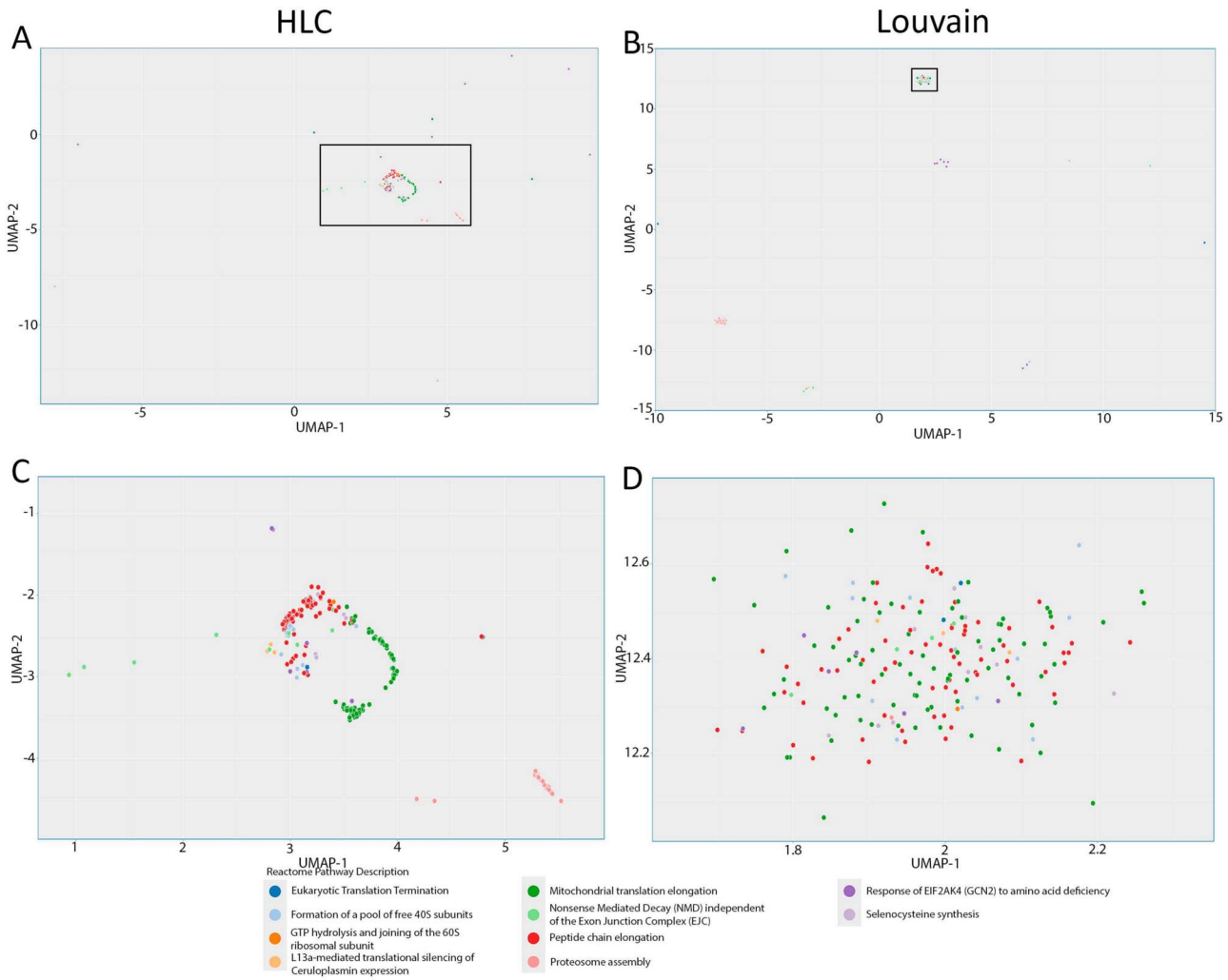

Figure 6. UMAP representation of the embeddings generated for huST700. The top ten largest Reactome biological pathways (with an internal edge density > 0.75) are depicted in different colors. (A and C) Embedding representation using HLC; (B and D) the representation obtained using Louvain clustering.

There were also a number of PI3K/AKT pathway members among the prioritized genes. These include *PIK3R1*, which encodes the p85α regulatory subunit of PI3K. Mutations in the C-terminal SH2 domain of p85α impair binding with p110α in PI3K complexes, which may indirectly affect Ras-pathway output [52]. These findings suggest cross talk between PI3K and Ras/MAPK signaling, which is to be expected given that PI3K is one of the main effectors of Ras [52]. Interestingly PI3K/AKT/mTOR pathway often cause developmental syndromes that share some similarity to RASopathies at the phenotypic level [53]. These findings show the potential use of our methodology to identify and investigate related pathways in the context of disease.

