Pattern Discovery in Multilayer Networks

Yuanfang Ren, Aisharjya Sarkar, Pierangelo Veltri, Ahmet Ay, Alin Dobra, and Tamer Kahveci

Abstract—*Motivation:* In bioinformatics, complex cellular modeling and behavior simulation to identify significant molecular interactions is considered a relevant problem. Traditional methods model such complex systems using single and binary network. However, this model is inadequate to represent biological networks as different sets of interactions can simultaneously take place for different interaction constraints (such as transcription regulation and protein interaction). Furthermore, biological systems may exhibit varying interaction topologies even for the same interaction type under different developmental stages or stress conditions. Therefore, models which consider biological systems as solitary interactions are inaccurate as they fail to capture the complex behavior of cellular interactions within organisms. Identification and counting of recurrent motifs within a network is one of the fundamental problems in biological network analysis. Existing methods for motif counting on single network topologies are inadequate to capture patterns of molecular interactions that have significant changes in biological expression when identified across different organisms that are similar, or even time-varying networks within the same organism. That is, they fail to identify recurrent interactions as they consider a single snapshot of a network among a set of multiple networks. Therefore, we need methods geared towards studying multiple network topologies and the pattern conservation among them. *Contributions:* In this paper, we consider the problem of counting the number of instances of a user supplied motif topology in a given multilayer network. We model interactions among a set of entities (e.g., genes) describing various conditions or temporal variation as multilayer networks. Thus a separate network as each layer shows the connectivity of the nodes under a unique network state. Existing motif counting and identification methods are limited to single network topologies, and thus cannot be directly applied on multilayer networks. We apply our model and algorithm to study frequent patterns in cellular networks that are common in varying cellular states under different stress conditions, where the cellular network topology under each stress condition describes a unique network layer. *Results:* We develop a methodology and corresponding algorithm based on the proposed model for motif counting in multilayer networks. We performed experiments on both real and synthetic datasets. We modeled the synthetic datasets under a wide spectrum of parameters, such as network size, density, motif frequency. Results on synthetic datasets demonstrate that our algorithm finds motif embeddings with very high accuracy compared to existing state-of-the-art methods such as G-tries, ESU (FANMODE) and mfinder. Furthermore, we observe that our method runs from several times to several orders of magnitude faster than existing methods. For experiments on real dataset, we consider Escherichia coli (*E. coli*) transcription regulatory network under different experimental conditions. We observe that the genes selected by our method conserves functional characteristics under various stress conditions with very low false discovery rates. Moreover, the method is scalable to real networks in terms of both network size and number of layers.

Index Terms—Multilayer networks, motif finding, biological networks

1 Introduction

Graph based models, or in general networks are used to represent and study interactions among various elements. Nodes and edges are used respectively to represent elements and interactions among those elements. For instance nodes can be molecules and edges represent interactions among them, or nodes can be proteins and edges model potential protein-protein interactions [17] well-known in literature. Depending on the underlying interaction type, biological networks can be represented using directed or undirected edges. Modeling, studying and simulating interactions is often used to represent, analyze and predict biological behaviour in the molecule-to-molecule interactions [14], [55]. Biological networks can thus be used to study and predict molecular interactions [1], [20], such as signaling pathways [51], protein functions [57] as well as drug molecular behavior in curing diseases [56], [61].

Real biological systems are often complex and thus simplifications have to be introduced in modeling. Simple instances of biological networks consist of binary networks where single edge connections among nodes encapsulate all interactions between them. However, these models are inadequate in capturing the complex cellular interactions as, a set of entities often interact with each other in patterns that may exhibit multiple types of relations. Indeed, different edges between the same couple of nodes may be necessary to capture different relations among nodes. For instance, modeling behaviors in human brain requires that neurons can be connected through different connectomes, such as synaptic, gap junction and monoamine, where each type of connection has a completely different dynamics [4]. Finally, biological networks may represent different configuration of node-to-

- *Yuanfang Ren, Aisharjya Sarkar, Alin Dobra, and Tamer Kahveci are with the Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611 USA. E-mail: {yuanfang, sarkar, adobra, tamer}@cise.ufl.edu.*
- *Pierangelo Veltri is with the Department of Surgical and Medical Science, University Magna Graecia of Catanzaro, 88100 Catanzaro, Italy. E-mail: veltri@unicz.it.*
- *Ahmet Ay is with the Departments of Biology and Mathematics, Colgate University, Hamilton, NY 13346 USA. E-mail: aay@colgate.edu.*

Manuscript received 27 Sept. 2020; revised 5 July 2021; accepted 8 Aug. 2021. Date of publication 16 Aug. 2021; date of current version 1 Apr. 2022. This work was partially supported by NSF under Award number 2111679. (Corresponding author: Yuanfang Ren.) Digital Object Identifier no. 10.1109/TCBB.2021.3105001

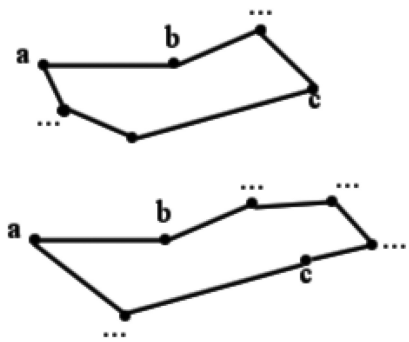

Fig. 1. An example of topology variation within biological networks. The above network shows an initial set of nodes and interactions between those nodes before external influence. In lower network (after external stimuli) the node b is now at a distance of 4 from node c, while distance among node a and b remains unchanged.

node interactions in situation where different topological configuration among nodes may influence the temporal-varying configuration. Fig. 1 reports such an example. In this figure, let us consider a network which contains a node a that interacts with a node b. Again, b is at a distance 2 from another node c (i.e., the shortest path from b to c contains two edges). In the presence of a stimulation received by external (environmental) conditions, the topology of this network may alter by rewiring a subset of interactions within the network. This may change the distance between b and c (for instance, distance equal to 4) while that between a and b remains unchanged. In order to study such networks with altering topologies under varying external factors, we need to enrich the biological modeling by means of using different instances of biological networks. Therefore, we focus on using multiple layers where each layer represent a network instance. We call such a system as *multilayer network*, where each node appears in a set of layers and each edge (interaction) may or may not be observed in a subset of these layers.

Fig. 2 depicts a hypothetical multilayer network having 10 nodes and 3 layers. Each layer represents a network defined on a set of nodes and edges representing relationships between those nodes. Moreover, these layers can interact among themselves for instance, by conserving part of the networks in a well defined topology. The interactions in such case for the corresponding in vivo biological model

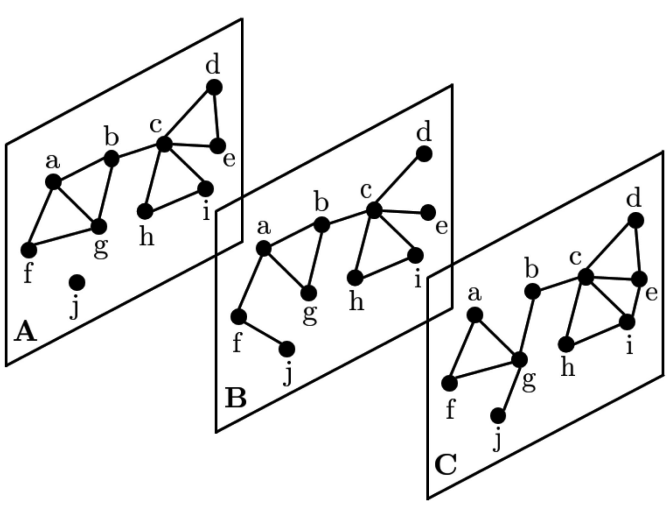

Fig. 2. A multilayer network with 3 layers and 10 nodes.

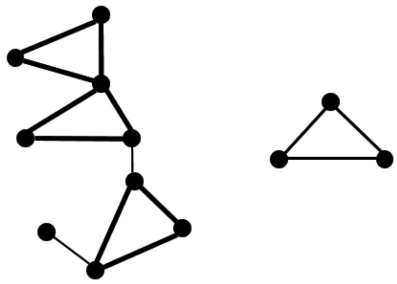

Fig. 3. Left: A hypothetical single layer network with nine nodes and 11 edge. Right: A motif in the form of triangle topology. Three instances of the triangle motif in the network are highlighted in bold.

may be represented by a subnetwork conservation (or invariant) among layers. This may represent for instance an invariant of the network in a time varying biological system evolution [12], [28], [46]. In our modelling, multilayer networks contain different layers, each node appears in all the layers. The nodes are homogeneous in type and the interactions happen only within a layer [11], [28], yet the same interaction can be observed in multiple layers (see Fig. 2).

The importance of using multilayer networks to model biological systems has been testified by several studies focused on mathematical modeling of these networks [26], [38], as well as efficient implementation of algorithms to extract information from them. Other studies report on defining network metrics by extending classical measures to multilayer networks. Such metrics are used to characterize network structures, including centrality [3], [53], clustering coefficient [10], [34], correlations [39], as well as defining operators to measure network topologies. Clustering of nodes based on similar network neighborhood within as well as hierarchically nearby layers is an important application area [62]. Another active research direction investigates various kinds of dynamical processes on multilayer topologies, such as percolation [6], diffusion processes [18], cooperation [19], synchronization [40], and information spreading [47]. An extensive review of the studies on multilayer networks can be found in [8], [28].

