Semi-Supervised Graph Learning for Cancer-Specific Protein Complex Discovery

Jennifer Li

Brown University

Advisor — Second Reader: Alper Uzun

Nominal Advisor — Reader: Ritambhara Singh

April 30, 2025

Abstract

This project proposes to leverage graph deep learning to identify differentially expressed protein communities, or complexes, in human cancers. Formulated as a transductive semi-supervised learning task, we propose cancer specific protein complex identification by curating the training data by cancer type. We start from the CORUM dataset of experimentally validated human protein complexes, then restrict to complexes with one or more subunits that are differentially expressed in tumor vs. control tissues, as indicated by the recently published CancerProteome database [1] [2]. Using this curated training data, we then leverage graph deep learning to uncover potentially unknown protein complexes which may play significant roles in cancer. By identifying protein complexes within these networks, we aim to enhance our understanding of the underlying biological mechanisms. Identification of such complexes is crucial for biological research as they inform drug discovery, disease treatment, and understanding of protein function. Given the inherent challenges of PPI network data, such as noise, biological variation, and sparsity, deep learning is particularly well suited to this task.

1 Background

1.1 Community Detection

Community detection is a key task in network analysis which aims to partition graphs into subgraphs, with nodes in a given subgraph more densely connected to one another than to nodes in different subgraphs. These subgraphs, or communities, may either overlap or remain distinct from one another, depending on the algorithm used. When applied to real-world networks, identifying these densely connected modules uncovers the hierarchical structure of the graph, providing valuable insights and facilitating deeper analysis [3].

Numerous community detection methods and algorithms have been developed over time. Traditional methods such as graph partitioning, divisive algorithms, and spectral clustering have laid the foundation for this field. Among these, the *Kernighan–Lin* algorithm is an early and still frequently used graph partitioning method, aiming to minimize the number of edges linking two connected communities [3]. Another such popular and well-known method is k-means clustering, which iteratively selects centroids based on a predefined number of clusters. Modularity-based approaches, which aim to optimize a modularity score that measures the strength of division of a network into communities, have also remained very popular. Furthermore, techniques have been developed which utilize spectral properties of graphs, such as the adjacency matrix and Laplacian matrix, for the community detection task [3].

Among the most effective and widely used algorithms are the Louvain and Leiden algorithms. Louvain's method is designed to maximize modularity while efficiently identifying communities, whereas Leiden builds on this by improving the quality and stability of the partitions [4] [5]. These techniques have proven essential for analyzing complex networks across various domains, including biology, social science, and information technology.

However, while traditional techniques have proven effective for small datasets, the increasing scale and complexity of modern networks have rendered these traditional approaches less practical [6]. In recent years, the rise in deep learning techniques has brought about a shift in the community detection task from traditional algorithms to models such as Graph Neural Networks (GNNs), Graph Attention Networks (GATs), and Autoencoders (AEs) [6].

As network size and complexity continue to grow, deep learning offers a number of significant advantages:

1. Allowing for the embedding of high-dimensional data into lower-dimensional representations [6].

2. More effective pattern detection and flexibility [6].

3. The ability to perform well despite sparsity in network data [7].

1.2 Protein-Protein Interaction Networks

One of the many applications in which community detection in networks plays a critical role is in the analysis of Protein-Protein Interaction networks (PPIs). In these networks, interacting and non-interacting protein pairs may be represented as a network graph, in which communities of interacting proteins may correspond to protein-protein complexes which perform some biological functions, including signal transduction, metabolic cycles, or metastasis, among others [8]. Thus, the identification of functional modules, or communities, in PPI networks may provide insights into biological interactions underlying complex disease.

As PPI network data may be sparse, and the underlying biological data complex and high-dimensional, deep learning has risen as a popular technique for predictive tasks related to PPI networks [8]. In existing research, the application of GNNs to PPI networks is often formulated as a link-prediction task. That is, deep learning has been successfully used in prediction tasks for interactions between proteins, such as protein binding or functional relationships [8].

More recently, however, research has turned to applying deep learning-based community detection models for identification of protein complexes. Approaches such as the Bernoulli-Poisson based Neural Overlapping Community Detection (NOCD) model, network embedding model, and protein information graph have demonstrated success in predicting protein complexes from large-scale protein-protein interaction (PPI) networks [9][10][11][12]. These methods have leveraged deep learning to uncover complex relationships between proteins, improving the accuracy of protein complex identification.

In this study, we look to extend this research by identifying differentially expressed protein complexes specific to various cancer types. To do so, we curate multiple training datasets tailored to cancer-associated protein communities. As semi-supervised graph learning models can perform well even with limited labeled data, we investigate whether modifying the

community detection task using targeted label subsets enables such models to learn cancer-associated complexes effectively. Given the success of deep learning in representing and imputing PPI network links and communities, this project seeks to evaluate this task from the perspective of cancer-specific protein communities.