Furthermore, our analysis identified several prioritized genes that represent potential drug targets. For instance, ALK, a RTK known to be implicated in multiple cancers, has been successfully targeted using inhibitors such as crizotinib, which has shown efficacy in non-small cell lung cancer [54]. In addition, the identification of PRKCA (protein kinase C alpha) suggests that pharmacological modulators of PKC activity might have therapeutic implications in disorders characterized by altered RAS/MAPK signaling [45]. These findings demonstrate that our method not only prioritizes known RASopathy-related genes but also uncovers additional candidates with potential therapeutic relevance.

When the methodology was applied to the genes associated with specific RASopathies, we found additional genes of interest among the priotization results. For example, when using the Noonan syndrome associated genes, *RASA2* emerged as a high-ranking gene; it encodes a Ras GTPase-activating protein whose loss-of-function mutations have been associated with Noonan syndrome-like phenotypes [55, 56]. We also found *HRAS* and *NF1*, important genes associated with different RASopathies (neurofibromatosis type 1 and Costello syndrome respectively) in the top 20. In addition, important RAS pathway related genes were also found, including *BRA* and various Guanine nucleotide exchange factors involved in Ras signaling. Interestingly, the results using the Costello syndrome associated genes were broadly similar, with the top twenty including multiple RAS genes involved in other RASopathies. For CFC syndrome, multiple additional RAS related genes are also found. Together, these findings highlight the molecular overlap between distinct RASopathies and support the idea that shared dysregulation within the RAS/MAPK pathway underlies their related phenotypic features, enabling the identification of novel common associations across the spectrum.

We also inspected the results of the huST900 network analysis. For prioritization involving all genes associated with RASopathies,