Motifs are connected patterns of nodes, which occur frequently in network instances [36]. They are often considered as the building blocks of the underlying biological system as they indicate robust functions executed through their specialized topologies. In classical single layer network, motif counting occurs on a network instance that may capture a single snapshot of a complex biological network. Fig. 3 shows an example of a single layer network (left) and a triangular motif (right). The network contains three instances of the triangle motif that are highlighted in bold. However, motif identification and counting on such a simplistic single layer network may fail to identify robust functional units ubiquitous across different layers, as a motif observed in one layer may not be observed in other layers. We focus on studying motifs in multilayer networks, searching for patterns repeated and present in biological networks with different layers. Motifs in biological network have been associated in different real applications, such as studying the biological transcription [52], finding genes associated with infections [58] or cancer [24], finding protein complexes [42] and host-pathogen interactions [7], or revealing relationship across species [20], [22]. Most of the existing

literature on motif identification and motif counting is limited to single layer deterministic [2], [13], [21], [25], [27], [29], [30], [31], [32], [41], [45], [59], [60], probabilistic [48], [49], [54], and dynamic [37] networks (See [44] for a survey on motifs).

Few works have been presented in a generalized definition of motifs for multilayer networks. In [4], [5], [33], the organization of edges between the same pair of nodes across all layers, is presented and defined as network motifs. Such multilayer motifs specify each layer's organization of edges among a set of nodes. However, the number of such distinct motifs grows exponentially with the number of nodes and layers. Thus, all these studies investigate motifs which contain limited number of layers and small set size of nodes. Whereas, real life biological networks often contains a large node set and considerable number of layers.

Our Contributions. Given a motif topology and a network, we say that two embeddings of this motif in the given network *overlap* if they have at least one common edge. In this paper, we solve the problem of identifying the largest set of nonoverlapping embeddings of a given motif topology in a given multilayer network, such that each embedding appears in more layers of the network than a user supplied threshold. We consider the generalized version of the motif counting problem, as we do not assume whether the network at each layer of the given multilayer network, and the motif topology is directed or undirected. The solution we propose applies to both directed and undirected networks. To simplify this concept, in the rest of the paper, we assume that the given network is directed and an undirected edge can be denoted with two edges in opposite directions. The motifs identified under our definition presents robust, independent and conserved structures within each layer as well as across different layers. We present a mathematically precise definition of motifs in multilayer networks in Section 2.2 after defining necessary terms and variables in Section 2.1. We develop a novel algorithm that counts embeddings of a given motif topology in multilayer networks. Our method first generates an aggregate network which summarizes all interactions in all layers in a single layer. Next, it locates all possible embeddings in the aggregate network. We calculate the *loss value* of each embedding, which quantifies the number of motif embeddings that cannot be selected along with the current embedding. Our method iteratively selects embeddings with the least loss until no more independent embedding can be chosen. We evaluate our methods on synthetic and real datasets. On synthetic data, our algorithm finds motif embeddings with near 100% accuracy for a wide set of network models with varying network size, density, motif frequency. Existing methods, such as G-tries, ESU (FANMODE), and mfinder fail to do that. Moreover, our method performs much more faster than existing methods. Experiments on Escherichia coli (E.coli) transcription regulatory network under different experimental conditions show that our method scales to real networks and more importantly can uncover conserved functional characteristics of genes participating in the network under various conditions with very low false discovery rates. In summary, the technical contributions of this paper are:

- We introduce the concept of independent motifs in multilayer networks.
- We propose the use of aggregate networks to adopt the motif counting for single layer networks to multi layer networks.
- We formulate a new loss function which works for multilayer networks.
- We develop a randomization strategy to compute the statistical significance of the results.

We organize the paper as follows. We formulate the problem definition and present our method in Section 2. Experimental results both on synthetic and real datasets are illustrated in Section 3. Section 4 concludes the paper.

2 Definitions and Methods

In this section, we present the formal definitions of motif and multilayer network (Section 2.1). Next, we define the independent motif counting problem (Section 2.2). We finally describe our proposed method used in experimental tests (Section 2.3).

2.1 Motif Definition and Notation

We model a network as a graph $G = (V, E)$ where V is the set of nodes and $E \subseteq V \times V$ is the set of edges connecting nodes. We define a multilayer network as a set of k layers, each containing one network containing same set of nodes, but possibly a different set of edges. Formally a multilayer network is a (k+1)-tuple $\mathcal{G} = (V, E_1, E_2, \ldots, E_k)$, where V is a set of nodes, and E_i denotes the set of edges in the ith-network contained at the ith layer of $\mathcal{G}$. We also use as notation a set of k unique labels to identify each layer. We denote such labels as $L = \{\pi_1, \pi_2, \ldots \pi_k\}$. Fig. 2 depicts a multilayer network having $k = 3$ layers, 10 nodes $V = \{a, b, c, d, e, f, g, h, i, j\}$ and three layers indicated with labels A, B, C (i.e., $L = \{\pi_1, \pi_2, \pi_3\}$, where $\pi_1 = A, \pi_2 = B, \pi_3 = C$). Without loss of generality we may use such a definition with nodes representing molecules and edges E_i representing interactions among molecules.

Given a multilayer network $\mathcal{G} = (V, E)$, we define a unique representation as a single network, and we indicate it as $\mathcal{A} = (\overline{V}, \overline{E}, \Omega)$. We call $\mathcal{A}$ an *aggregate network*. In other context, this concept is represented as superposition network [18]. The aggregate network is defined on the same set of nodes as the corresponding multilayer network, thus $\overline{V} = V$. The set of edges is the union of all the edges belonging to layers in the multilayer. In other words, an edge $(u, v) \in \overline{E}$ if and only if $\exists \pi_i \in L$ such that $(u, v) \in E_i$. i.e., if there is a layer where the edge (u, v) connect two nodes.

The function $\Omega : \overline{E} \rightarrow \mathcal{P}(L)$, where $\mathcal{P}(L)$ represents the power set of a L, shows the set of layers which contain a given edge of the aggregate network. That is $\Omega(u, v) = \{\pi_i \mid (u, v) \in E_i\}$. Fig. 4 presents the aggregate network $\mathcal{A}$ of the multilayer network in Fig. 2, where the edges are the union of edges contained in each layers, and values associated with each edge represent the layers containing it.

Given a multilayer network instance, we focus on finding (and counting) pattern of (sub)network (or graph) that can be found and repeated in different layers. We define a *motif pattern* as a connected graph $M = (V', E')$, where V' and E'

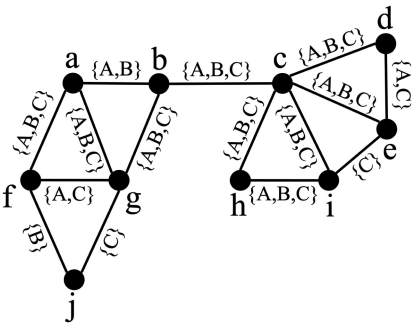

Fig. 4. The aggregate network $\mathcal{A}$ of the multilayer network in Fig. 2. The value associated with each edge represents the layers containing it.

denote the set of motif nodes and edges respectively. Fig. 5a represents an example of motif pattern.

Let $\mathcal{G}$ be a multilayer network and let M be a motif pattern. Let H_i be a set of edges, i.e., $H_i \subseteq E_j$. We say that H_i is an instance (or embedding) of M in $\mathcal{G}$, if at least one layer of the given multilayer network contains H_i, i.e., formally, (i) $\exists \pi_j \in L$, such that $H_i \subseteq E_j$, and (ii) the subnetwork defined by H_i is topologically isomorphic to M. Notice that H_i can appear in multiple layers. For example, consider the triangle pattern (see Fig. 5a) and the multilayer network in Fig. 2. The subgraph $H_i = \{(a, f), (a, g), (f,g)\}$ is an embedding of M (see Fig. 5b) and appears in the two layers denoted with A and C. We denote the set of all possible embeddings of M in $\mathcal{G}$ with $\mathcal{H}(M|\mathcal{G})$. Given an embedding $H_i \in \mathcal{H}(M|\mathcal{G})$, we indicate with $f(H_i|\mathcal{G}) = \{\pi_j | H_i \subseteq E_j\}$ the set of layers containing H_i.

Given a subset $\mathcal{H}'$ of $\mathcal{H}(M|\mathcal{G})$, it is possible to calculate the set of layers containing all embeddings in $\mathcal{H}'$ as $F(\mathcal{H}'|\mathcal{G}) = \bigcap_{H_i \in \mathcal{H}'} f(H_i|\mathcal{G})$. E.g., let us consider the multilayer network in Fig. 2 and the triangle pattern (see Fig. 5a; as shown Fig. 5, there are five embeddings of triangle pattern. With reference to the Fig. 5, let us now consider the subset of embeddings, $\mathcal{H}' = \{H_1, H_2\}$. $f(H_1|\mathcal{G}) = \{A, C\}$ and $f(H_2|\mathcal{G}) = \{A, B\}$. Thus, $F(\mathcal{H}'|\mathcal{G}) = \{A, B\} \cap \{A, C\} = \{A\}$, which implies that embeddings H_1 and H_2 appear simultaneously only in layer A.

