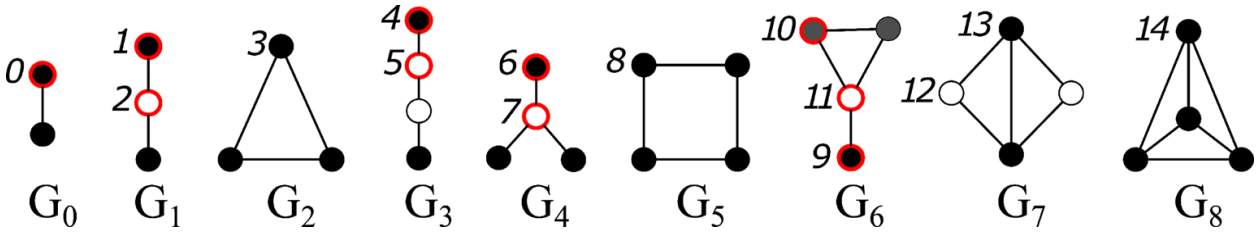
Supplementary Information for:

Linear functional organization of the omic embedding space

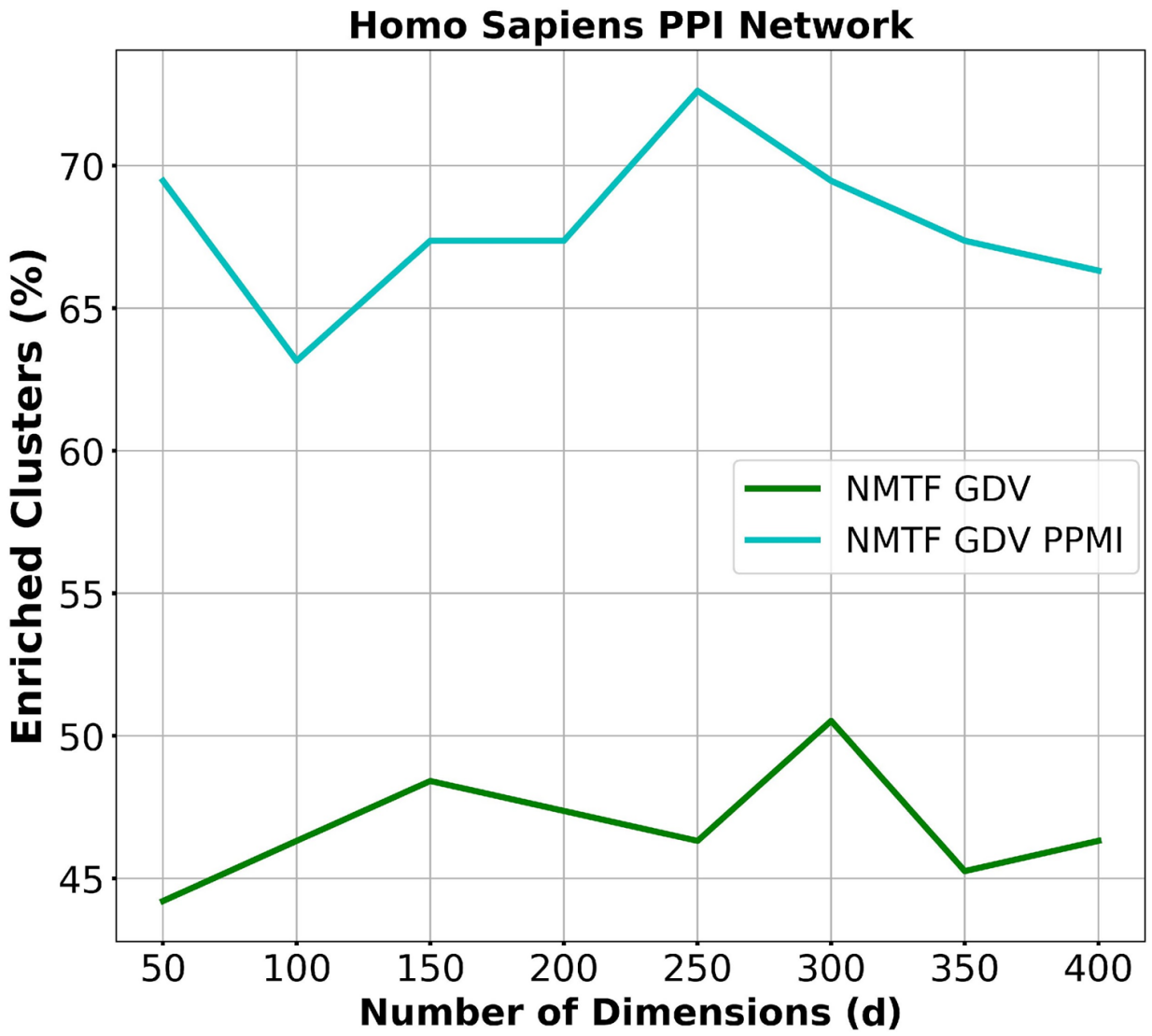
Alexandros Xenos, Nöel Malod-Dognin, Stevan Milinković and Nataša Pržulj

