# Exploring the building blocks of cell organization as high-order network motifs with graph isomorphism network

**Yang Yu**
Department of Electrical Engineering and Computer Science,
Christopher S. Bond Life Sciences Center
University of Missouri
Columbia, MO 65211, USA
yykk3@missouri.edu

**Shuang Wang**
Department of Computer Science,
Luddy School of Informatics, Computing, and Engineering
Indiana University Bloomington
Bloomington, IN 47405, USA
sw152@iu.edu

**Dong Xu***
Department of Electrical Engineering and Computer Science,
Christopher S. Bond Life Sciences Center
University of Missouri
Columbia, MO 65211, USA
xudong@missouri.edu

**Juexin Wang***
Department of BioHealth Informatics,
Luddy School of Informatics, Computing, and Engineering
Indiana University Indianapolis
Indianapolis, IN 46202, USA
wangjuex@iu.edu

## Abstract

The spatial arrangement of cells within tissues plays a pivotal role in shaping tissue function. A critical spatial pattern is network motif as cell organization. Network motifs can be represented as recurring significant interconnections in a spatial cell-relation graph, i.e., the occurrences of isomorphic subgraphs in the graph, which is computationally infeasible to have an optimal solution with high-order (>3 nodes) subgraphs. We introduce Triangulation Network Motif Neural Network (TrimNN), a neural network-based approach designed to estimate the prevalence of network motifs of any order in a triangulated cell graph. TrimNN simplifies the intricate task of occurrence regression by decomposing it into several binary present/absent predictions on small graphs. TrimNN is trained using representative pairs of predefined subgraphs and triangulated cell graphs to estimate overrepresented network motifs. On typical spatial omics samples within thousands of cells in dozens of cell types, TrimNN robustly infers high-order network motifs in seconds. TrimNN

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

provides an accurate, efficient, and robust approach for quantifying network motifs, which helps pave the way to disclose the biological mechanisms underlying cell organization in multicellular differentiation, development, and disease progression.

# 1 Background

Deciphering the relationship between structure and function in tissues is the cornerstone of tissue biology and pathology[1]. With advancements in spatial omics, such as spatially resolved transcriptomics[2] and proteomics[3], researchers have access to an unprecedented resource to explore how distinct cell types are organized to perform specialized roles at the cellular level[4]. However, identifying the building blocks of the cell organization and determining which spatial cellular interconnection patterns are informative to tissue function remain challenging[5].

Network motifs as recurring significant interconnections represent network characteristics as conservative patterns[6]. The studies of network motif have greatly enhanced the knowledge of network functions in social networks[7] and biological networks[8]. We hypothesize network motifs can be treated as building blocks of cell organization that invariantly across different samples, and they connect with key functions in a biologically meaningful context. Currently, most existing network motif analyses are limited to 1-3 orders[9]. However, in spatial omics studies, biologists have observed the prevalence of high-order network motifs significantly correlated with patient survival in colorectal cancer[10], brain tumor[11], and lung cancer[12].

The biological problem of identifying overrepresented network motifs can be modeled mathematically by identifying the most overrepresented subgraphs. This problem usually consists of two subproblems: subgraph matching[13] and pattern growth[8]. It is proven that subgraph matching is NP-complete[14], which makes it computationally infeasible to count high-order isomorphic subgraphs in a polynomial time. Even though many methods adopted heuristic strategies such as edge sampling (e.g., MFinder[15]), node sampling (e.g., FANMOD[16]), and global pruning (e.g., Ullmann[17], VF2[18]) to address this challenge, their practical utility remains limited due to the scalability issue. Neural Subgraph Isomorphism Counting (NSIC)[19] is the first deep learning model to predict subgraph occurrences, but its far-reaching goal on universal graphs and its limited accuracy narrow its practical utility.