In the literature, there are three frequency measures to count motifs: (i) $\mathcal{F}_1$; (ii) $\mathcal{F}_2$ and (iii) $\mathcal{F}_3$. $\mathcal{F}_1$ counts all possible embeddings of motifs regardless of whether they overlap with each other or not. $\mathcal{F}_2$ counts the edge disjoint embedding of motifs. $\mathcal{F}_3$ is similar but more restrictive since it counts node disjoint embeddings [13], [50]. For example, let us calculate the measures with respect to embeddings of the Fig. 5. For simplicity, assume that all the five embeddings shown in this figure appear at the same network layer. $\mathcal{F}_1$ count is 5 as there are five embeddings. $\mathcal{F}_2$ count is 3 which are H_1, H_3 and H_5. H_2 shares edge with H_1; H_4 overlap with H_3 and H_5. $\mathcal{F}_3$ count is 2 as H_1 shares nodes with H_2 and the remaining three embeddings also share nodes.

Without loss of generality, we focus on counting motifs using the $\mathcal{F}_2$ measure in the rest of this paper. Our algorithm can easily be adapted to the other two measures.

2.2 Independent Motif

Here, we focus on counting *independent* motifs in a multilayer network. Given a subset $\mathcal{H}'$ of $\mathcal{H}(M|\mathcal{G})$, we define a function $\phi()$ on $\mathcal{H}'$ as follows: $\phi(\mathcal{H}') = 1$ if no two embeddings in $\mathcal{H}'$ share an edge, and 0 otherwise. Using these notations, we formally define the independent motif counting problem in multilayer networks.

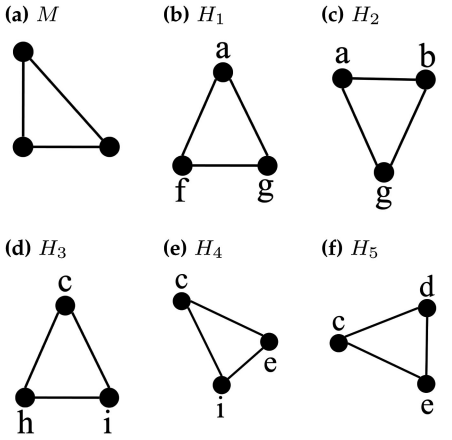

Fig. 5. (a) A motif pattern M. (b) - (f) five embeddings of M in the multilayer network (Fig. 2).

Definition 1 (Independent motif counting in multilayer networks). *Consider a multilayer network $\mathcal{G} = (V, E_1, E_2, \ldots, E_k)$ and a motif pattern M. Let $\alpha \in (0, 1]$ be a parameter, the independent motif counting in multilayer networks problem aims to find the maximum independent set of embeddings $\mathcal{H}'$ which appear in at least $\alpha \cdot k$ layers simultaneously, which is* $\arg\max_{\mathcal{H}'} |\mathcal{H}'|$ *subject to* $|F(\mathcal{H}'|\mathcal{G})| \geq \alpha \cdot k$ *and* $\phi(\mathcal{H}') = 1$.

As an example, let us consider the multilayer network $\mathcal{G}$ in Fig. 2. It has five embeddings of triangle pattern (see Fig. 5). Let us consider $\alpha = 0.6$. We want to find a set of non-overlapping motif embeddings which exist in at least two (i.e., $\lceil 0.6 \cdot 3 \rceil$) layers in $\mathcal{G}$. Consider two subsets of motif embeddings $\mathcal{H}'_1 = \{H_2, H_3\}$ $(F(\mathcal{H}'_1|\mathcal{G}) = \{A, B\})$, $\mathcal{H}'_2 = \{H_1, H_3, H_5\}$ $(F(\mathcal{H}'_2|\mathcal{G}) = \{A, C\})$ which satisfy the requirement of being non-overlapping and appearing in at least two layers. Thus, $\mathcal{H}'_2$ is the better solution among the two as it contains more embeddings than $\mathcal{H}'_1$.

Counting independent motifs in multilayer networks is a challenging task. Indeed, counting independent motifs in a single layer is a well known NP-complete problem that requires solving two NP complete problems: subgraph isomorphism problem [9] and the maximum independent set problem [16]. Counting independent motifs in multilayer networks can be done in two steps: (i) Enumerate all possible embedding sets for each layer independently, and (ii) count the number of layers containing each one. This strategy however does not scale to large networks as well as networks with many layers. The proposed algorithm is based on an heuristic to tackle this problem.

2.3 The Proposed Algorithm

To count independent motifs in multilayer networks, we propose an algorithm, (**Algorithm 1** in pseudo-code), which works as follows. It takes as input:

- a multilayer network $\mathcal{G} = (V, E_1, E_2, \ldots, E_k)$,
- a motif pattern M, and
- a value α representing the minimum motif frequency.

It consists of the following four steps:

1) It builds an aggregate network $\mathcal{A} = (V, E, \Omega)$ corresponding to $\mathcal{G}$ (Line 1);
2) Then, it finds the set of all possible embeddings $\mathcal{H}(M|\mathcal{G})$ in $\mathcal{G}$ using $\mathcal{A}$ and selects candidate embeddings denoted with $\mathcal{H}^o \subseteq \mathcal{H}(M)$ (see Lines 2-3);
3) It builds an overlap graph $\bar{\mathcal{A}}$ for $\mathcal{H}^o$, where each node corresponds to an embedding (Line 4);
4) Finally, it uses a heuristic strategy to count independent motifs (Lines 5-18).

Algorithm 1. Independent Motif Counting in Multilayer Networks

Input: Multilayer network $\mathcal{G} = (V, E_1, E_2, \ldots, E_k)$, motif pattern M and parameter $\alpha \in (0, 1]$
Output: A set S of independent motif embeddings each appearing in at least $\lceil \alpha \cdot k \rceil$ layers of $\mathcal{G}$
1: $\mathcal{A} \leftarrow$ Construct the aggregated network of $\mathcal{G}$
2: $\mathcal{H}(M|\mathcal{G}) \leftarrow$ Find all embeddings of M in $\mathcal{G}$ using $\mathcal{A}$
3: $\mathcal{H}^o \leftarrow$ Select the set of embeddings from $\mathcal{H}(M|\mathcal{G})$ that $\forall H_i \in \mathcal{H}^o, |f(H_i|\mathcal{G})| \geq \alpha \cdot k$
4: Build the overlap graph $\bar{\mathcal{A}}$ for the embedding set $\mathcal{H}^o$
5: $S \leftarrow \emptyset$; $\mathcal{L} \leftarrow L$;
6: **while** $\bar{\mathcal{A}}$ is not empty **do**
7: **for** each node corresponding embedding H_i in $\bar{\mathcal{A}}$ **do**
8: $\rho_i \leftarrow$ calculate the loss value of H_i
9: **end for**
10: $H_r \leftarrow$ select the embedding with least loss value ρ_r
11: $S \leftarrow S \cup \{H_r\}$
12: $\mathcal{L} \leftarrow \mathcal{L} \cap f(H_r|\mathcal{G})$
13: **for** each node corresponding embedding H_i in $\bar{\mathcal{A}}$ **do**
14: $f(H_i|\mathcal{G}) \leftarrow f(H_i|\mathcal{G}) \cap \mathcal{L}$
15: **end for**
16: Remove nodes corresponding to H_r, its neighbors and other embeddings with $f(H_i|\mathcal{G}) < \lceil \alpha \cdot k \rceil$ from $\mathcal{A}$
17: **end while**
18: **return** S

We now explain the steps of the algorithm in detail. Starting from the multilayer graph $\mathcal{G}$, it constructs the aggregate graph $\mathcal{A} = (V, E, \Omega)$ (see Fig. 4). Thus, the motif counting problem mapped reduces to the motif counting for a single layer. Aggregate network maintains information regarding patterns and motif as in multilayer one. Nevertheless, it can generate false positive embeddings (i.e., the aggregate network may contain an embedding of the given motif which does not appear on any layer of the multilayer network). As we explain below, we eliminate such false positives in the subsequent steps of our algorithm. It costs less to eliminate false positive instead of motif counting on multilayer network.

We locate all possible embeddings $\mathcal{H}(M|\mathcal{G})$ using $\mathcal{A}$. An embedding H_i belongs to $\mathcal{H}(M|\mathcal{G})$ if $H_i \subseteq E$ is topologically isomorphic to M and $|f(H_i|\mathcal{G})| \geq \alpha \cdot k$. We calculate $|f(H_i|\mathcal{G})|$ using the aggregate graph $\mathcal{A}$ as the cardinality of the set $\bigcap_{(u,v)\in H_i} \Omega(u, v)$'. For example, consider the aggregated network in Fig. 4. The subnetwork {(f, g), (f, j), (g, j)} is topologically isomorphic to the triangle motif pattern (Fig. 5a). It however is a false positive since all the three edges of this subnetwork does not appear in any layer at the same time (see Fig. 2). We compute this as Ω(f, g) = {A, C}, Ω(f, j) = {B}, and Ω(j, g) = {C}. Thus the intersection of these three sets yields $f() = \emptyset$. As mentioned in Definition 1, we seek to find the largest independent subset of embeddings that appear in at least $\alpha \cdot k$ layers. Thus, we include an embedding H_i to the candidate embedding set $\mathcal{H}^o$ only if $|f(H_i|\mathcal{G})| \geq \alpha \cdot k$. For example, for the multilayer network in Fig. 2, when $\alpha = 0.6$, we have $k = 3$ and thus $\lceil \alpha \cdot k \rceil = 2$. The candidate embeddings in this example are $\mathcal{H}^o = \{H_1, H_2, H_3, H_5\}$. We exclude embedding H_4 from the candidate set since it appears in only one layer.