This Supplementary Information document contains Supplementary Figures 1 to 6 and Supplementary Tables 1 to 3. Additionally, it contains the NMTF Multiplicative Update Rules as well as the details for the enrichment analysis.

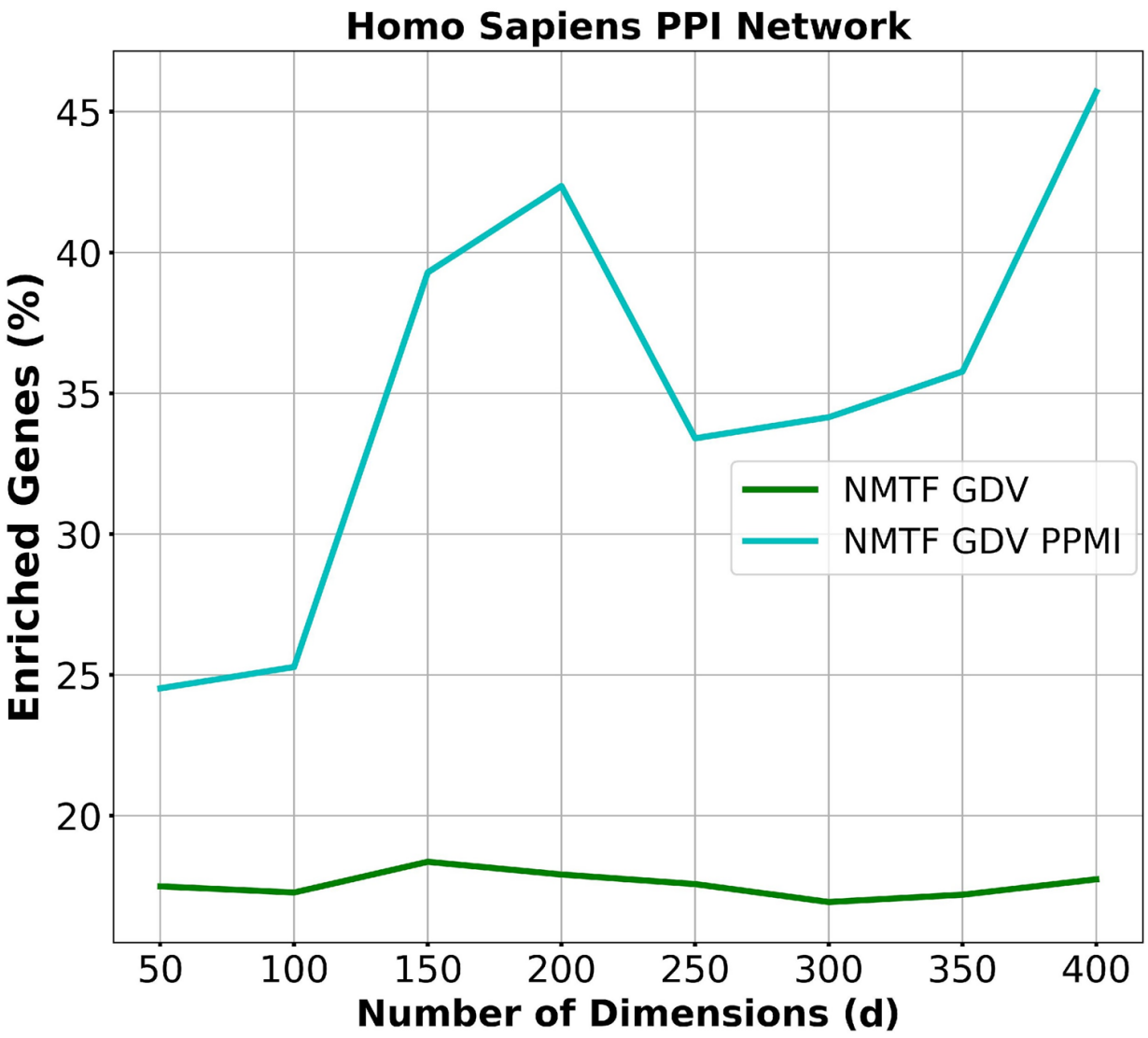
Supplementary Figures



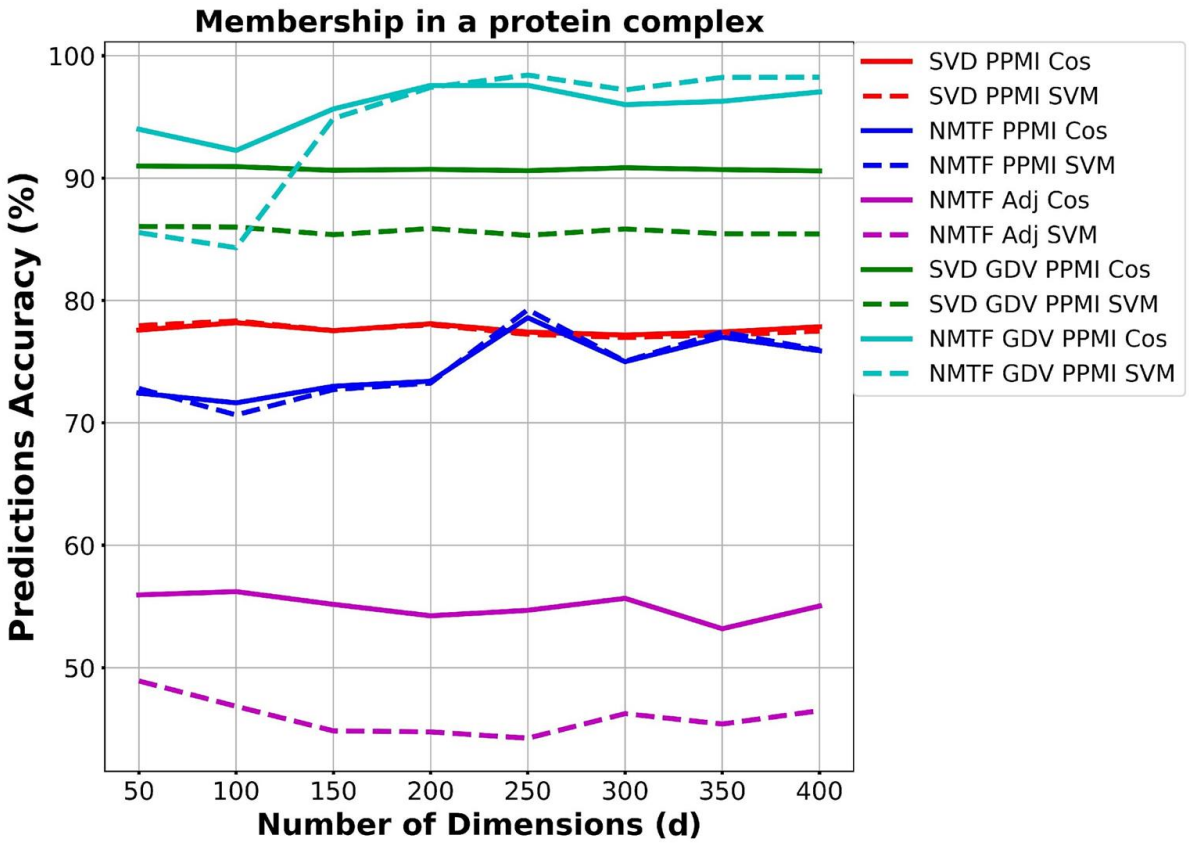
Supplementary Figure 1. The nine 2- to 4-node graphlets and their 15 orbits. Within each graphlet, nodes belonging to the same orbit have the same color (either white, black or grey). The ten non-redundant orbits (Yaveroğlu et al., 2014), whose counts cannot be derived from the counts of the other orbits, are highlighted in red.