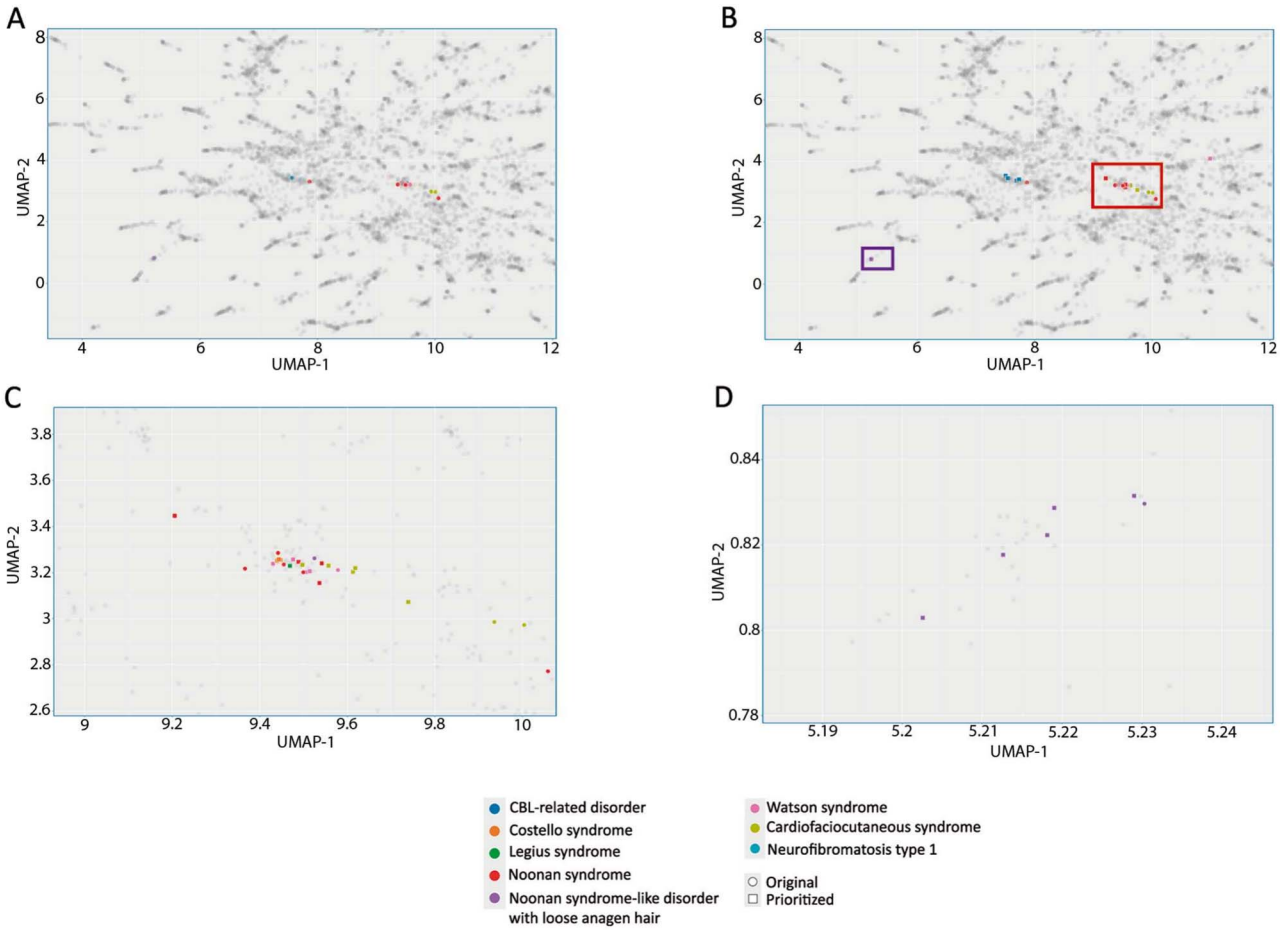

Figure 7. UMAP representation of the embeddings generated for huST700. Note that we have zoomed into a small part of the entire UMAP to better visualize the distances between these genes and other, non RASopathy-related genes. (A) All RASopathy-related genes. (B) All RASopathy-related genes, plus the top 5 prioritized genes for each RASopathy. (C) Zoomed in on the main cluster of RASopathy related genes (red box in B). (D) Zoomed in on the single Noonan gene found far from the other RASopaty genes, along with the other priotized Noonan-related genes that lie close in the UMAP representation (violet box in B).

results were broadly simiar to the huST700 analysis, with some differences. For example, RASA2, the Noonan-associated target was found. More generally, our analysis highlighted several other genes, including *RAP1B*, *RASGRF1*, *OSBPL3*, *RASGRF2*, *ARHGEF12*, *IL3RA*, *CNKSR1*, *CSF2RB*, *KIAA1549*, *GAB2*, and *RAPGEF5* that, although not classically linked to RASopathies, play integral roles in modulating Ras signaling. Given their involvement in the regulation of the Ras-MAPK pathway, these genes may represent novel risk factors or mechanistic contributors to developmental disorders associated with dysregulation of Ras signaling [33].

In addition to examining RASopathy-associated genes, we investigated the relative positions of the pathways to which these genes belong using UMAP visualization (Fig. 8). Notably, the genes involved in MAPK1 activation can be found in close proximity to those participating in the CSF3 signaling pathway and signaling by phosphorylated juxtamembrane, extracellular and kinase domain KIT mutants pathway genes. KIT signaling have been shown to activate the RAS/MAPK cascade, and that GM-CSF and c-Kit together promote proliferation [57, 58]. Given that MAPK1 (also known as ERK2) is a key effector in the RAS pathway, such spatial proximity suggests that cross-talk or co-regulation between these pathways may contribute to the abnormal signal transduction observed in RASopathies. However, it should be noted that the actual RASopathy-associated genes from the MAPK1 activation pathway (MAPK1 and MAP2K2) are found relatively far from the other members of this pathway in the UMAP visualization.

Moreover, the UMAP visualization also revealed that genes related to FGFR4 signaling, PDGFRA, and insulin signaling are positioned near a MAPK1 activation gene and, again, in the vicinity of CSF3 pathway genes. This arrangement is significant because FGFR4 is an established upstream activator of RAS/MAPK signaling [59], PDGFRA has been linked to RAS in a proliferation model, and the insulin signaling cascade is known active the Ras-MAPK pathway following the binding of Grb2 and SOS to IRS proteins [60]. Again, proximity of these pathways implies a potential convergence of signals that could amplify or modulate RAS activity. Further analysis could involved examining and investigating the genes that do not belong to these pathways, but fall between them in the UMAP representation.

Discussion

In this study, we developed and evaluated two enhanced implementations of the HLC algorithm to address challenges in embedding PPI networks. Our results show that clustering the network before embedding improves pathway representation by increasing coherence—meaning that functionally related genes are grouped more closely together in the embedding space, making it easier to identify biologically meaningful relationships. These results