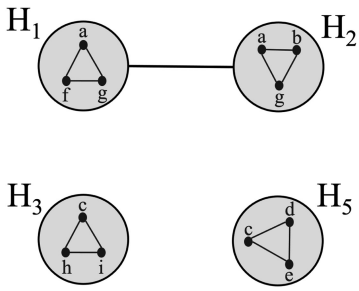

Fig. 6. The overlap graph for the embeddings $\{H_1, H_2, H_3, H_5\}$ (Fig. 5) in the multilayer network (Fig. 2). Each node in the overlap graph corresponds to an embedding of the target motif. The edge between nodes H_1 and H_2 indicates that the two embeddings share at least one edge (edge (a, g)).

Once we identify the set of all potential embeddings $\mathcal{H}^o$ using $\mathcal{A}$, we build a new graph, called *overlap graph*, denoted with $\bar{\mathcal{A}}$. In the overlap graph, each node represents an embedding $H_i \in \mathcal{H}^o$. An edge is inserted between two nodes if their corresponding embeddings share at least one edge in the given multilayer network. Fig. 6 illustrates the overlap graph for $\mathcal{H}^o = \{H_1, H_2, H_3, H_5\}$. For instance, in this figure, the node for H_1 connects with that for H_2 because H_1 and H_2 both contain the edge (a, g). Note that the concept of overlap graphs was defined earlier in the context of classic single layer [13] and probabilistic networks [48]. Here, we extend it to multilayer networks.

We use a heuristic strategy to select the subset of independent embeddings. We keep the set of layers denoted with $\mathcal{L}$ on which the resulting embeddings can appear simultaneously (see Line 5 of the Algorithm 1). This set is initialized to $\mathcal{L} = L$, which means that all layers can be the target layers on which the resulting embeddings can appear. Each embedding added to the solution set imposes a restriction on the set $\mathcal{L}$. For instance, if we include H_r in the solution, then $\mathcal{L} = \mathcal{L} \cap f(H_r|\mathcal{G})$. Thus, the set $\mathcal{L}$ monotonically gets smaller, with the increasing number of embeddings inserted in the solution set.

The relationship between the solution size (i.e., the number of frequent independent motif embeddings) and the number of layers containing that solution set is defined by a value we call *loss value*. For each embedding $H_i \in \mathcal{H}^o$, the loss value, denoted with ρ_i, is the sum of two parameters: (i) the number of neighbors of H_i in the overlap graph. This is because, each neighbor of H_i shares an edge with H_i. Therefore, including H_i in the solution set prevents us from selecting its neighbors. (ii) the potential loss for the remaining embeddings, i.e., the embeddings which do not share an edge with H_i. We calculate *potential loss value* as follows. Let us consider one embedding, which is not a neighbor of H_i

in the overlap graph, denoted with H_r. If H_i is selected in the result set and $\mathcal{L}$ is updated as explained above, only the layers in $\mathcal{L}$ which are common with $f(H_i|\mathcal{G})$ have the potential to contain H_r. Recall that the set of layers which contain H_r without this constraint is $f(H_r|\mathcal{G})$. We denote the set of layers which contain H_r under the constraint that H_i is already selected with $f'(H_r|\mathcal{G}) = f(H_r|\mathcal{G}) \cap \mathcal{L}$. If $f'(H_r|\mathcal{G}) < \lceil \alpha \cdot k \rceil$, we remove H_r and calculate its potential loss to be equal to 1. Otherwise, we calculate it as the fraction of layers which cannot add contribution to H_r (due to selection of H_i) and compute it as

$$\frac{(|f(H_r|\mathcal{G})| - |f'(H_r|\mathcal{G})|)}{|f(H_r|\mathcal{G})|}$$

Finally, we iteratively pick the embedding with least loss value (see Line 10 of the Algorithm 1), we update $\mathcal{L}$ and the layer set associated with each embedding ($f(H_r|\mathcal{G})$) (see Lines 12-15 of the Algorithm 1), and remove the corresponding node in the overlap graph along with other nodes that conflict with this selection, which consists of its neighbors and other nodes for which are contained in less than $\lceil \alpha \cdot k \rceil$ layers in $\mathcal{L}$ (see Line 16 of the Algorithm 1).

2.4 Final Discussions on the Method

Example. Let us consider the overlap graph in Fig. 6 corresponding to the multilayer network in Fig. 2, the triangle motif in Fig. 5a, and $\alpha = 0.6$. The initial target layer set is $\mathcal{L} = \{A, B, C\}$. We first calculate the loss value of H_1. If we include H_1, the target layer set is $\mathcal{L}' = \{A, C\}$ since H_1 does not occur in layer B. Including H_1 requires to remove its neighbour H_2, which increases the loss value by 1. For the other two embeddings H_3 and H_5, because of the shrink of the target layer set, H_3 still can exist but loses the one in layer B, which increases the loss value by 1/3; H_5 however has no effect. Thus, the loss value of H_1 is 1 + 1/3 + 0 = 4/3. Similar to H_1, including H_2 leads to removing H_1 and losing one H_3 in layer B. Moreover, picking H_2 also leads to removing H_5 since it does not exist in at least two layers of the target layer set $\{A, B\}$. Thus the loss value of H_2 is 1 + 1/3 + 1 = 7/3. Using the same strategy, we can calculate the loss values of H_3 and H_5 as 0 + 0 + 0 = 0 and 0 + 1 + 1/3 = 4/3 respectively. Thus, in the first iteration, we pick H_3 as it has the least loss value among the four options.

Special Case of Motif Counting. We observe a special case when $\alpha = 1$ (i.e., the motifs counted is required to appear in all the layers). In this case, the multilayer motif counting problem becomes identical to single layer motif counting. This is because the aggregate graph now contains only the edges which appear in all the layers of the network. As a result $f(H_i|G)$ becomes identical for all the H_i (that is all embeddings H_i appear in all network layers), and thus the loss function does not have the second term (i.e., potential loss).

Complexity. The construction of aggregated graph costs $\theta(\sum_i |E_i|)$ in terms of time. Finding all embeddings depends on the motif topologies. As explained in [13], the number of embeddings of a motif does not have downward closure property. In other words, the motif count does not change monotonically with increasing motif size; it depends on the network topology more than the network size. For example, for the triangle pattern, in the worst case it takes $\mathcal{O}(|V|^2)$ as the graph topology approaches to almost complete graph. The cost of finding overlap graph depends on the number of embeddings contained in the aggregate graph. It builds the overlap graph for $\mathcal{H}^o$ in $\mathcal{O}(|\mathcal{H}^o|^2)$. For the last step, we denote the final embedding set with S. Let us denote the number of edges in the given motif shape with m. The upper bound for the size of this set is $O(|S|) = O(\sum_i |E_i|)/(m \cdot \alpha \cdot k)$. This happens when all the embeddings found exist in $\geq \alpha \cdot k$ layers. As each iteration only picks one embedding, it executes $|S|$ times. The most time consuming process for each iteration is calculating the *loss value*. Since it considers all other embeddings to calculate the loss value of one embedding, this step considers $|\mathcal{H}^o|^2$ pairs of nodes in the worst case in each iteration. Thus, heuristic strategy costs $\mathcal{O}(|S||\mathcal{H}^o|^2)$ in terms of time.

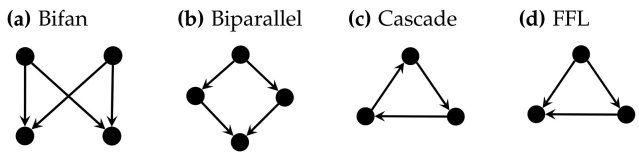

Fig. 7. Four conserved motifs studied frequently in the literature.

3 Experimental Results

We evaluate performance of our method on both synthetic and real datasets. We consider four motif topologies, namely bifan, biparallel, cascade and feed forward loop (FFL) (see Fig. 7). These four motifs are commonly studied in the literature and have been shown to be over-represented in many biologic networks [36], [52]. In the following, we describe in detail the datasets used in our experiments and implementation details.

Synthetic Dataset Description. To observe the performance of our method under controlled dataset characteristics, we perform extensive experiments on synthetically generated directed multilayer networks. To guarantee that there exists a set of independent motifs across $\alpha \cdot k$ layers, we plant a set of independent embeddings in each synthetic multilayer networks. To better understand our synthetic dataset, we first define some notations. We use the *size* and *average degree* of the multilayer network to represent the number of nodes and the number of edges per node in each layer of the multilayer network respectively, which are denoted with $|V|$ and d respectively. We also use the *edge ratio* to represent the ratio of the number of edges constituting planted motif embeddings to the total number of edges denoted with $c \in (0, 1]$. Thus, given the network size, average degree, the edge ratio, and a motif pattern with m edges, the number of motif embeddings in the multilayer network is $(|V| \cdot d \cdot c)/(2m)$. We run experiments on synthetic directed networks under four varying parameters: (i) network size, (ii) average degree, (iii) edge ratio and (iv) parameter α. For each parameter setting and each motif pattern, we construct ten multilayer networks each with ten layers as follows. We first construct a layer with $(|V| \cdot d \cdot c)/(2m)$ independent embeddings and randomly generate remaining edges. We then generate another nine layers by performing topological perturbations on it. We do this using the degree preserving edge shuffling method [35] with a given mutation rate of

TABLE 1
Real E.Coli Multilayer Networks Used in Our Experiments; Number of Nodes, Average Number of Edges Per Layer and Average Degree Per Layer

Conditions	Nodes	Edges	Degree
Cold	1373	2699	3.93
Heat	1375	2467	3.59
Oxidative	1527	3085	4.04
Lactose	1542	3204	4.15
Control	1340	2643	3.95
All conditions	1428	2820	3.95

All networks have five layers.