2 Methodology

2.1 Datasets

The model is trained on the hu.MAP 2.0 human protein interaction network [13]. We use the CORUM dataset of human protein complexes, which offers manually curated and experimentally validated communities as ground truth [1].

To analyze differentially expressed protein complexes across cancer types, we incorporate the CancerProteome dataset, which captures both protein abundance and post-translational modification data [2]. This resource is derived from publicly available mass spectrometry (MS)-based experimental data, enabling a broad characterization of the proteome landscape by cancer type.

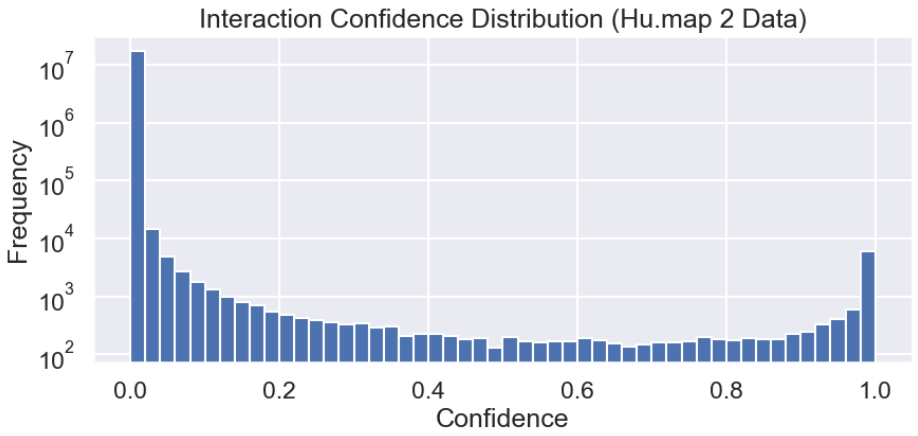


Figure 1: Confidence distribution of the hu.MAP 2.0 interaction network.

Given the large number of low-confidence interactions in the hu.MAP network, we filter to include only edges with confidence scores > 0.5 [Fig. 1]. If this step yields a disconnected graph, we select the largest connected graph as the training network and modify the included CORUM communities accordingly.

2.2 Preprocessing

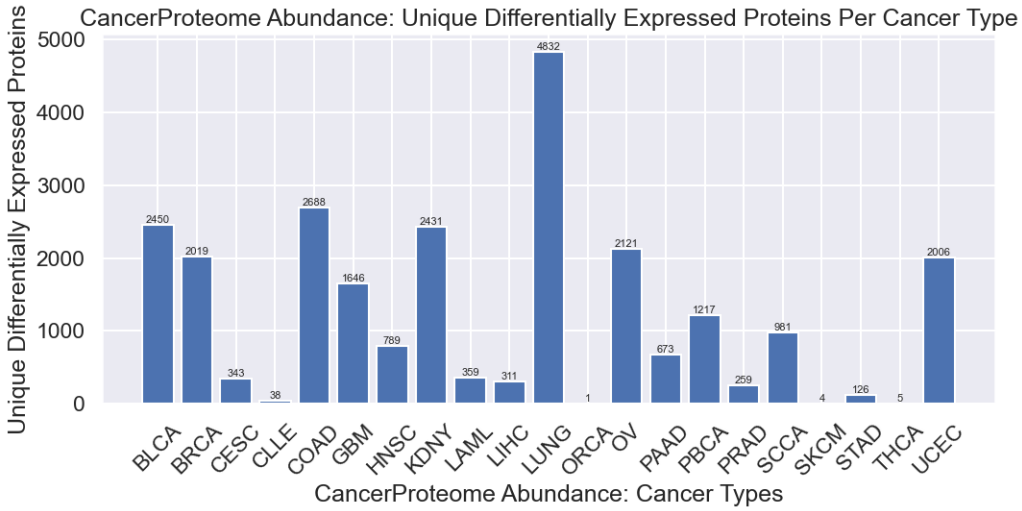
The ground truth CORUM data is initially filtered to identify proteins present in one or more CORUM complexes but absent from the hu.MAP interaction network. CancerProteome data was then processed to include all differentially expressed proteins — defined as those with significant abundance changes and post-translational modifications (PTMs) between tumor and control tissue samples — across 21 cancer types [Fig. 2][Table 1].