Here, we present Triangulation Network Motif Neural Network (TrimNN), a neural network-based approach to estimate the prevalence of network motifs of any order in a graph. TrimNN aims to address the subgraph matching problem in triangulated graphs after Delaunay triangulation derived from spatial omics. TrimNN decomposes the occurrence regression challenge into several binary classification problems modeled by the sub-TrimNN module. Inspired by NSIC, TrimNN is trained on representative pairs of the predefined subgraphs and the triangulated cell graphs. TrimNN aggregates the sub-TrimNN module's results and outputs the subgraph's relative abundance, which can be used to estimate the prevalence of overrepresented network motifs.

Our major contribution is formulating and simplifying the subgraph matching challenge in the context of spatial omics. Avoiding predicting the absolute occurrences of network motifs in the universal graphs in the original setting, TrimNN decomposes the challenge to a serial binary classification problem in a well-defined set of biological meaningful triangulated graphs, where it performs binary presence/absence predictions at a similar scale.TrimNN is publicly available at https://github.com/yuyang-0825/TrimNN.

# 2 Methods

## 2.1 Problem setting and definition

Formally, we define the k-order subgraph with $k$ nodes as $M_k$, the triangulated graph as $G = \{V, E\}$. The biological problem of identifying the overrepresented network motifs can be modeled mathematically in finding the most overrepresented subgraph $M_k^*$ in $G$, where $M_k^* \in M_k$, and $M_k \in G$. This challenge consists of a subgraph matching problem and a pattern growth problem built on it. The whole workflow is shown in Figure 1.

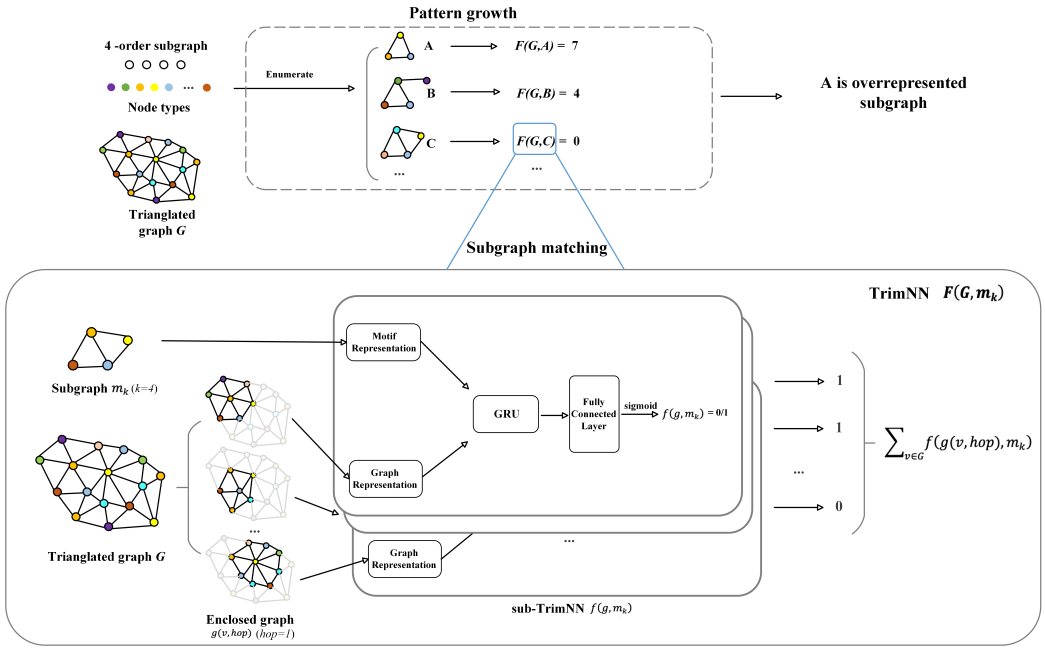


Figure 1: Flowchart of the motif identification problem and the TrimNN framework. In this figure, a 4-order subgraph is taken as an example.