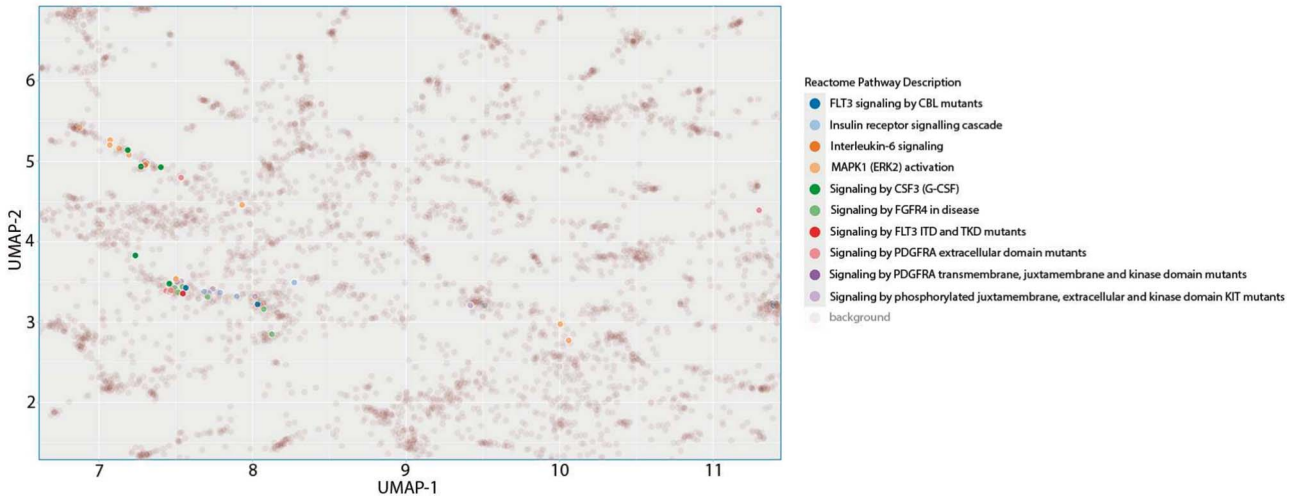

Figure 8. UMAP representation of the embeddings generated for huST700. Various Reactome biological pathways (with an internal edge density > 0.75) to which RASopathy associated genes belong are shown in different colors.

emphasize the value of incorporating community-aware strategies into network representation learning, as they enhance the ability to capture biologically relevant structures.

To address the computational limitations of HLC when applied to large biological networks, we introduced two novel implementations: HLCnp in Python and LKCMsp in R. LKCMsp adheres more closely to the original Ahn formulation while improving execution time and memory efficiency. HLCnp, integrated with CDLIB, extends functionality by optimizing dense network processing, providing greater flexibility in clustering, and enabling efficient filtering of the Tanimoto distance matrix. Benchmarking these tools on diverse weighted networks, including large biological datasets and dense graphs like Wikipedia, demonstrated their scalability and effectiveness.

We used our HLC implementation to enhance the embedding of the human proteome network. Among the tested approaches, integrating HLC clusters yielded the best performance (Fig. 4A and B). In contrast, the Louvain algorithm performed poorly, likely due to distortions in its embedding space (Supplementary Fig. S7). (Supplementary Fig. S7). More specifically, Louvain tends to identify a few large clusters, which can artificially separate functionally related genes if a Reactome pathway is spread across multiple clusters. Since the embedding process preserves local structures within clusters but not necessarily between them, nodes assigned to different Louvain clusters are distributed in isolated regions of the embedding space. As a result, the distances between pathway members increase, disrupting biological coherence and making it harder to capture meaningful functional relationships. In addition, within the clusters identified by Louvain, the algorithm was unable to reflect any meaningful internal structure, instead displaying a homogeneous distribution pattern.

For example, in UMAP analyses of the top five largest pathways (Fig. 5), the HLC method clearly differentiates pathways such as "Formation of a pool of free 40S subunits," "Mitochondrial translation elongation," and "L13a-mediated translational silencing of Ceruloplasmin expression," whereas the Louvain method groups them into a single cluster. This pattern persists when representing the full set of the top ten largest pathways (Fig. 6), where additional related pathways appear connected in the HLC embedding but remain indistinguishable in the Louvain embedding. Notably, in the "Proteome assembly" pathway, although members cluster together with both methods, the HLC algorithm positions them closer to the previously described pathways, indicating that Louvain clustering distorts the embedding space and fails to accurately capture the relationships between biological pathways.

Furthermore, while both methods grouped protein degradation pathways (e.g. proteasome-related) in close proximity, only HLC positioned them near protein synthesis processes, accurately reflecting their functional relationship. In contrast, Louvain left them isolated (Fig. 6). These findings illustrate how biological insights can be extracted from the UMAP representation of network embeddings.

Interestingly, Louvain clustering appears to perform worse than no clustering at all, which is in contrast with previous findings, such as those reported in CRARE [21]. To better understand this discrepancy, future work should explore different clustering parameters for embedding across a range of networks. Reducing the weight assigned to clustering integration might help mitigate distortions in the embedding space. In fact, a similar pattern was observed in CRARE, where performance deteriorated when a high weight was given to clustering. Furthermore, testing alternative clustering methods could provide additional insights into whether the issue arises from the Louvain algorithm itself or from the specific way it interacts with the embedding process.

These findings demonstrate that HLC produces a biologically meaningful organization of pathways, highlighting the value of overlapping cluster structures in complex biological networks. Consequently, this improved embedding could enhance deep learning applications in genomics and proteomics, where embedding serves as an essential preprocessing step [6, 12–14]. Moreover, the embedding representation itself holds significant potential, as demonstrated in this study. Researchers can map genes of interest onto the embedding space, explore their connectivity to biological pathways, and interpret their functional roles within this context.

In benchmark comparisons with other node embedding methods, our strategy consistently outperformed the alternatives, with node2vec [5] and DeepWalk [11] ranking as the next best performers. Interestingly, Role2Vec [28] exhibited substantial variability in its results. This may be due to the heterogeneous nature of biological pathways, with some of them having common structural properties that are more favorable to the Role2Vec methodology.