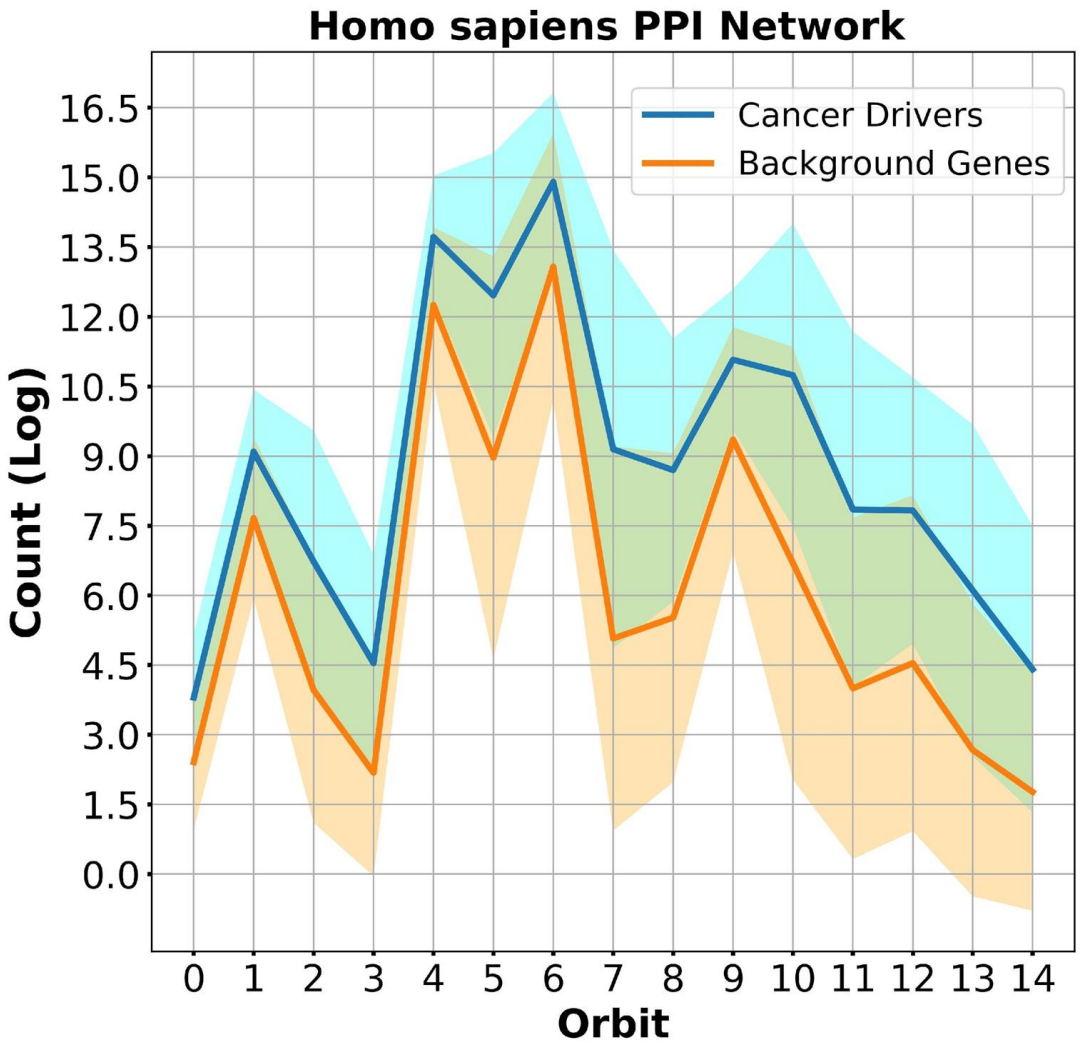
Supplementary Figure 2. For each number of dimensions, d (x-axis), the percentage of enriched clusters (y-axis) obtained in the embedding spaces generated by the NMTF GDV decomposition (green) and NMTF GDV PPMI decomposition (cyan).