The goal of TrimNN is the subgraph matching problem, which aims to define $F(G, m_k) \in N$, estimating the relative occurrence of the given $m_k$ in $G$. The problem can be quasi-divided and conquered by summarization of many sub-TrimNN problems. The goal of sub-TrimNN is to build a reliable binary prediction model $f(g, m_k) \in [0, 1]$, where 0 presents $m_k$ is absent in $g$, and 1 represents presence. Here $g \in G$ and $g$ is in a similar scale of $m_k$. With sub-TrimNN on enclosed graphs of each node, TrimNN is the summarization of results from all sub-TrimNN in the graph. Finally, $F(G, m_k) = \sum_{v \in G} f(g(v, hop), m_k)$ , where $g(v, hop)$ is the enclosed graph as the neighborhoods of node $v \in V$ with $hop \in [1, 2, 3, \dots]$, and $g(v, hop) \in G$. After we get a fast and reliable $F(G, m_k)$ from TrimNN, we can use it in the problem of pattern growth. Using enumeration or other searching processes, the final target top overrepresented set $M_k^*$ has the maximum relative abundance identified by $F(G, m_k)$, where $F(G, M_k^*) = \max(F(G, M_k))$.

### 2.2 TrimNN model architecture

We decompose the regression problem of TrimNN to many binary classification problems solved by sub-TrimNN. The input of sub-TrimNN is a pair of subgraph $m_k$ and the triangulated graph g, both of which can be extracted and learned by Graph Isomorphism Network (GIN)[20]. Then this pair of representations are aligned in the interaction module, which consists of gated recurrent units (GRU) with dynamic memory. After fully connected layers and activated by the sigmoid function, sub-TrimNN outputs the binary predictions (presence/absence). The training process minimizes the loss function on cross-entropy of known presence/absence relations. After trained sub-TrimNN $f(g, m_k)$, TrimNN estimates the abundances of $F(G, m_k)$ by summerizting sub-TrimNN predictions on each node's enclosed graph.

### 2.3 Constructing the training set as pairs between subgraphs and triangulated graphs

We simulated the training set on known presence/absence relations of pairs between subgraphs and triangulated graphs. 8 distinct subgraphs in various topologies were generated, including 3-order and high-order subgraphs up to 9-order. Given the context of routine spatial omics for each network motif, we constructed the corresponding triangulated graphs with varying node sizes of 16, 32, 64, and 128, and node types of 8, 16, and 32. To preserve the diversity, we generated 50 extended subgraphs permutating node types for each subgraph, and generated corresponding 1,000 triangulated graphs

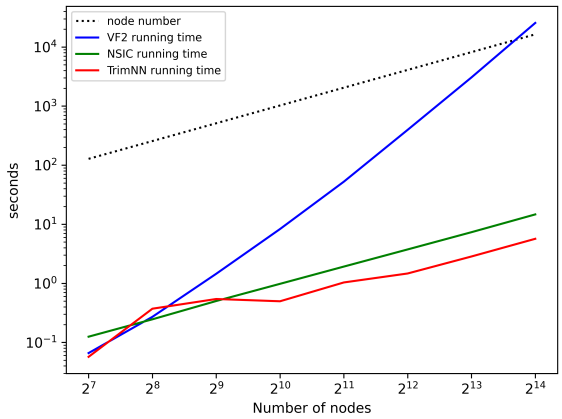


Figure 2: Comparison in scalability between TrimNN and competitors.

permutating node types. We controlled the proportion of positive to negative samples at 1:1 in data generation and split the generated data into training, validation, and test sets in a ratio of 8:1:1.

# 3 Results

## 3.1 TrimNN accurately identifies whether the network motifs present in the triangulated graph

We first tested the performance of TrimNN on a modified task of subgraph matching, which predicts whether the network motif existed in the triangulated graph by sub-TrimNN. As a binary prediction task, the performance is evaluated by precision, recall, F1 score, and MCC on the generated test set varying size and node types. For NSIC regresses occurrence as integers, we treat NSIC's prediction on 0 as not existing in the graph, and any value equal or larger than one as existing in the graph. We did not include VF2 here because it enumerates with huge computational cost. We observed that TrimNN outperforms competitors' bias in nearly all the scenarios (Table 1). Notably, we value the near-perfect performance on recall criteria, indicating TrimNN is confident when it predicts the network motif presented in the triangulated graph.