Abbreviation	Full Description
BLCA	Bladder Cancer
BRCA	Breast Cancer
CESC	Cervical Cancer
CLLE	Chronic Lymphocytic Leukemia
COAD	Colon Adenocarcinoma (Colon Cancer)
GBM	Glioblastoma
HNSC	Head and Neck Squamous Cell Carcinoma
KDNY	Kidney Cancer
LAML	Acute Myeloid Leukemia
LIHC	Liver Hepatocellular Carcinoma (Liver Cancer)
LUNG	Lung Cancer
ORCA	Oral Squamous Cell Carcinoma
OV	Ovary Cancer
PAAD	Pancreatic Adenocarcinoma (Pancreatic Cancer)
PBCA	PediatricAYA Brain Tumors
PRAD	Prostate Adenocarcinoma (Prostate Cancer)
SCCA	Squamous Cell Carcinoma of the Anus (Anal Canal Cancer)
SKCM	Skin Cutaneous Melanoma (Skin Cancer)
STAD	Stomach Adenocarcinoma (Gastric Cancer)
THCA	Thyroid Cancer
UCEC	Uterine Corpus Endometrial Carcinoma (Endometrial Carcinoma)

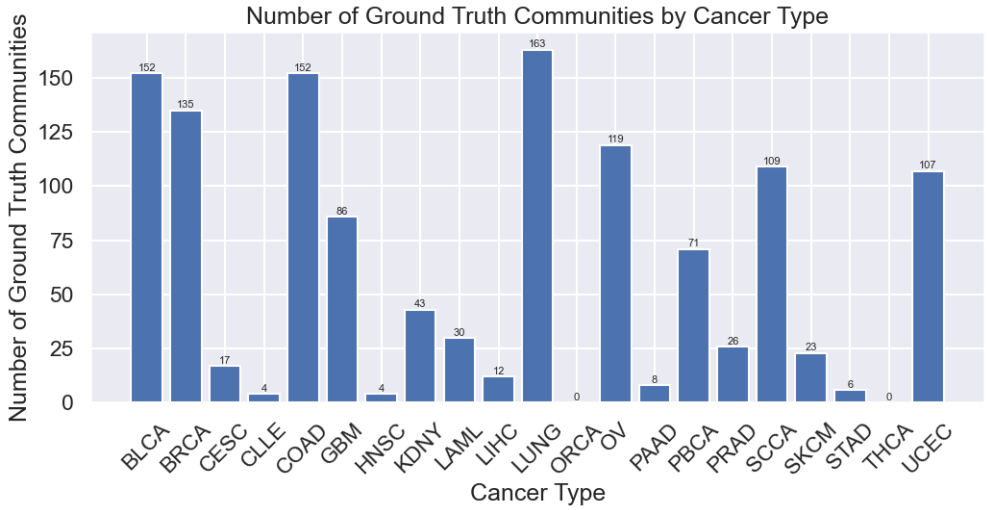
Table 1: The 21 cancer types included in the CancerProteome database with abbreviations listed.

Differentially expressed proteins from CancerProteome were then mapped to CORUM's ground truth protein complexes. For each cancer type, a complex was excluded if none

of its subunits exhibited differential expression between tumor and control tissues. The remaining cancer-specific ground truth complexes were subsequently used for model training.



(a) Visualizing the total number of unique differentially expressed proteins provided via CancerProteome for each of the 21 provided cancer types.



(b) Visualizing the total number of ground truth communities for each of the 21 provided cancer types.

Figure 2: Ground truth was constructed using differentially expressed proteins from CancerProteome combined with experimentally validated CORUM complexes to produce communities by cancer type.

Of the 21 cancer types, we select the 9 with the highest number of ground truth communities: BLCA, BRCA, COAD, GBM, LUNG, OV, PBCA, SCCA, and UCEC [Table 1] [Fig. 2].

2.3 Model Selection

We adapt the Attention Overlapping Community Detection (AOCD) model, a modified version of the NOCD model previously discussed [14] [12]. While graph convolution, as in NOCD, weights a given node's neighbors equally at each layer, graph attention learns attention weights that assign a measure of relevance to the neighbors of said node. Thus, graph attention models are able to learn more complex relationships and have demonstrated improved performance on a number of benchmark community detection datasets [14].

In accordance with the Bernoulli-Poisson probabilistic model, we sample the entries A_{uv} from the graph's adjacency matrix i.i.d. as

$$A_{uv} \sim \text{Bernoulli}(1 - \exp(-F_u F_v^T))$$

where $F \in \mathbb{R}_{\geq 0}^{N \times C}$ is the graph's community affiliation matrix for N nodes and C communities [9]. Based on this generative model, we instead train the model to output F given an adjacency matrix A and inputs X as follows:

$$F := \text{GAT}_\theta(A, X)$$

The negative log-likelihood is then calculated as:

$$-\log \Pr(A|F) = -\sum_{(u,v)\in E} \log(1 - \exp(-F_u F_v^T)) + \sum_{(u,v)\notin E} F_u F_v^T$$

where E is the set of all edges in the network. To address imbalance between due to edge sparsity, we sample from the sets of edges and non-edges to balance the two terms, as described in Shchur (2019) and Sismanis (2025).