$\beta \in [0, 1]$. Given the edge set of a network denoted with E', a mutation rate of β means that $\beta \cdot |E|'/2$ edge pairs in the network are shuffled. In our experiment, we fix β to 0.3. Given the parameter α, in addition to the first layer containing embedding set, for each of the remaining $\lceil \alpha \cdot k \rceil - 1$ layers, we perform perturbation on the edges which do not contain planted embeddings and keep the edges of planted embeddings. Thus, we obtain $\lceil \alpha \cdot k \rceil$ layers with each containing the same set of independent motif embeddings but also having some topological difference. As for the remaining layers, we construct them by performing perturbation on the entire edge set of the first layer.

Real Dataset Description. For real dataset, we use *Escherichia coli (E.coli)* transcription regulatory network downloaded from RegulonDB Database [15], [36]. This network contains 4400 nodes and 4407 edges. We use the E.coli gene expression dataset, GSE20305, obtained from the GEO database to determine the existence of each interaction under different time points and different conditions [23]. The dataset contains five different stress conditions including cold, heat, oxidative, lactose diauxie and stationary phase (control). For each interaction under different time points (from 3 to 7) under specific condition, we include the interaction from RegulonDB between two genes if the expression level of the reactant gene is greater than a user supplied threshold. Thus, we construct five multilayer networks (one per condition) with five layers (each layer representing a time point). We also construct a network with each layer representing a specific condition; an interaction under the specific condition happens if the average gene expression of the reactant gene across all time points is larger than the threshold. In this experiment, we set the threshold for gene expression value to 8. We remove the nodes that are isolated at all layers from the multilayer network as they are guaranteed to not contribute to motif count. Table 1 describes the details of the six multilayer networks.

Methods Compared Against. To the best of our knowledge, this is the first extended study to count independent motifs in multilayer networks (see [43] for the preliminary version of this study). Most motif finding algorithm focus on the single layer network. To better evaluate the performance of our method, we use three state of the art algorithms, mfinder [25], ESU (FANMOD) [60] and G-Tries [45], to count motifs in multilayer networks. Same with most motif finding methods, these algorithms are limited to single layer network. To address this limitation, we propose a solution and briefly summarize it as follows. It works in three steps: (i) It first feeds these algorithms to each layer of the multilayer network one by one, which locate all embeddings in each layer. Thus, it produces an embedding set with each embedding associated with a set of layers. Different from our method first locating embeddings from aggregate network, this step requires the algorithm running k times for a multi-layer network with k layers. (ii) It then randomly picks one embedding which appear on the largest number of target layers while removing embeddings sharing the same edges. Same with our method, the set of target layers is initialized to all layers of network and then changes to the common layers of all selected embeddings. (iii) It repeats step (ii) until there are no embeddings or the remaining embeddings do not appear on more than specified number of target layers.

Implementation and System Details. We implement the algorithm in C++. We perform all the computational experiments on a Linux machine equipped with Intel core i7 processor 3.6 GHz CPU and 12GBs RAM.

3.1 Evaluation on Synthetic Networks

We compared our method with respect to three state of the art methods under a wide spectrum of parameter values using synthetic datasets. We vary the network size, average degree, edge ratio and parameter α. In each experiment, we vary one of these parameters and fix other ones. We repeat each experiment on 10 networks. We measure the accuracy and running time for each inference method and report the average result. We calculate accuracy as the ratio of the number of embeddings discovered to the number of embeddings planted.

Effect of Network Size. We run experiments using networks of sizes of 200, 400, 800 and 1600. We set the average degree to 4, the edge ratio to 0.2 and α to 0.6. Fig. 8a plots the results. We first focus on the accuracy. Our results demonstrate that our method achieves almost 100% accuracy rate for all network sizes across all motif types. Besides, we observe that the gap between our method and other methods gradually decreases with growing network sizes. The possible reason is that the network becomes sparser with increasing number of nodes while fixing the network degree. Sparse networks reduce the risk of missing potential embeddings when using other methods to find embeddings in each layer as the motif embeddings are less likely to overlap with each other. Moreover, we observe that the accuracy rate of other three methods in finding motifs of cascade and feed forward loop is much higher than that of bifan and biparallel patterns. This is mainly because the topologies of the former are much simpler than the latter (i.e., the cascade pattern consists of three nodes and three edges; while the bifan contains four nodes and four edges). As a result, other three methods have a smaller chance to miss simple motif patterns when locating all possible embeddings in each layer.

Next, we evaluate the running time. Our method either achieves the best or the second best running time across all network sizes for all motif types. Even for the largest network (i.e., 1600 nodes), our method runs very fast (in only a few seconds). Thus, our method has the potential to scale to large networks. We observe that mfinder has by far the

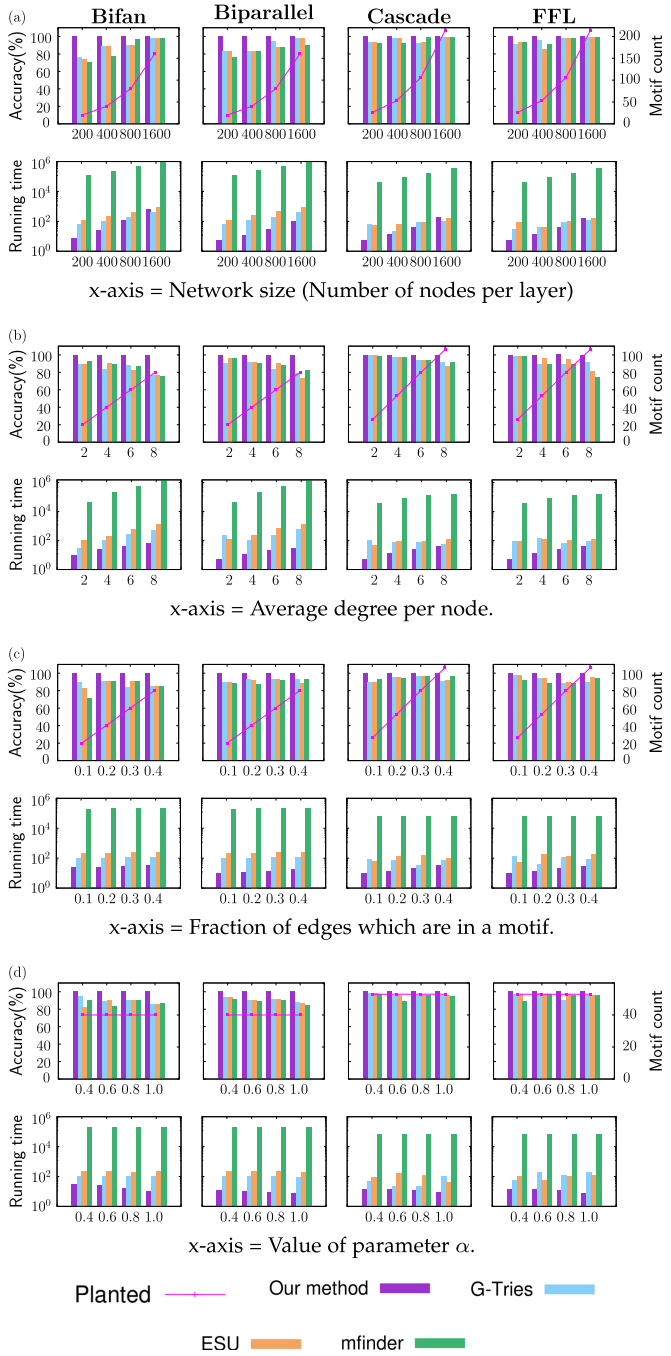

Fig. 8. Accuracy and running time of methods on multilayer networks with various parameters for four motif patterns. The x-axis shows (a) network sizes; (b) network degrees; (c) edge ratios (i.e., the fraction of edges that are part of a motif embedding); (d) α values. The y-axis on the left of the first figure in (a) - (d) represents the accuracy of methods; the y-axis on the right represents the number of planted embeddings, which is shown with the red line in the bar charts. The running times are reported in million seconds and represented in log-scale.

largest running time, resulting from running some other tasks such as calculating statistical significance. The running times of all methods increases with the increasing network sizes, yet they remain feasible particularly for the three methods: i.e., ours, FANMOD, and G-Tries. The running times of cascade and feed forward loop are slighter lower than those of bifan and biparallel. One possible reason is that the topologies of the former are relatively simpler, leading to smaller amount time of discovering all possible embeddings regardless of the number of embeddings planted.

Effect of Network Degree. We fix the network size to 400, the edge ratio to 0.2 and α equal to 0.6, and we vary the network degrees as 2, 4, 6 and 8. Fig. 8b reports the results. Consistent with the previous experiment, our method achieves the best performance with almost 100% accuracy rate and the lowest running time consistently for all degrees and motif topologies. More importantly, we observe that the gap between the accuracy of our method and those of other methods increases with the growing network degrees across all motifs. This is consistent with the results in Fig. 8a since increasing degree complicates the network topologies (i.e., the network becomes denser) and the chance that the embeddings overlap with each other.