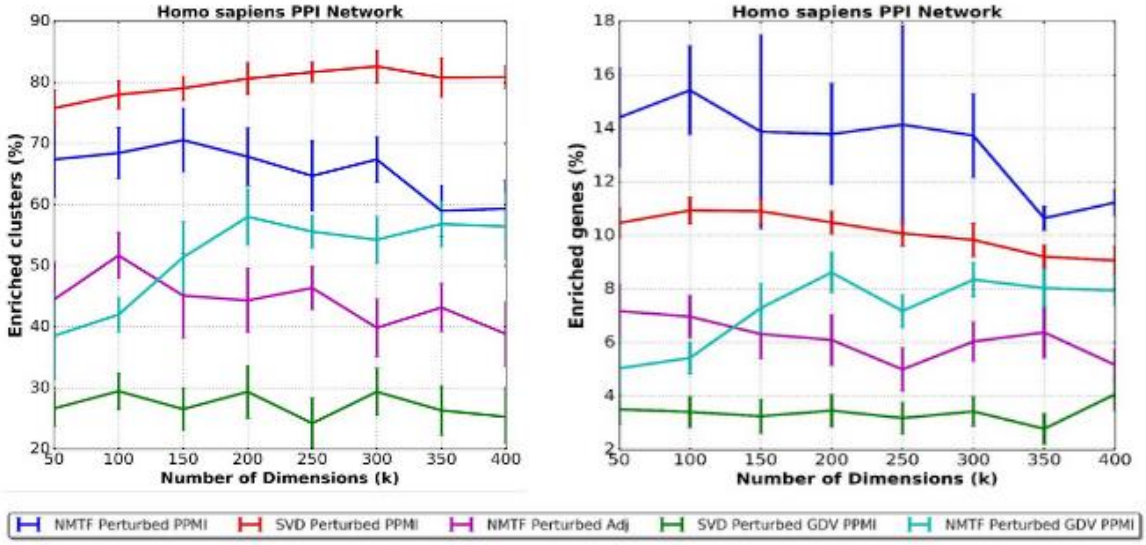
Supplementary Figure 3. GDV decompositions, gene enrichments. For each number of dimensions, d (x-axis), the percentage of enriched genes (y-axis) in the clusters obtained in the embedding spaces generated by the NMTF-based decompositions of the GDV similarity matrix (green) and the GDV PPMI matrix (cyan).



Supplementary Figure 4. Protein complex membership prediction as binary classification problem. For each number of dimensions, d (x-axis), the lines show the accuracy (y-axis) of the predictions based on the cosine based prediction strategy and the dashed lines the accuracy of the SVM predictions that are obtained in the embedding spaces generated by the SVD PPMI decomposition (red), NMTF PPMI decomposition (blue), NMTF Adj decomposition (purple), SVD GDV PPMI decomposition (green) and NMTF GDV PPMI decomposition (cyan).



Supplementary Figure 5. GDV signatures of the cancer driver genes and the background genes. For the cancer driver genes (blue) and the background genes (orange), the lines show the average GDV vectors in log-scale (y-axis) and the shades their variation for each orbit (x-axis).



Supplementary Figure 6. **Functional coherence of the embedding spaces of perturbed networks.** For 15 permutations, for each dimension (x-axis), the average percentage of clusters (y-axis, panel A) with at least one GO BP function enriched and the average percentage of enriched genes (y-axis, panel B) in the clusters that are obtained in the embedding space generated by each of the NMTF PPMI decomposition (blue), SVD PPMI decomposition (red), NMTF Adj decomposition (purple), SVD GDV PPMI decomposition (green) and NMTF GDV PPMI decomposition (cyan).

Supplementary Tables

Dimension, *d*	Number of shared enriched GO BP terms in both embeddings, obtained by: factorizing the Adj matrix, and factorizing the GDV PPMI matrix	Number of unique enriched GO BP terms when factorizing the Adjacency matrix	Number of unique enriched GO BP terms when factorizing GDV PPMI matrix	Overlap Coefficient
50	1169	3031	1069	0.52
100	1446	2645	1292	0.53
150	1117	2877	1404	0.44
200	1146	3331	1499	0.43
250	733	3257	1025	0.42
300	1010	2743	1283	0.44
350	976	2799	1254	0.44
400	813	3147	1081	0.43

Supplementary Table 1. For each number of dimensions, *d* (the first column), the number of shared enriched GO BP terms (the second column) and uniquely enriched GO BP terms for the two methods: NMTF-based embeddings obtained by factorizing the Adjacency matrix of the human protein-protein interaction (PPI) network (the third column), the GDV PPMI matrix of the human PPI network (the fourth column) and the Overlap Coefficient of the two different enriched sets (the fifth column).

Dimension, *d*	Number of shared enriched GO BP terms in both embeddings, obtained by: factorizing the PPMI matrix, with NMTF and SVD decompositions	Number of unique enriched GO BP terms when factorizing the PPMI matrix with NMTF	Number of unique enriched GO BP terms when factorizing PPMI matrix with SVD	Overlap Coefficient
50	2496	842	1278	0.75
100	3097	916	969	0.77
150	3082	1127	1188	0.73
200	2881	1454	1348	0.68
250	3017	1390	1122	0.73
300	2944	1687	1345	0.69
350	2962	1323	1290	0.70
400	3016	1700	1176	0.72

Supplementary Table 2. For each dimension, *d* (the first column), the number of shared enriched GO BP terms (the second column) and uniquely enriched GO BP terms for the two decompositions: NMTF-based embeddings (the third column), the SVD embeddings (the fourth column) obtained by factorizing the PPMI matrix of the human protein-protein interaction (PPI) network and the Overlap Coefficient of the two different enriched sets (the fifth column).