$$L_{\text{BP}} = -\mathbb{E}_{(u,v)\sim p_E}\left[\log(1 - \exp(-F_u F_v^T))\right] + \mathbb{E}_{(u,v)\sim p_N}\left[F_u F_v^T\right]$$

However, at this point, we aim to introduce some measure of supervision using ground truth communities derived from differentially expressed proteins by cancer type. To do so, we include an additional asymmetric loss (ASL) term, which addresses positive-negative label

imbalance in multi-label classification, as described in Ridnik (2021). ASL loss is defined as follows:

$$L_{\text{ASL}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\left[(1-p)^{\gamma+}\log(p)\right] - (1-y_{i,c})\left[(p_m)^{\gamma-}\log(1-p_m)\right] \tag{1}$$

where $\gamma+$ and $\gamma-$ are decoupled focusing parameters for the positive and negative loss terms, p is the output probability of node i belonging to community c, $y_{i,c}$ is the true label of node i belonging to community c, and p_m is the shifted probability [15]. Then, to increase the contribution of rare positive samples while decreasing that of negative samples, we set $\gamma- > \gamma+$ and the probability margin m to some value greater than zero [15].

We then construct the overall loss as:

$$L = L_{\text{BP}} + \lambda L_{\text{ASL}}$$

where $0 < \lambda < 1$ weights the degree of supervised asymmetric loss incorporated.

3 Results

Cancer Type	Extended-Q	CON	TPR
BLCA	0.030	0.64	0.69
BRCA	0.030	0.66	0.69
COAD	0.03	0.64	0.7
GBM	0.018	0.68	0.76
LUNG	0.035	0.63	0.68
OV	0.034	0.63	0.7
PBCA	0.00	0.67	0.69
SCCA	0.024	0.66	0.74
UCEC	0.016	0.66	0.7

Table 2: Ground truth community modularity (Extended-Q), conductance (CON), and triangle participation ratio (TPR) metrics.

Cancer Type	Extended-Q	CON	TPR
BLCA	0.014	0.39	0.57
BRCA	0.0075	0.50	0.45
COAD	0.0052	0.53	0.39
GBM	0.02	0.24	0.65
LUNG	0.011	0.41	0.54
OV	0.0087	0.41	0.5
PBCA	0.015	0.4	0.51
SCCA	0.0093	0.34	0.56
UCEC	0.013	0.37	0.58

Table 3: Predicted community modularity (Extended-Q), conductance (CON), and triangle participation ratio (TPR) metrics.

The modified AOCD model was trained separately on the nine datasets with highest number of ground truth communities. Following model training, we then measure predicted community robustness using structural metrics in comparison to ground truth communities [Table 2] [Table 3].

Note that although we use differentially expressed protein communities as ground truth, we do not rely on evaluation metrics that require such ground truth labels. Given that we chose not to hold out on nodes with ground truth labels for the purposes of validation and testing, such methods would not accurately reflect model performance. The reasoning for this choice is as follows:

- Holding out labels would further reduce an already limited set of known labels, especially for cancer types such as PBCA and GBM, with 71 and 86 ground truth communities, respectively.
- We aim to maximize information from our given ground truth labels. All available protein complex data is used to "seed" the Bernoulli-Poisson probabilistic model with known cancer-specific communities to tune the generated outputs.
- Since our graphs are not attributed and our focus is more on graph structure than generalization, traditional evaluation paradigms are less applicable.
- The Bernoulli-Poisson model is inherently unsupervised. While we do introduce some degree of supervision through the inclusion of asymmetric loss, we do not believe this is sufficient to justify evaluation based on the limited ground truth labels.

Instead, we evaluate using structural metrics for community detection methods. These include Extended-Q (Extended Modularity), CON (Conductance), and TPR (Triangle Participation Ratio).

Briefly, Extended-Q, or modularity for overlapping communities, is defined as [16]:

$$M_{c_r}^{ov} = \sum_{i \in c_r} \frac{\sum_{j \in c_r, i \neq j} a_{ij} - \sum_{j \notin c_r} a_{ij}}{d_i \cdot s_i} \cdot \frac{n_{c_r}^e}{n_{c_r} \cdot \binom{n_{c_r}}{2}}$$

where c_r refers to the rth community, a_{ij} refers to the whether nodes i and j are connected in the graph's adjacency matrix a, d_i refers to the degree of ndoe i, s_i refers to the number

of communities node i belongs to, n_{c_r} refers to the number of nodes within cluster c_r and $n^e_{c_r}$ refers to the number of edges within cluster c_r [16].

Conductance measures community quality through the number of inter-community edges and the number of edges leaving the community [17][18]:

$$Conductance = \frac{O_c}{2I_c + O_c}$$

where O_c refers to the number of edges pointing outside the community c, and I_c refers to the number of edges within the community c.