Effect of Edge Ratio. We use the network with varying edge ratio (i.e., ratio of embedded motif edges to all edges) from 0.1 to 0.4 at increments of 0.1. We set the network size, network degree and α to 400, 4 and 0.6 respectively. Fig. 8c presents the result. Similar to the previous experiments, our method achieves the best performance in terms of both accuracy and running time. Increasing edge ratios leads to increasing number of embeddings planted. However, increasing the number of embeddings planted has limited effect on the running time. This implies that locating all possible embeddings contributes more to the running times than iteratively picking up independent embeddings (see Section 2.3 for the time complexity analysis). One possible reason is that such randomly generated multilayer networks do not have many overlapping embeddings. We investigate this further on real multilayer networks in the next section.

Effect of α Value. Here, we vary α from 0.4 to 1.0 at increments of 0.2. We fix the network size, network degree and edge ratio to 400, 4 and 0.2 respectively. Fig. 8d shows the result. Our results are consistent with those in our previous experiment that our method is both more accurate and faster than competing methods. We also observe that increasing the α value does not substantially affect the accuracy of other methods. Especially for the value 1.0, the impact of heuristic strategy of picking embeddings is reduced to zero since all layers have been planted same set of embeddings and the target layer set will not shrink when picking embeddings. Obviously, the accuracies for other methods have not increased. We observe that searching embeddings by analyzing layer by layer fails to report all planted embeddings, proving the advantage of our method which finds embeddings on a single layer network describing all the layers collectively (i.e., aggregate network).

3.2 Evaluation on Real Networks

Topological characteristics of the real networks can exhibit substantial differences with respect to those of synthetic networks. We performed tests of the proposed counting motifs algorithm on six E.coli multilayer networks.

Motif Count and Running Time. We run the here proposed algorithm by looking for unknown motifs counts and compare performance with available methods in terms results and running time. Experimental results are reported in

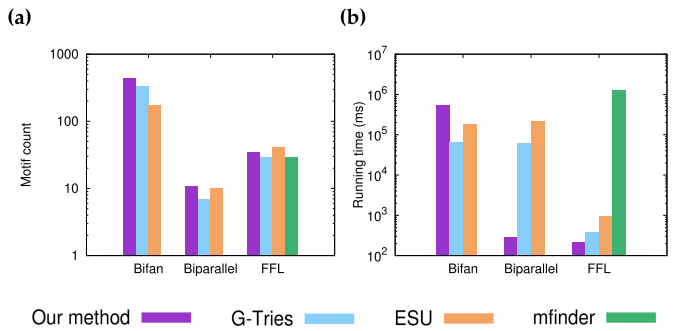

Fig. 9. Motif count (on the left) and running time (on the right) of different methods running on the multilayer network under cold stress. Running times are in milliseconds and reported in log-scale. Experiments run with $\alpha = 0.6$.

Fig. 9. Results show that the proposed method performs better than other methods (left part of Fig. 9). ESU method identifies slightly larger number of embeddings for the feed forward loop (FFL, in orange the ESU Motif count results in left part of Fig. 9), even if the difference is very small. This result in terms of motif count is consistent with what obtained with synthetic networks. In terms of efficiency (i.e., running time in the right part of Fig. 9), the here proposed methos is faster than all the other methods but Bifan case, probably due to the cost of constructing and maintaining the overlap graph for E.Coli. Nevertheless, the running is comparable even for bifan. Given such results on real networks, we conclude that by using the proposed method, finding embeddings in multilayer networks is fast and highly accurate.

We also run methods varying α value from 0.6 to 1.0. Table 2 reports the motif counts of three patterns on the E. coli multilayer networks at different experimental conditions. E.coli network contains a large number of bifan patterns compared to other two ones across all conditions. Standard deviation of motif counts is very small on networks in cold, heat and control conditions, meaning that the largest set of embeddings are conserved at all layers varying conditions through time. Changes in motif counts arise under oxidative and lactose stress during the same interval, suggesting that such conditions are more disruptive to the organization and abundance of motifs in E.coli network than other experimental conditions. Finally, results reported in Table 2 show that motif counts change when checking all conditions together (see bifan patterns for instance), which indicates that conditions influences network topologies.

Statistical Observations. Results Statistical significance have been calculated using *p-value* and *z-score*. To design a suitable null-model, we consider the dependencies between layers. We construct a random bijective function $\psi() : |V| \times |V| \rightarrow |V| \times |V|$. A random network layer is generated for each existing network layer i, by replacing all $(u, v) \in E_i$ with $\psi(u, v)$. This allows preserving dependency between layers. Given two layers sharing an edge (u, v) in the original network, they also share edge $\psi(u, v)$ in the randomized network. By repeating several time the randomizing process, we evaluate mean and the standard deviation of the motif counts in the randomly generated multilayer networks.

TABLE 2
Motif Count of Three Motif Patterns on the E.Coli Multilayer Networks Under Various Experimental Conditions

Conditions	Bifan	Biparallel	FFL
Control	430.33 ± 1.25	10.00 ± 0.00	36.33 ± 0.47
Cold	434.00 ± 0.00	35.00 ± 0.00	11.00 ± 0.00
Heat	365.00 ± 4.24	9.00 ± 0.00	41.33 ± 0.47
Oxidative	452.00 ± 12.33	12.33 ± 2.05	44.33 ± 2.87
Lactose	497.67 ± 20.24	21.00 ± 3.56	72.33 ± 9.74
All	394.67 ± 40.88	8.67 ± 4.11	30.67 ± 4.19

Each result is reported as mean $\pm$ standard deviation.

Let N^* be the number of time a given motif in the original network appears; let μ and σ be respectively mean and standard deviation of the motif in the random networks. We calculate the z-score as

$$Z = \frac{N^* - \mu}{\sigma}.$$

We claim that a motif is over- or under-represented if its z-score is ≥ 2 or ≤ -2 respectively. Fig. 11 presents the results for three motif patterns on the E.coli multilayer networks under various stress conditions using different minimum numbers of target layers. Large z-score values are obtained in all conditions, for both bifan and feed forward loop tests that are thus the building blocks of E.coli network, in line with [36]. Biparallel is significant only on oxidative, lactose and all conditions, which show substantial change of motif counts while changing the minimum numbers of target layers in the previous experiment. Finally, a *p-value* < 0.01 is obtained for all significant patterns. We also investigate the running time of statistical significance analysis experiment and report the result on the multilayer network under cold stress with α = 0.6. By generating 1000 randomized multilayer networks, we obtain the significance values of three motif patterns (bifan, biparallel and FFL) in approximate 446, 10, and 7 seconds respectively. Compared to the running time of counting motif of those three motif patterns (550, 0.3, and 0.2 seconds, Fig. 9a), the performance of our algorithm of calculating statistical significance largely depends on the network size and topology but remains practical.

Experiments on Genes Motif. We performed tests on genes forming motif patterns. We start focusing on five condition multilayer networks, where each layer represents one time point after the application of the stress condition. For each condition specific multilayer network, we consider all the genes in the embeddings obtained by using our method. For each motif related gene, we count the number of condition specific multilayer networks where it appears on the embeddings. Fig. 10 reports the distribution of motif related genes. We observe that, as we increase the number of conditions (i.e., as we move from left to right in Fig. 10), the number of genes observed first decreases, then increases again. As a result, either one condition (left most bar) or all conditions (rightmost bar) yields the largest number of genes. This suggests that most genes tend to characterize either behavior unique to one condition specific network (discriminative function), or consistent behavior across all conditions (robust function). To

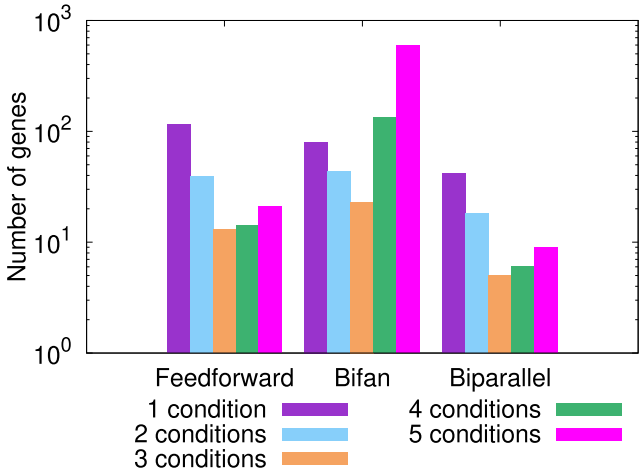

Fig. 10. The distribution of motif related genes that are belong to different number of condition specific multilayer networks.

understand the impact of this behavior, we have also performed gene ontology analysis on genes that are related to all conditions. In Table 3, we report the top five Gene Ontology (GO) terms, that show stress condition related biological processes for the genes linked to dicriminative or robust functions found in Fig. 10 and their False Discovery Rates (FDRs). We also report the FDRs of the same functions for the genes found using single layer networks. We observe that bifan genes appearing in one or all networks play a significant role on stress related biological process, such as the oxidation-reduction process. We also observe that multilayer network analysis yields much better FDR values for all the functions. Single layer network analysis not only yields genes with higher FDR values but it also completely misses two statistically highly significant functions.

TABLE 3
Gene Ontology Analysis of the Genes Appearing in Condition Specific Networks Under One or All Conditions

GO Term	Biological process	FDR (m)	FDR (s)
0008150	biological process	3.93E-31	NA
0044699	single-organism process	1.27E-27	NA
0055114	oxidation-reduction process	3.06E-25	1.53E-15
0045333	cellular respiration	2.16E-20	1.15E-12
0015980	energy derivation by oxidation of organic compounds	8.33E-20	7.78E-13

FDR (m) and FDR (s) show the false discovery rates for the genes found in multi and single layer networks respectively. NA indicates that the corresponding biological process is not found.