Orbit	Avg Degree (Log-Scale) Cancer Drivers	Avg Degree (Log-Scale) Background Genes	Mann Whitney U p-value
0	3.80	2.41	0
1	9.10	7.67	0
2	6.73	3.96	0
3	4.54	2.18	0
4	13.72	12.26	0
5	12.46	8.96	0
6	14.90	13.08	0
7	9.15	5.07	0
8	8.70	5.52	0
9	11.07	9.36	0
10	10.74	6.69	0
11	7.85	4.00	0
12	7.83	4.54	0
13	6.11	2.67	0
14	4.41	1.77	0

Supplementary Table 3. For each orbit for up to 4-node graphlets (the first column), the average graphlet orbit degree in log scale of cancer drivers (the second column) and background genes (the third column) as well as the corresponding p-value of the Mann W hitney U test (the fourth column) for each pair of distributions. Regarding the p-value, their values are close to 0, but due to the fact that p-values in Python are float64 objects (i.e. 16 decimals are reported), they are rendered to 0.

NMTF Multiplicative Update Rules

To solve the optimization problem, we use a fixed point method that initializes the matrix factors *G, S* and *P*, as follows, and iteratively uses the multiplicative update rules to converge towards a locally optimal solution.

Initial Solution: To generate initial G, S and P matrices, we use the Singular Value Decomposition (SVD) based strategy (Qiao, 2015). We initialize the diagonal elements of the S matrix with the singular values of the Σ matrix. However, U and V matrices of the SVD (see details Main Manuscript Section 2.3.1) can contain negative entries and thus, to comply with the non-negativity constraint of the NMTF, we use only their positive entries (we replace the negative entries with 0) to initialize G and P matrices, respectively. This strategy makes the solver deterministic and also reduces the number of iterations that are needed to achieve convergence (see Qiao (2015) for more details).

Multiplicative Update Rules:

$$S \leftarrow S \sqrt{\frac{G^T X P}{G^T G S P^T P}},$$

$$G \leftarrow G \sqrt{\frac{X P S^T}{G S P^T P S^T}},$$

$$P \leftarrow P \sqrt{\frac{X^T G S}{P P^T X^T G S}}$$

Evaluation Criteria: To measure the quality of the factorization, we compute the Relative Square Error (RSE) between the input matrix, X, and its corresponding decomposition, $X \approx GSP^T$, as:

$$RSE = \frac{\|X - GSP^T\|_F^2}{X_F^2}$$

We stop the iterative solver when the value of the RSE is not decreasing anymore, or after 500 iterations.

Enrichment Analysis

To assess if the genes that are close in the embedding space have similar biological functions, we perform clustering and enrichment analysis for the embeddings of the human PPI network. To obtain the clusters of genes, on the embedding vectors obtained from matrix factor *G* for the NMTF and to $U_d\sqrt{\Sigma_d}$ for SVD, we apply *k*-means clustering, which assigns each gene to the cluster whose centroid is the closest to the gene. To define the number of clusters, k, we use the heuristic rule of thumb, where n is the number of data points (Kodinariya and Makwana, 2013). Following that rule, the number of clusters is *k = 95* in human PPI network.

The probability that an annotation is enriched in a cluster is computed with the hypergeometric test (sampling without replacement strategy):

$$p = 1 - \sum_{i=0}^{X-i} \binom{K}{i} \binom{M-K}{N-i} / \binom{M}{N}$$

where *N* is the number of annotated genes in the cluster, *X* is the number of genes in the cluster that are annotated with the given annotation, *M* is the number of annotated genes in the network and *K* is the number of genes in the network that are annotated with the annotation in question. An annotation is considered to be statistically significantly enriched if its enrichment p-value, after Benjamini and Hochberg correction (Benjamini and Hochberg, 1995) for multiple hypothesis testing, is lower than or equal to 5%.

We measure the quality of the clustering by computing the percentage of genes having at least one of their annotations enriched in their clusters over all annotated genes.