Triangle Participation Ratio measures community connectivity using the fraction of inter-community nodes that belong to at least one triangle [19][18]:

$$TPR = \frac{T_c}{N_c}$$

where T_c refers to the number of nodes within community c which form a triangle with other nodes in c, and N_c refers to the total number of nodes in c.

The three metrics, Extended-Q, CON, and TPR, were evaluated and aggregated for both ground truth and predicted communities [Table 2] [Table 3]. The values are similar, though we see a slightly lower conductance and TPR on predicted communities.

To further evaluate whether the semi-supervised AOCD approach in predicting communities structurally reflective of the ground truth, we run three unsupervised community detection algorithms, CoAch, Angel, and Graph Entropy, on the preprocessed hu.MAP network. We then calculate Extended-Q, CON and TPR for the resultant communities [Table 4].

Overlapping Community Detection Algorithm	Extended-Q	CON	TPR
CoAch [20]	0.18	0.51	1.0
Angel [21]	0.42	0.23	0.95
Graph Entropy [22]	0.036	0.81	0.14

Table 4: Extended-Q, CON, and TPR metrics for communities constructed from the CoAch, Angel, and Graph Entropy algorithms using default parameters on the hu.MAP network dataset.

Of these, CoAch and Graph Entropy are algorithms designed specifically for the protein complex detection task, whereas Angel is a more recently published, generalized, and efficient node-centric approach [20][21][22].

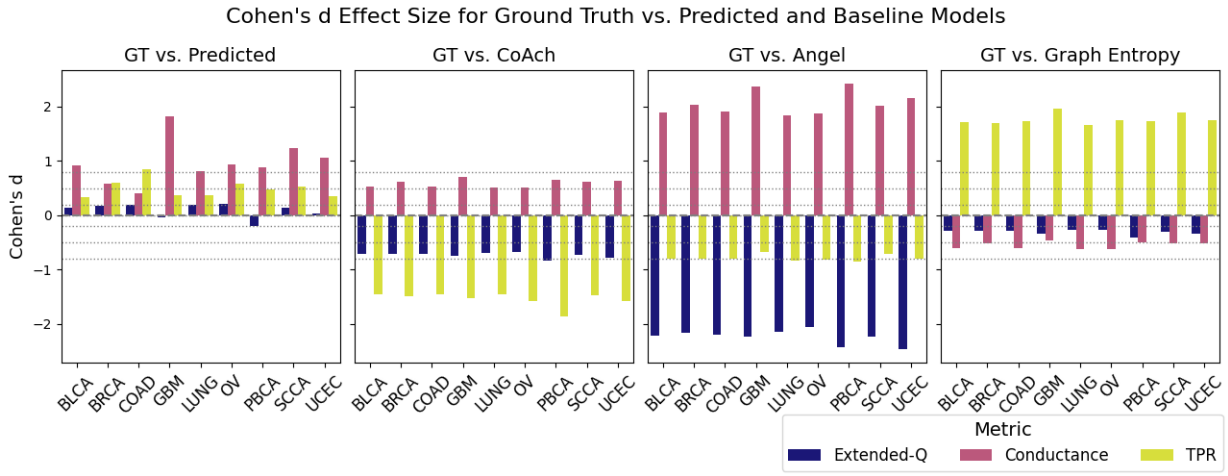


Figure 3: Cohen's d effect size for ground truth vs. predicted and baseline communities. Dotted lines at $x = \pm 0.2$, $x = \pm 0.5$, $x = \pm 0.8$ indicate general benchmarks for small, moderate, and large effect sizes, respectively.

To compare the difference between mean metrics, we then calculate Cohen's d effect size for ground truth and each of the predicted, CoAch, Angel, and Graph Entropy community output metrics [Fig. 3]. Effect sizes for Extended-Q are lowest when comparing ground truth vs. predicted communities by our semi-supervised GAT model, with absolute values ranging from 0.025 to 0.20 across the nine cancer datasets. When comparing ground truth communities to CoAch, Angel, and Graph Entropy community outputs, Extended-Q effect sizes were moderate to high, high, and low to moderate, respectively. Additionally, we note that *GT vs. Predicted* was the only case in which Extended-Q shows positive Cohen's d values, indicating a slight decrease in modularity from the ground truth to the predicted communities.

Effect sizes for conductance and TPR, on the other hand, range from moderate to high when comparing ground truth vs. predicted communities. All Cohen's d values are positive and indicate a moderate to high decrease in conductance and TPR between ground truth and predicted community means. As lower conductance is reflective of more isolated, well-defined communities, we note that predicted communities exhibit more distinct groupings. On the

other hand, though Cohen's d TPR values are similarly positive for the nine datasets, a decrease in TPR is reflective of a lower triangle density and more loosely connected nodes in predicted communities.