In summary, results on both synthetic and real datasets show that our method is robust to the growing network sizes, network degrees, edge ratios and α value. Also, the method is more efficient in terms of execution time than existing methods. The parameters which influence network, such as the network size, average degree, and the number of candidate embeddings, and motif topology have great impact on the running time.

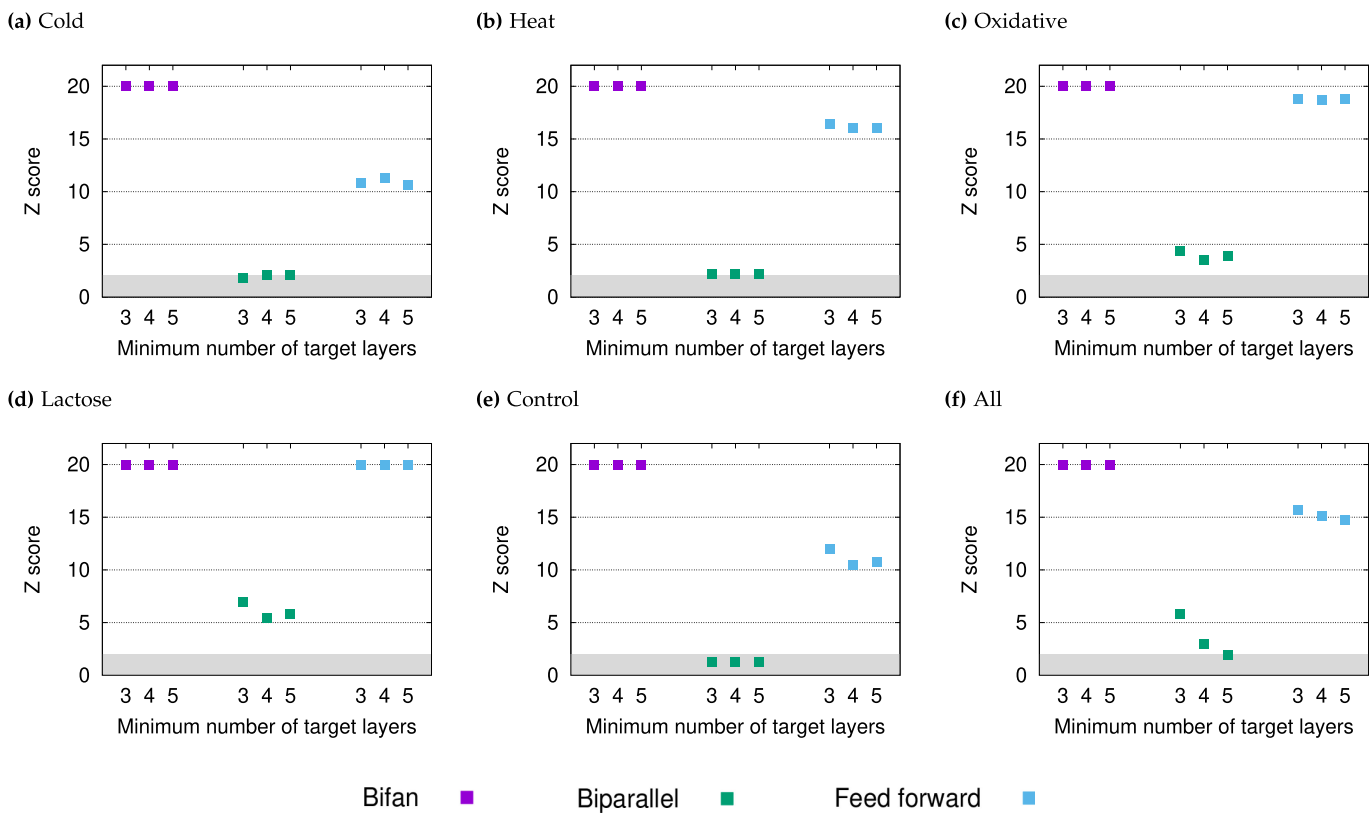

Fig. 11. The z-score values of three patterns on E.coli multilayer networks under various experimental conditions using different minimum numbers of target layers. If the z-score is greater than 20, we report it as 20 to reduce the y-axis scale of the plots. The shaded area represents z-score values that are not significant.

4 Conclusion

Traditional methods in the existing literature often represent complex systems as a single, static, and binary network. These models are inadequate in capturing complex cellular interactions which vary under different conditions as well as over time. Furthermore, the same set of molecules can interact in varying patterns across different interactomes. In this paper, we considered one of the most fundamental problems in network analysis. We extended the classical network motif identification problem to multilayer networks. We developed an efficient and accurate method to solve this problem. Our experimental results on both synthetic and real datasets demonstrated that our method identifies motifs in multilayer networks with high accuracy and scales to large networks in practical time. Our results on E. coli transcription regulatory networks demonstrate that our method helps in uncovering key functional characteristics of biological networks.

References

[1] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[2] N. Alon, P. Dao, I. Hajirasouliha, F. Hormozdiari, and S. Cenk Sahinalp, "Biomolecular network motif counting and discovery by color coding," *Bioinformatics*, vol. 24, no. 13, pp. i241–i249, Jul. 2008.

[3] F. Battiston, V. Nicosia, and V. Latora, "Structural measures for multiplex networks," *Phys. Rev. E*, vol. 89, no. 3, 2014, Art. no. 032804.

[4] B. Bentley *et al.*, "The multilayer connectome of caenorhabditis elegans," *PLoS Comput. Biol.*, vol. 12, no. 12, 2016, Art. no. e1005283.

[5] G. Bianconi, "Statistical mechanics of multiplex networks: Entropy and overlap," *Phys. Rev. E*, vol. 87, no. 6, 2013, Art. no. 062806.

[6] G. Bianconi and S. N. Dorogovtsev, "Multiple percolation transitions in a configuration model of a network of networks," *Phys. Rev. E*, vol. 89, no. 6, 2014, Art. no. 062814.

[7] S. Biswas, S. Ray, and S. Bandyopadhyay, "Colored network motif analysis by dynamic programming approach: An application in host-pathogen interaction network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 2, pp. 550–561, Mar./Apr. 2021.

[8] S. Boccaletti *et al.*, "The structure and dynamics of multilayer networks," *Phys. Rep.*, vol. 544, no. 1, pp. 1–122, 2014.

[9] S. A. Cook, "The complexity of theorem-proving procedures," in *Proc. 3rd Annu. ACM Symp. Theory Comput.*, 1971, pp. 151–158.

[10] R. Criado, J. Flores, A. García del Amo, J. Gómez-Gardenes, and M. Romance, "A mathematical model for networks with structures in the mesoscale," *Int. J. Comput. Math.*, vol. 89, no. 3, pp. 291–309, 2012.

[11] M. De Domenico *et al.*, "Mathematical formulation of multilayer networks," *Phys. Rev. X*, vol. 3, no. 4, 2013, Art. no. 041022.

[12] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable representation learning for heterogeneous networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2017, pp. 135–144.

[13] R. Elhesha and T. Kahveci, "Identification of large disjoint motifs in biological networks," *BMC Bioinf.*, vol. 17, 2016, Art. no. 408.

[14] M. Flajolet *et al.*, "A genomic approach of the hepatitis C virus generates a protein interaction map," *Gene*, vol. 242, no. 1, pp. 369–379, 2000.

[15] S. Gama-Castro *et al.*, "Regulondb version 9.0: High-level integration of gene regulation, coexpression, motif clustering and beyond," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D133–D143, 2015.

[16] M. R. Garey and D. S. Johnson, "Computers and intractability: A guide to the theory of npcompleteness (series of books in the mathematical sciences)," in *Computers and Intractability*, San Francisco, CA, USA: Freeman, 1979.

[17] E. Golemis and P. David Adams, *Protein-Protein Interactions: A Molecular Cloning Manual*. Cold Spring Harbor, NY, USA: Cold Spring Harbor Lab. Press, 2002.

[18] S. Gomez, A. Diaz-Guilera, J. Gomez-Gardenes, C. J. Perez-Vicente, Y. Moreno, and A. Arenas, "Diffusion dynamics on multiplex networks," *Phys. Rev. Lett.*, vol. 110, no. 2, 2013, Art. no. 028701.

[19] J. Gómez-Gardenes, I. Reinares, A. Arenas, and L. Mario Floría, "Evolution of cooperation in multiplex networks," *Sci. Rep.*, vol. 2, 2012, Art. no. 620.

[20] M. L. Green and P. D. Karp, "A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases," *BMC Bioinf.*, vol. 5, no. 1, 2004, Art. no. 1.

[21] J. A. Grochow and M. Kellis, "Network motif discovery using subgraph enumeration and symmetry-breaking," in *Proc. 11th Annu. Int. Conf. Res. Comput. Mol. Biol.*, 2007, pp. 92–106.

[22] Y. Hu *et al.*, "An integrative approach to ortholog prediction for disease-focused and other functional studies," *BMC Bioinf.*, vol. 12, no. 1, 2011, Art. no. 1.

[23] S. Jozefczuk *et al.*, "Metabolomic and transcriptomic stress response of *Escherichia coli*," *Mol. Syst. Biol.*, vol. 6, no. 1, 2010, Art. no. 364.