The results of the application of our method to RASopathy related genes found multiple novel candidate genes that may contribute to the pathogenesis of RASopathies. Several of these

genes were found across multiple RASopathies, highlighting a significant overlap in the underlying biological processes shared by these syndromes.. The identification of RASA2, a gene already implicated in Noonan syndrome-like phenotypes, supports the algorithm's validity, while the discovery of additional modulators of Ras signaling such as RAP1B, RASGRF1, and others opens new avenues for research into the molecular mechanisms underlying these disorders. Moreover, the recognition of druggable targets such as ALK and PIK3CD shows the potential for identifying existing targeted therapies for repurposing in RASopathy contexts, although efforts to validate these candidate genes through functional assays and clinical studies are essential to translate these findings into effective therapeutic strategies.

While our approach demonstrates strong performance across diverse networks, several limitations should be acknowledged. First, the accuracy of biological interpretation depends on the quality and completeness of the input PPI data, which may carry biases due to underrepresented or poorly annotated proteins [61, 62]. Second, although our method performs well on the human proteome network, its generalizability to other species or to highly sparse or noisy biological graphs remains to be evaluated.

Future work will focus on extending our approach to multi-omics integration, enabling embeddings that simultaneously capture transcriptomic, epigenomic, and metabolomic relationships. Moreover, applying these methods across a broader range of diseases and biological systems will help to assess the robustness and transferability of our framework. Finally, deeper validation of candidate gene predictions through experimental and clinical collaborations will be essential to fully realize the translational potential of our approach.

In conclusion, our work underscores the importance of overlapping clustering in biological networks and introduces computationally efficient tools to enhance graph embeddings. By leveraging community-aware embeddings, we establish a robust framework for uncovering biologically meaningful insights, particularly in pathway and disease analysis within human PPI networks.

Key Points

- Optimized Edge-Based Clustering for Large Networks.We introduce HLCnp (Python/CDLIB) and LKCMsp (R/linkcomm), optimized implementations of the Hierarchical Link Clustering (HLC) algorithm. These tools enhance efficiency and accuracy in clustering large, weighted, undirected networks, incorporating complete linkage and Tanimoto filtering for improved performance.
- Enhanced Graph Embedding via Community-Constrained Random Walks. By integrating overlapping clustering with network embedding, our approach improves pathway representation in the human interactome. Constraining random walks to pre-computed communities improve embedding space with respect to biological patwhay.
- Application to RASopathies for Gene Prioritization. We apply our method to multiple RASopathies, including Noonan and Costello syndrome, improving the visualization of gene associations and prioritizing novel causal genes and additional pathway-related genes.
- Accessible Implementations and Benchmarking We provide open-source tools and benchmarking workflows, enabling researchers to apply and extend our methods in diverse biological and biomedical network studies.

Acknowledgments

The authors thank the Supercomputing and Bioinnovation Center (SCBI) of the University of Málaga for their provision of computational resources and technical support (http://www.scbi.uma.es/site). We thank to Junta de Andalucía its support to PAIDI BIO-267 group. Funding for open access charge: Universidad de Málaga / CBUA.

Author contributions

J.A.G.R., F.G.C., J.R.P., and P.S. conceived the methodology. F.G.C., P.S., and J.R.P. developed the software that implements the protocol. J.A.G.R., F.G.C., J.R.P., P.S., and E.R. analyzed the results and provided interpretation. P.S.Z., F.G.C., J.R.P., and E.R. wrote the manuscript. J.A.G.R., J.R.P., and P.S. were involved in planning of the study, contributed to the acquisition of research funding and headed the project. All authors read and approved the final version of the manuscript.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: No competing interest is declared.

Funding

This work is supported by funds from Spanish Ministry of Economy and Competitiveness [PID2019-108096RB-C21 and PID2022-140047OB-C21]; the Institute of Health Carlos III (project IMPaCT-Data, exp. IMP/00019), co-funded by the European Union, European Regional Development Fund (ERDF, "A way to make Europe"); and the European Union through the project EURAS (Horizon Europe HORIZON-HLTH-2022-DISEASE-06, Project ID: 101080580) to J.A.G.R. The EURAS project receives funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No. 101080580 (HORIZON-HLTH-2022-DISEASE-06). Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. Biomedicine research projects [PI-0075-2017 and PEER-0019-2020] are funded by Fundación Progreso y Salud. J.R.P. and E.R. hold a research grant from the Instituto de Investigación Biomédica de Málaga (IBIMA-Plataforma BIONAND) [PI RARE 24-03]. F.G.C. is a predoctoral researcher from "Ayudas para contratos predoctorales para la Formación del Profesorado Universitario" (FPU21/01449) supported by the Ministerio de Ciencia, Innovación y Universidades. The "CIBER de Enfermedades Raras" is an initiative from the ISCIII (Spain). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Data availability

HLCnp is available in the CDLIB library (https://github.com/GiulioRossetti/cdlib), and LKCMsp can be accessed at https://github.com/jimrperkins/linkcomm. Workflow with the benchmarking process and datasets corresponding to HLC implementations and network embedding are available at https://github.com/seoanezonjic/network_hlc_benchmark.git and https://github.com/federedef/embedding_hlc_benchmark.