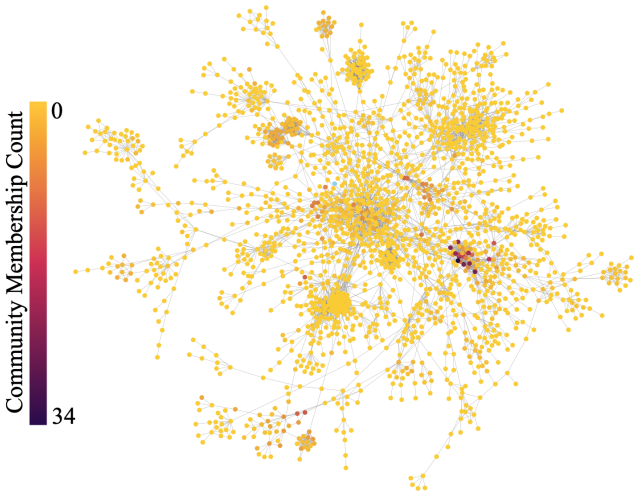


Figure 4: Ground truth community membership graph for BLCA dataset.

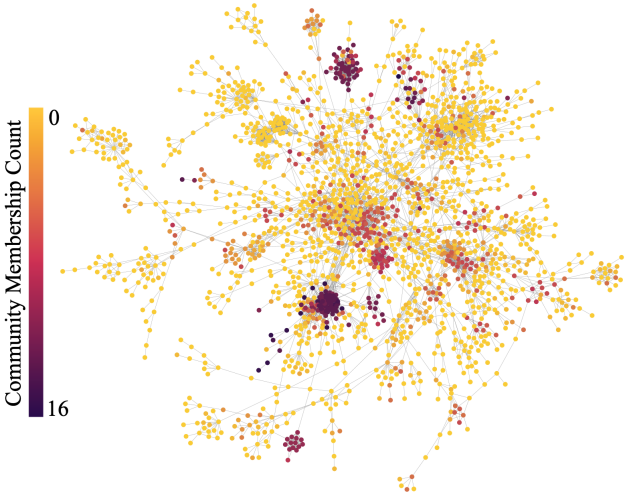


Figure 5: Predicted community membership graph for BLCA dataset.

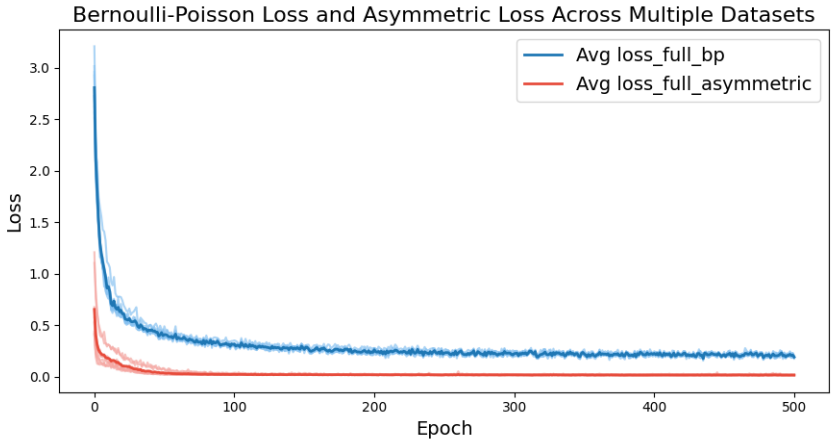


Figure 6: Comparing bernoulli-poisson loss vs. asymmetric loss across all 9 datasets used. Individual dataset loss is shown overlaid in lighter blue and red for BP loss and ASL loss, respectively.

Model training was stable, and both bernoulli-poisson (BP) and asymmetric (ASL) loss converged [Fig. 6]. Loss curves were similar for all 9 datasets trained on, though we did see

higher initial ASL loss and slower convergence for datasets with fewer ground truth labels (i.e. GBM, PBCA). This is unsurprising, as we would expect more prediction inaccuracies at the outset when restricting to a limited set of labels.

Community membership visualization in Cytoscape indicated comparable community membership count between ground truth and predicted complexes for the BLCA dataset [Fig.4][Fig.5]. Moreover, we see higher community density in more diverse areas of the prediction network, suggesting that the model has not overfit to the ground truth dataset.

4 Discussion

Evaluation of structural graph metrics indicates that the semi-supervised GAT model produced communities with comparable modularity, but with lower conductance and triangle participation ratio (TPR) relative to ground truth datasets.

Interestingly, both the ground truth and the predicted communities exhibit relatively low modularity scores. This suggests that, from a modularity perspective, the observed community structure is only marginally better than random. In contrast, baseline algorithms such as CoAch and Angel achieve much higher modularity scores, indicating well-separated community structures. However, these high-modularity communities fail to align with the ground truth, revealing a disconnect between structural quality and biological relevance. Our goal is not to maximize modularity but to reflect biologically relevant groupings. The low ground truth and prediction values are expected given the inherent overlap and biological complexity of cancer-related protein interactions.