Table 1: Performances of TrimNN and NSIC on test data

| Network motif | Cell type | Number of nodes in the graph | TrimNN | | | | NSIC | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Precision | Recall | F1 | MCC | Precision | Recall | F1 | MCC |
| 3-order | 8 | 64 | **0.8499** | 0.9755 | **0.9084** | **0.8126** | 0.5018 | **0.9842** | 0.6647 | 0.0387 |
| | | 128 | **0.795** | 0.9381 | **0.8607** | **0.7075** | 0.503 | **0.9962** | 0.6685 | 0.0548 |
| | 16 | 64 | **0.8988** | 0.9917 | **0.943** | **0.885** | 0.5008 | **0.9991** | 0.6671 | 0.0563 |
| | | 128 | **0.8399** | 0.9817 | **0.9053** | **0.8054** | 0.5016 | **1** | 0.668 | 0 |
| | 32 | 64 | **0.9043** | 0.9945 | **0.9473** | **0.8938** | 0.5004 | **0.9999** | 0.667 | 0.0355 |
| | | 128 | **0.8485** | 0.9891 | **0.9134** | **0.8242** | 0.499 | **1** | 0.6658 | 0 |
| 4-order | 8 | 64 | **0.8655** | 0.9688 | **0.9142** | **0.8241** | 0.5015 | **1** | 0.668 | 0.0553 |
| | | 128 | **0.7898** | 0.9332 | **0.8555** | **0.6984** | 0.4973 | **1** | 0.6642 | 0.0122 |
| | 16 | 64 | **0.8936** | 0.9895 | **0.9391** | **0.8773** | 0.502 | **0.9992** | 0.6683 | 0.0878 |
| | | 128 | **0.8342** | 0.9781 | **0.9004** | **0.7963** | 0.5013 | **0.9994** | 0.6677 | 0.0697 |
| | 32 | 64 | **0.9141** | 0.9958 | **0.9532** | **0.9061** | 0.4985 | **1** | 0.6653 | 0.007 |
| | | 128 | **0.8559** | 0.9892 | **0.9177** | **0.8331** | 0.4994 | **1** | 0.6661 | 0 |

## 3.2 TrimNN accurately identifies top overrepresented network motifs

Then we tested whether TrimNN identifies the correct overrepresented network motifs in the triangulated graph as a pattern growth problem. As we only care about the biological meaningful top overrepresented network motifs, we used criteria of Mean Average Recall at K (MAR@K) and

Mean Squared error (MSE) to evaluate the performances. As in Section 3.1, we excluded VF2 for it enumerates to generate exact results. The performances of TrimNN and NSIC are not shown due to the page limit. We can see that TrimNN outperforms competitors in nearly all the scenarios again with a large margin.

### 3.3 TrimNN is highly scalable in identifying high-order network motifs

We compared the computational time on triangulated graphs varying different node sizes. In the experiments, the inquiry subgraph contains 9 nodes, both the subgraph and the triangulated graph have 32 node types. All the experiments were performed on a workstation equipped with AMD EPYC 7713 CPU and one NVIDIA A100 GPU. Figure 2 shows that both TrimNN (red line) and NSIC (green line) exhibit linear scalability with increasing node sizes (black dot line), while TrimNN continuously consumes even lower computational time when the graph size is in the scale of typical spatial omics samples (>512 nodes). In contrast, the VF2 method (blue line) grows exponentially. When the number of graph nodes exceeds 10k, VF2's runtime surpasses 20k seconds, making it unacceptable in most scenarios.

## 4 Discussion

The advent of spatial omics has revolutionized our capacity to explore the nuanced organization of cells within tissues at the cellular level. Based on graph isomorphism network, TrimNN provides an accurate, unbiased, efficient, and robust approach to quantify the network motifs as interpretable building blocks of cell organization. In the future, the identified enriched network motifs will be evaluated with downstream analysis using biological validations[21] from multiple independent data sources.