References

1. Szklarczyk D, Kirsch R, Koutrouli M. *et al.* The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* 2023;**51**:D638–46. https://doi.org/10.1093/nar/gkac1000
2. Eberhard P, Kern M, Aichem M. *et al.* PathwayNexus: a tool for interactive metabolic data analysis. *Bioinformatics* 2024;**40**:btae310. https://doi.org/10.1093/bioinformatics/btae310
3. Díaz-Santiago E, Jabato FM, Rojano E. *et al.* Phenotype-genotype comorbidity analysis of patients with rare disorders provides insight into their pathological and molecular bases. *PLoS Genet* 2020;**16**:e1009054. https://doi.org/10.1371/journal.pgen.1009054
4. Hu JX, Thomas CE, Brunak S. Network biology concepts in complex disease comorbidities. *Nat Rev Genet* 2016;**17**:615–29.
5. Grover A, Leskovec J. Node2vec: scalable feature learning for networks. In: Krishnapuram B, Shah M, Smola AJ, Aggarwal CC, Shen D, Rastogi R (eds.), *KDD: Proceedings International Conference on Knowledge Discovery & Data Mining*, **2016**. New York, NY: ACM, 2016, 855–64.
6. Wang N, Zeng M, Li Y. *et al.* Essential protein prediction based on node2vec and XGBoost. *J Comput Biol* 2021;**28**:687–700. https://doi.org/10.1089/cmb.2020.0543
7. Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput* 2003;**15**:1373–96. https://doi.org/10.1162/089976603321780317
8. Ezzat A, Min W, Li X-L. *et al.* Drug-target interaction prediction using ensemble learning and dimensionality reduction. *Methods* 2017;**129**:81–8. https://doi.org/10.1016/j.ymeth.2017.05.016
9. Nikolentzos G, Vazirgiannis M. Learning structural node representations using graph kernels. *IEEE Trans Knowl Data Eng* 2021;**33**:2045–56.
10. Jabato F, M, Rojano E, Perkins JR. *et al.* Kernel based approaches to identify hidden connections in gene networks using NetAnalyzer. In: Goos G, Hartmanis J (eds.), *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Heidelberg, Germany: Springer Nature, 2020.
11. Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online learning of social representations. In: Leskovec J, Wang W (eds.), *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, pp. 701–10. New York, NY, USA: Association for Computing Machinery, 2014.
12. Chen Z-H, You Z-H, Guo Z-H. *et al.* Predicting drug-target interactions by node2vec node embedding in molecular associations network. In: Huang DS, Bevilacqua V, Hussain A (eds.), *Intelligent Computing Theories and Application: 16th International Conference, ICIC 2020, Bari, Italy, October 2–5, 2020, Proceedings, Part II*, pp. 348–58. Berlin, Heidelberg: Springer-Verlag, 2020.
13. Bernett J, Blumenthal DB, List M. Cracking the black box of deep sequence-based protein–protein interaction prediction. *Brief Bioinform* 2024;**25**:bbae076.
14. Zou H-T, Ji B-Y, Xie X-L. A multi-source molecular network representation model for protein–protein interactions prediction. *Sci Rep* 2024;**14**:6184. https://doi.org/10.1038/s41598-024-56286-w
15. Fang H, Liu J, Li L. *et al.* Community detection in complex networks using node2vec with spectral clustering. *Physica A* 2020;**545**:123633.
16. Kojaku S, Radicchi F, Ahn Y-Y. *et al.* Network community detection via neural embeddings. *Nat Commun* 2024;**15**:9446. https://doi.org/10.1038/s41467-024-52355-w
17. Kovács B, Kojaku S, Palla G. *et al.* Iterative embedding and reweighting of complex networks reveals community structure. *Sci Rep* 2024;**14**:17184. https://doi.org/10.1038/s41598-024-68152-w
18. Dayu H, Liang K, Zhou S. *et al.* scDFC: a deep fusion clustering method for single-cell RNA-seq data. *Brief Bioinform* 2023;**24**:bbad216.
19. Dayu H, Dong Z, Liang K. *et al.* High-order topology for deep single-cell multiview fuzzy clustering. *IEEE Trans Fuzzy Syst* 2024;**32**:4448–59. https://doi.org/10.1109/TFUZZ.2024.3399740
20. Dong Y, Chawla NV, Swami A. Metapath2vec: Scalable representation learning for heterogeneous networks. In: Matwin S, Yu S, Farooq F (eds.), *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 135–44. Halifax NS Canada: ACM, 2017.
21. Zhang H, Kou G, Peng Y. *et al.* Role-aware random walk for network embedding. *Inform Sci* 2024;**652**:119765. https://doi.org/10.1016/j.ins.2023.119765
22. Guo K, Wang Q, Lin J. *et al.* Network representation learning based on community-aware and adaptive random walk for overlapping community detection. *Appl Intell* 2022;**52**:9919–37. https://doi.org/10.1007/s10489-021-02999-8
23. Keikha MM, Rahgozar M, Asadpour M. Community aware random walk for network embedding. *Knowledge-Based Syst* 2018;**148**:47–54. https://doi.org/10.1016/j.knosys.2018.02.028
24. Rozemberczki B, Davies R, Sarkar R. *et al.* GEMSEC: Graph embedding with self clustering. In: Spezzano F, Chen W, Xiao X (eds.) *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '19*, pp. 65–72. New York, NY, USA: Association for Computing Machinery, 2020.
25. Ahn YY, Bagrow JP, Lehmann S. Link communities reveal multiscale complexity in networks. *Nature* 2010;**466**:761–4. https://doi.org/10.1038/nature09182
26. Wang Y, Chen Q, Yang L. *et al.* Overlapping structures detection in protein–protein interaction networks using community detection algorithm based on neighbor clustering coefficient. *Front Genet* 2021;**12**:689515, 1–14. https://doi.org/10.3389/fgene.2021.689515
27. Huang L, Wang G, Wang Y. *et al.* Link clustering with extended link similarity and EQ evaluation division. *PLoS One* 2013;**8**:e66005. https://doi.org/10.1371/journal.pone.0066005
28. Ahmed NK, Rossi R, Lee JB. *et al. Learning Role-Based Graph Embeddings* 2018. arXiv preprint arXiv:1802.02896; 2018.
29. Yang D, Rosso P, Li B. *et al.* NodeSketch: Highly-efficient graph embeddings via recursive sketching. In: Teredesai A, Kumar V, Li Y, Rosales R, Terzi E, Karypis G (eds.), *proceedings of the 25th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, pp. 1162–72. Anchorage AK USA: ACM, 2019.
30. Cao S, Lu W, Qiongkai X. GraRep: learning graph representations with global structural information. In: Bailey J, Moffat A, Aggarwal CC, de Rijke M, Kumar R, Murdock V, Sellis T, Yu JX (eds.), *Proceedings of the 24th ACM international on conference on*