We see a similar pattern emerge for conductance and TPR of the CoAch and Angel baselines in comparison to ground truth. Both CoAch and Angel community outputs are characterized by lower conductance and higher TPR, indicating well-formed predictions with higher triangle density and proportion of inter-community edges. Unlike modularity, however, predictions of the semi-supervised GAT model show similarly large effect sizes, albeit in the opposite direction for TPR. Also of note is that Graph Entropy also exhibits low effect size for modularity, though predicted communities exhibit a large decrease in TPR, indicating communities that are neither modular nor triangle-dense.

4.1 Limitations and Future Work

Here, we highlight a fundamental limitation of this evaluation approach. We assume that community structure metrics are reflective of biologically meaningful protein complexes. To robustly evaluate the model's output communities, we need to empirically validate the predicted complexes to assess their biological significance. We do not have the experimental resources to evaluate whether predicted communities truly represent differentially expressed protein complexes by cancer type. However, this may be a possible consideration in future work.

Another limitation of the current model choice is the fixed number of communities C. Although we may choose to run the GAT model with multiple different output community values, this is not reflective of real world scenarios as the number of total communities C is not predetermined. To address this issue, Novel Class Discovery (NCD) related models may be an option, though further investigation is needed.

5 Conclusion

In this study, we propose a method to identify differentially expressed protein complexes by cancer type using a semi-supervised GAT model with combined bernoulli-poisson and asymmetric losses. Multiple sources were processed to create nine ground truth cancer datasets to be used for model training. Results indicate predicted communities with similar modularity, lower conductance, and lower TPR in comparison to ground truth communities, and visualizations suggest generalization across the entire hu.MAP network used for training. Although further parameter tuning and biological validation are needed, results suggest that combining the Bernoulli-Poisson model with an asymmetric loss may offer a possible framework for introducing supervision into a probabilistic model that has already proven successful in community detection tasks.

6 Acknowledgments

First and foremost, I would like to express my gratitude to my advisor, Dr. Uzun, for his unwavering support throughout my time at Brown as undergraduate student. His guidance and encouragement, from taking me on as a first-year student, to encouraging me to pursue honors in computer science, has been truly invaluable. I genuinely could not have asked for a better mentor.

I also thank my reader, Dr. Singh, for her time and absolutely incredible teaching. Deep Learning in Genomics was — without a doubt — one of the best classes I have taken at Brown, and this project would never have taken shape without it.

References

[1] G. Tsitsiridis, R. Steinkamp, M. Giurgiu, *et al.*, "CORUM: The comprehensive resource of mammalian protein complexes–2022," *Nucleic Acids Research*, vol. 51, pp. D539–D545, D1 Jan. 6, 2023, ISSN: 0305-1048. DOI: 10.1093/nar/gkac1015. [Online]. Available: https://doi.org/10.1093/nar/gkac1015 (visited on 04/30/2025).

[2] D. Lv, D. Li, Y. Cai, *et al.*, "CancerProteome: A resource to functionally decipher the proteome landscape in cancer," *Nucleic Acids Research*, vol. 52, pp. D1155–D1162, D1 Jan. 5, 2024, ISSN: 0305-1048. DOI: 10.1093/nar/gkad824. [Online]. Available: https://doi.org/10.1093/nar/gkad824 (visited on 04/30/2025).

[3] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, Feb. 1, 2010, ISSN: 0370-1573. DOI: 10.1016/j.physrep.2009.11.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0370157309002841 (visited on 04/30/2025).

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, P10008, Oct. 2008, ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/p10008. [Online]. Available: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

[5] V. A. Traag, L. Waltman, and N. J. van Eck, "From louvain to leiden: Guaranteeing well-connected communities," *Scientific Reports*, vol. 9, no. 1, Mar. 2019, ISSN: 2045-2322. DOI: 10.1038/s41598-019-41695-z. [Online]. Available: http://dx.doi.org/10.1038/s41598-019-41695-z.

[6] X. Su, S. Xue, F. Liu, *et al.*, "A comprehensive survey on community detection with deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 4, pp. 4682–4702, Apr. 2024, ISSN: 2162-2388. DOI: 10.1109/TNNLS.2021.3137396. [Online]. Available: https://ieeexplore.ieee.org/document/9732192 (visited on 04/30/2025).

[7] F. Liu, S. Xue, J. Wu, *et al.*, "Deep learning for community detection: Progress, challenges and opportunities," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, ser. IJCAI-PRICAI-2020, International Joint Conferences on Artificial Intelligence Organization, Jul. 2020, 4981–4987. DOI: 10.24963/

ijcai.2020/693. [Online]. Available: http://dx.doi.org/10.24963/ijcai.2020/693.