## References


[1] Giovanni Palla, David S Fischer, Aviv Regev, and Fabian J Theis. Spatial components of molecular tissue biology. *Nature Biotechnology*, 40(3):308–318, 2022.

[2] Anjali Rao, Dalia Barkley, Gustavo S França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021.

[3] Emma Lundberg and Georg HH Borner. Spatial proteomics: a powerful discovery tool for cell biology. *Nature Reviews Molecular Cell Biology*, 20(5):285–302, 2019.

[4] Dario Bressan, Giorgia Battistoni, and Gregory J Hannon. The dawn of spatial omics. *Science*, 381(6657):eabq4964, 2023.

[5] Salil S Bhate, Graham L Barlow, Christian M Schürch, and Garry P Nolan. Tissue schematics map the specialization of immune tissue motifs and their appropriation by tumors. *Cell Systems*, 13(2):109–130, 2022.

[6] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[7] Lewi Stone, Daniel Simberloff, and Yael Artzy-Randrup. Network motifs and their origins. *PLoS computational biology*, 15(4):e1006749, 2019.

[8] Sabyasachi Patra and Anjali Mohapatra. Review of tools and algorithms for network motif discovery in biological networks. *IET systems biology*, 14(4):171–189, 2020.

[9] Elisabeth Wong, Brittany Baur, Saad Quader, and Chun-Hsi Huang. Biological network motif detection: principles and practice. *Briefings in bioinformatics*, 13(2):202–215, 2012.

[10] Christian M Schürch, Salil S Bhate, Graham L Barlow, Darci J Phillips, Luca Noti, Inti Zlobec, Pauline Chu, Sarah Black, Janos Demeter, David R McIlwain, et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell*, 182(5):1341–1359, 2020.

[11] Elham Karimi, Miranda W Yu, Sarah M Maritan, Lucas JM Perus, Morteza Rezanejad, Mark Sorin, Matthew Dankner, Parvaneh Fallah, Samuel Doré, Dongmei Zuo, et al. Single-cell spatial immune landscapes of primary and metastatic brain tumours. *Nature*, 614(7948):555–563, 2023.

[12] Mark Sorin, Elham Karimi, Morteza Rezanejad, W Yu Miranda, Lysanne Desharnais, Sheri AC McDowell, Samuel Doré, Azadeh Arabzadeh, Valerie Breton, Benoit Fiset, et al. Single-cell spatial landscape of immunotherapy response reveals mechanisms of cxcl13 enhanced antitumor immunity. *Journal for Immunotherapy of Cancer*, 11(2), 2023.

[13] Zhao Sun, Hongzhi Wang, Haixun Wang, Bin Shao, and Jianzhong Li. Efficient subgraph matching on billion node graphs. *arXiv preprint arXiv:1205.6691*, 2012.

[14] Michael R Garey and David S Johnson. Computers and intractability. *A Guide to the*, 1979.

[15] Nadav Kashtan, Shalev Itzkovitz, Ron Milo, and Uri Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, 2004.

[16] Sebastian Wernicke and Florian Rasche. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006.

[17] Julian R Ullmann. An algorithm for subgraph isomorphism. *Journal of the ACM (JACM)*, 23(1):31–42, 1976.

[18] Luigi P Cordella, Pasquale Foggia, Carlo Sansone, and Mario Vento. A (sub) graph isomorphism algorithm for matching large graphs. *IEEE transactions on pattern analysis and machine intelligence*, 26(10):1367–1372, 2004.

[19] Xin Liu, Haojie Pan, Mutian He, Yangqiu Song, Xin Jiang, and Lifeng Shang. Neural subgraph isomorphism counting. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1959–1969, 2020.

[20] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[21] Shimrit Mayer, Tomer Milo, Achinoam Isaacson, Coral Halperin, Shoval Miyara, Yaniv Stein, Chen Lior, Meirav Pevsner-Fischer, Eldad Tzahor, Avi Mayo, et al. The tumor microenvironment shows a hierarchy of cell-cell interactions dominated by fibroblasts. *Nature Communications*, 14(1):5810, 2023.