information and knowledge management, pp. 891–900. Melbourne Australia: ACM, 2015.
31. Kalinka AT, Tomancak P. Linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. *Bioinformatics* 2011;**27**:2011–2. https://doi.org/10.1093/bioinformatics/btr311
32. Rossetti G, Milli L, Cazabet R. CDLIB: A python library to extract, compare and evaluate communities from complex networks. *Appl Network Sci* 2019;**4**:1–26. https://doi.org/10.1007/s41109-019-0165-9
33. Zenker M. Clinical overview on RASopathies. *Am J Med Genet C Semin Med Genet* 2022;**190**:414–24. https://doi.org/10.1002/ajmg.c.32015
34. Pagano-Márquez R, Córdoba-Caballero J, Martínez-Poveda B. *et al.* Deepening the knowledge of rare diseases dependent on angiogenesis through semantic similarity clustering and network analysis. *Brief Bioinform* 2022;**23**:1–15. https://doi.org/10.1093/bib/bbac220
35. Bahar ME, Kim HJ, Kim DR. Targeting the RAS/RAF/MAPK pathway for cancer therapy: From mechanism to clinical studies. *Signal Transduct Target Ther* 2023;**8**:455.
36. Mikolov T, Sutskever I, Chen K. *et al. Distributed Representations of Words and Phrases and their Compositionality*. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds.), *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. Lake Tahoe, NV, USA: Curran Associates, Inc., 2013, 3111–19.
37. Bottou L, Laboratories TB. Stochastic gradient learning in neural networks. In: *Proceedings of Neuro-Nîmes 1991*. Nîmes, France: EC2, 1991.
38. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: Moschitti A, Pang B, Daelemans W (eds.) *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–43. Doha, Qatar: Association for Computational Linguistics, 2014.
39. Qiu J, Dong Y, Ma H. *et al.* Network embedding as matrix factorization: Unifying DeepWalk, LINE, PTE, and node2vec. In: Chang Y, Zhai C, Liu Y, Maarek Y (eds.) *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 459–67. Marina Del Rey CA USA: ACM, 2018.
40. Ou M, Cui P, Pei J. *et al.* Asymmetric transitivity preserving graph embedding. In: Krishnapuram B, Shah M, Smola AJ, Aggarwal CC, Shen D, Rastogi R (eds.) *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1105–14. San Francisco California USA: ACM, 2016.
41. Sun DL, Févotte C. Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy: Institute of Electrical and Electronics Engineers, 2014, 6201–5.
42. Rozemberczki B, Kiss O, Sarkar R. Karate Club: an API oriented open-source python framework for unsupervised learning on graphs. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, Vol. 29, pp. 3125–32, 2020.
43. Fabregat A, Jupe S, Matthews L. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res* 2018;**46**:D649–55. https://doi.org/10.1093/nar/gkx1132
44. Tidyman WE, Rauen KA. Pathogenetics of the RASopathies. *Hum Mol Genet* 2016;**25**:R123–32. https://doi.org/10.1093/hmg/ddw191
45. Rauen KA, Tidyman WE. RASopathies - what they reveal about RAS/MAPK signaling in skeletal muscle development. *Dis Model Mech* 2024;**17**. p. 1-13. https://doi.org/10.1242/dmm.050609