[8] F. Soleymani, E. Paquet, H. Viktor, W. Michalowski, and D. Spinello, "Protein–protein interaction prediction with deep learning: A comprehensive review," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 5316–5341, Jan. 1, 2022, ISSN: 2001-0370. DOI: 10.1016/j.csbj.2022.08.070. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037022004044 (visited on 04/30/2025).

[9] N. Zaki, H. Singh, and E. A. Mohamed, "Identifying protein complexes in protein-protein interaction data using graph convolutional network," *IEEE Access*, vol. 9, pp. 123 717–123 726, 2021, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3110845. [Online]. Available: https://ieeexplore.ieee.org/document/9530526 (visited on 04/30/2025).

[10] P. Zhou, Y. Zhang, Z. Li, K. Pang, and D. Zhao, "Protein complex identification based on heterogeneous protein information network," *Journal of Computational Biology*, vol. 30, no. 9, pp. 985–998, Sep. 2023, Publisher: Mary Ann Liebert, Inc., publishers. DOI: 10.1089/cmb.2023.0081. [Online]. Available: https://www.liebertpub.com/doi/10.1089/cmb.2023.0081 (visited on 04/30/2025).

[11] J. Zhu, Z. Zheng, M. Yang, G. P. C. Fung, and C. Huang, "Protein complexes detection based on semi-supervised network embedding model," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 2, pp. 797–803, Mar. 2021, ISSN: 1557-9964. DOI: 10.1109/TCBB.2019.2944809. [Online]. Available: https://ieeexplore.ieee.org/document/8855005 (visited on 04/30/2025).

[12] O. Shchur and S. Günnemann, *Overlapping community detection with graph neural networks*, Sep. 26, 2019. DOI: 10.48550/arXiv.1909.12201. arXiv: 1909.12201[cs]. [Online]. Available: http://arxiv.org/abs/1909.12201 (visited on 04/30/2025).

[13] K. Drew, J. B. Wallingford, and E. M. Marcotte, "Hu.MAP 2.0: Integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies," *Molecular Systems Biology*, vol. 17, no. 5, e10016, May 2021, Publisher: John Wiley & Sons, Ltd, ISSN: 1744-4292. DOI: 10.15252/msb.202010016. [Online]. Available: https://www.embopress.org/doi/full/10.15252/msb.202010016 (visited on 04/30/2025).

[14] K. Sismanis, P. Potikas, D. Souliou, and A. Pagourtzis, "Overlapping community detection using graph attention networks," *Future Generation Computer Systems*, vol. 163, p. 107 529, Feb. 1, 2025, ISSN: 0167-739X. DOI: 10.1016/j.future.2024.107529. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X2400493X (visited on 04/30/2025).

[15] T. Ridnik, E. Ben-Baruch, N. Zamir, *et al.*, "Asymmetric loss for multi-label classification,"

[16] A. Lázár, D. Ábel, and T. Vicsek, "Modularity measure of networks with overlapping communities," *EPL (Europhysics Letters)*, vol. 90, no. 1, p. 18 001, Apr. 2010, ISSN: 1286-4854. DOI: 10.1209/0295-5075/90/18001. [Online]. Available: http://dx.doi.org/10.1209/0295-5075/90/18001.

[17] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000. DOI: 10.1109/34.868688.

[18] J. Mothe, K. Mkhitaryan, and M. Haroutunian, "Community detection: Comparison of state of the art algorithms," in *2017 Computer Science and Information Technologies (CSIT)*, Sep. 2017, pp. 125–129. DOI: 10.1109/CSITechnol.2017.8312155. [Online]. Available: https://ieeexplore.ieee.org/document/8312155 (visited on 04/30/2025).

[19] J. Yang and J. Leskovec, *Defining and evaluating network communities based on ground-truth*, 2012. arXiv: 1205.6233 [cs.SI]. [Online]. Available: https://arxiv.org/abs/1205.6233.

[20] M. Wu, X. Li, C.-K. Kwoh, and S.-K. Ng, "A core-attachment based method to detect protein complexes in PPI networks," *BMC Bioinformatics*, vol. 10, no. 1, p. 169, Jun. 2, 2009, ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-169. [Online]. Available: https://doi.org/10.1186/1471-2105-10-169.

[21] G. Rossetti, "Exorcising the demon: Angel, efficient node-centric community discovery," in *Complex Networks and Their Applications VIII*, H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, and L. M. Rocha, Eds., Cham: Springer International Publishing, 2020, pp. 152–163, ISBN: 978-3-030-36687-2.

[22] E. C. Kenley and Y.-R. Cho, "Detecting protein complexes and functional modules from protein interaction networks: A graph entropy approach," *PROTEOMICS*, vol. 11, no. 19, pp. 3835–3844, 2011, ISSN: 1615-9861. DOI: 10.1002/pmic.201100193. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.201100193 (visited on 04/30/2025).