46. D'Silva NJ, Mitra RS, Zhang Z. *et al.* Rap1, a small GTP-binding protein is upregulated during arrest of proliferation in human keratinocytes. *J Cell Physiol* 2003;**196**:532–40. https://doi.org/10.1002/jcp.10331
47. Jaśkiewicz A, Pajak B, Orzechowski A. The many faces of Rap1 GTPase. *Int J Mol Sci* 2018;**19**:2848. https://doi.org/10.3390/ijms19102848
48. Tsai I-C, McKnight K, McKinstry SU. *et al.* Small molecule inhibition of RAS/MAPK signaling ameliorates developmental pathologies of kabuki syndrome. *Sci Rep* 2018;**8**:10779.
49. Lee H, Noh H, Mun J. *et al.* Anks1a regulates COPII-mediated anterograde transport of receptor tyrosine kinases critical for tumorigenesis. *Nat Commun* 2016;**7**:12799. https://doi.org/10.1038/ncomms12799
50. Raab M, Smith X, Matthess Y. *et al.* SKAP1 protein PH domain determines RapL membrane localization and Rap1 protein complex formation for T cell receptor (TCR) activation of LFA-1. *J Biol Chem* 2011;**286**:29663–70. https://doi.org/10.1074/jbc.M111.222661
51. Alharatani R, Griffin JN, Liu KJ. Expression of the guanine nucleotide exchange factor, RAPGEF5, during mouse and human embryogenesis. *Gene Expr Patterns* 2019;**34**:119057. https://doi.org/10.1016/j.gep.2019.119057
52. Cuesta C, Arévalo-Alameda C, Castellano E. The importance of being PI3K in the RAS signaling network. *Genes* 2021;**12**:1094. https://doi.org/10.3390/genes12071094
53. Canaud G, Hammill AM, Adams D. *et al.* A review of mechanisms of disease across PIK3CA-related disorders with vascular manifestations. *Orphanet J Rare Dis* 2021;**16**:306.
54. Shaw AT, Kim D-W, Mehra R. *et al.* Ceritinib in ALK-rearranged non-small-cell lung cancer. *N Engl J Med* 2014;**370**:1189–97. https://doi.org/10.1056/NEJMoa1311107
55. Aoki Y, Niihori T, Inoue S-I. *et al.* Recent advances in RASopathies. *J Hum Genet* 2016;**61**:33–9. https://doi.org/10.1038/jhg.2015.114
56. Chen P-C, Yin J, Hui-Wen Y. *et al.* Next-generation sequencing identifies rare variants associated with Noonan syndrome. *Proc Natl Acad Sci U S A* 2014;**111**:11473–8. https://doi.org/10.1073/pnas.1324128111
57. Sheikh E, Tran T, Vranic S. *et al.* Role and significance of c-KIT receptor tyrosine kinase in cancer: a review. *Bosn J Basic Med Sci* 2022;**22**:683–98.
58. Lennartsson J, Rönnstrand L. Stem cell factor receptor/c-kit: from basic science to clinical implications. *Physiol Rev* 2012;**92**:1619–49. https://doi.org/10.1152/physrev.00046.2011
59. Hebron KE, Hernandez ER, Yohe ME. The RASopathies: from pathogenetics to therapeutics. *Dis Model Mech* 2022;**15**:dmm049107.
60. Taniguchi CM, Emanuelli B, Ronald Kahn C. Critical nodes in signalling pathways: insights into insulin action. *Nat Rev Mol Cell Biol* 2006;**7**:85–96. https://doi.org/10.1038/nrm1837
61. de Silva E, Thorne T, Ingram P. *et al.* The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol* 2006;**4**:39. https://doi.org/10.1186/1741-7007-4-39
62. Aj C, David A, Mje S. PhenoRank: Reducing study bias in gene prioritization through simulation. *Bioinformatics (Oxford, England)* 2018;**34**:2087–95. https://doi.org/10.1093/bioinformatics/bty028

Briefings in Bioinformatics, 2025, **26(4)**, bbaf320
https://doi.org/10.1093/bib/bbaf320
Case Study