[24] S. Kai, G. Lin, and W. B. Bo, "An integrated network motif based approach to identify colorectal cancer related genes," in *Proc. 34th Chin. Control Conf.*, 2015, pp. 8573–8578.

[25] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs," *Bioinformatics*, vol. 20, no. 11, pp. 1746–1758, 2004.

[26] J. Yeol Kim and K.-I. Goh, "Coevolution and correlated multiplexity in multiplex networks," *Phys. Rev. Lett.*, vol. 111, no. 5, 2013, Art. no. 058702.

[27] W. Kim, M. Li, J. Wang, and Y. Pan, "Biological network motif detection and evaluation," *BMC Syst. Biol.*, vol. 5, no. Suppl 3, Dec. 2011, Art. no. S5.

[28] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *J. Complex Netw.*, vol. 2, no. 3, pp. 203–271, 2014.

[29] X. Li, Rebecca J. Stones, H. Wang, H. Deng, X. Liu, and G. Wang, "Netmode: Network motif detection without nauty," *PLoS One*, vol. 7, no. 12, pp. 1–9, Dec. 2012.

[30] J. Luo, L. Ding, C. Liang, and N. H. Tu, "An efficient network motif discovery approach for co-regulatory networks," *IEEE Access*, vol. 6, pp. 14151–14158, 2018.

[31] S. Mbadiwe and W. Kim, "ParaMODA: Improving motif-centric subgraph pattern search in PPI networks," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2017, pp. 1723–1730.

[32] L. A. A. Meira, V. R. Máximo, Á. L. Fazenda, and A. F. da Conceição, "acc-motif: Accelerated network motif detection," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 5, pp. 853–862, Sep./Oct. 2014.

[33] G. Menichetti, D. Remondini, P. Panzarasa, R. J. Mondragón, and G. Bianconi, "Weighted multiplex networks," *PLoS One*, vol. 9, no. 6, 2014, Art. no. e97857.

[34] T. Michoel and B. Nachtergaele, "Alignment and integration of complex networks by hypergraph-based spectral clustering," *Phys. Rev. E*, vol. 86, no. 5, 2012, Art. no. 056111.

[35] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon, "On the uniform generation of random graphs with prescribed degree sequences," 2004, *arXiv:cond-mat/0312028*.

[36] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Sci.*, vol. 298, no. 5594, pp. 824–827, 2002.

[37] K. Mukharjee, M. Mahmudul Hasan, C. Boucher, and T. Kahveci, "Counting motifs in dynamic networks," *BMC Syst. Biol.*, vol. 12, no. 1, 2018, Art. no. 6.

[38] V. Nicosia, G. Bianconi, V. Latora, and M. Barthelemy, "Growing multiplex networks," *Phys. Rev. Lett.*, vol. 111, no. 5, 2013, Art. no. 058701.

[39] V. Nicosia and V. Latora, "Measuring and modeling correlations in multiplex networks," *Phys. Rev. E*, vol. 92, no. 3, 2015, Art. no. 032805.

[40] V. Nicosia, M. Valencia, M. Chavez, A. Díaz-Guilera, and V. Latora, "Remote synchronization reveals network symmetries and functional modules," *Phys. Rev. Lett.*, vol. 110, no. 17, 2013, Art. no. 174102.

[41] S. Omidi, F. Schreiber, and A. Masoudi-Nejad, "Moda: An efficient algorithm for network motif discovery in biological networks," *Genes Genet. Syst.*, vol. 84, pp. 385–95, Oct. 2009.

[42] S. Patra and A. Mohapatra, "Protein complex prediction in interaction network based on network motif," *Comput. Biol. Chem.*, vol. 89, 2020, Art. no. 107399.
[43] Y. Ren, A. Sarkar, A. Ay, A. Dobra, and T. Kahveci, "Finding conserved patterns in multilayer networks," in *Proc. ACM BCB Conf. Proc.*, 2019, pp. 97–102.
[44] P. Ribeiro, P. Paredes, M. E. P. Silva, D. Aparício, and F. Silva, "A survey on subgraph counting: Concepts, algorithms and applications to network motifs and graphlets," *ACM Comput. Surveys (CSUR)*, ACM New York, NY, USA, vol. 54, no. 2, pp. 1–36, 2021.
[45] P. Ribeiro and F. Silva, "G-tries: An efficient data structure for discovering network motifs," in *Proc. ACM Symp. Appl. Comput.*, 2010, pp. 1559–1566.
[46] R. A. Rossi *et al.*, "Heterogeneous network motifs," 2019, *arXiv: 1901.10026*.
[47] M. Salehi, R. Sharma, M. Marzolla, M. Magnani, P. Siyari, and D. Montesi, "Spreading processes in multilayer networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 2, no. 2, pp. 65–83, Apr.–Jun. 2015.
[48] A. Sarkar, Y. Ren, R. Elhesha, and T. Kahveci, "A new algorithm for counting independent motifs in probabilistic networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 4, pp. 1049–1062, Jul./Aug. 2019.
[49] A. Sarkar, Y. Ren, R. Elhesha, and T. Kahveci, "Counting independent motifs in probabilistic networks," in *Proc. 7th ACM Int. Conf. Bioinf., Comput. Biol., Health Inform.*, 2016, pp. 231–240.
[50] F. Schreiber and H. Schwöbbermeyer, "Frequency concepts and pattern detection for the analysis of motifs in networks," *Trans. Comput. Syst. Biol. III*, vol. 3, pp. 89–104, 2005.
[51] J. Scott, T. Ideker, R. M. Karp, and R. Sharan, "Efficient algorithms for detecting signaling pathways in protein interaction networks," *J. Comput. Biol.*, vol. 13, no. 2, pp. 133–144, 2006.
[52] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of *Escherichia coli*," *Nature Genet.*, vol. 31, no. 1, pp. 64–68, 2002.
[53] L. Solá, M. Romance, R. Criado, J. Flores, A. García del Amo, and S. Boccaletti, "Eigenvector centrality of nodes in multiplex networks," *Chaos: An Interdiscipl. J. Nonlinear Sci.*, vol. 23, no. 3, 2013, Art. no. 033131.
[54] A. Todor, A. Dobra, and T. Kahveci, "Counting motifs in probabilistic biological networks," in *Proc. 6th ACM Conf. Bioinf. Comput. Biol. Health Inform.*, 2015, pp. 116–125.
[55] P. Uetz *et al.*, "A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae," *Nature*, vol. 403, no. 6770, pp. 623–627, 2000.
[56] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLoS Comput. Biol.*, vol. 6, no. 1, 2010, Art. no. e1000641.
[57] H. Wang, H. Huang, and C. Ding, "Function–function correlated multi-label protein function prediction over interaction networks," *J. Comput. Biol.*, vol. 20, no. 4, pp. 322–343, 2013.
[58] P. Wang, J. Lü, and X. Yu, "Identification of important nodes in directed biological networks: A network motif approach," *PLoS One*, vol. 9, no. 8, 2014, Art. no. e106132.
[59] T. Wang, J. Peng, Q. Peng, Y. Wang, and J. Chen, "FSM: Fast and scalable network motif discovery for exploring higher-order network organizations," *Methods*, vol. 173, pp. 83–93, 2020.
[60] S. Wernicke, "Efficient detection of network motifs," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 3, no. 4, pp. 347–359, Fourth quarter 2006.
[61] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Mol. Syst. Biol.*, vol. 4, no. 1, 2008, Art. no. 189.
[62] M. Zitnik and J. Leskovec, "Predicting multicellular function through multi-layer tissue networks," *Bioinformatics*, vol. 33, no. 14, pp. i190–i198, 2017.

Yuanfang Ren received the BE degree in software engineering from Nanjing University, Nanjing, China, in 2010, the ME degree in computer science from Tongji University, Shanghai, China, and the PhD degree from the Computer and Information Science and Engineering Department, University of Florida, Gainesville, Florida, in December 2018. Her research focuses on bioinformatics.

Aisharjya Sarkar received the BTech degree from Information Technology Department, West Bengal University of Technology, Kalyani, India, the ME degrees Indian Institute of Engineering, Science and Technology, Shibpur, India, and the PhD degree from Computer and Information Science and Engineering Department, University of Florida, Gainesville, Florida, in 2020. Her research interests include bioinformatics and machine learning.

Pierangelo Veltri received the PhD degree in computer science from Paris XI, Orsay, France, in 2002. He was a researcher with INRIA-Rocquencourt from 1998 to 2002 on database models and query languages for semistructured data. From 2000 to 2002, he was an adjunct professor with Paris Villetaneuse, where teach database and programming language. He joined the Faculty of Medicine, Italy, in 2002 and was an assistant professor until 2010 and as an associate professor until 2020. He is currently a full professor. His main interests include data modeling, protein and molecular modeling, spatial and geographic database systems, and health informatics. He teaches database and clinical informatics systems. He is the editor of ACM SIGBIO newsletter and an associate editor for the *BMC Medical Informatics and Decision Making* and *Journal of Healthcare and Informatics Research*.

Ahmet Ay received the BS degree in mathematics from Bilkent University, Ankara, Turkey, in 2002, and the PhD degree in mathematics and quantitative biology from Michigan State University, East Lansing, MI, USA, in 2009. He is currently an associate professor with the Departments of Biology and Mathematics, Colgate University. His main research interests include systems biology, bioinformatics, and mathematical biology. He has worked on construction, analysis, and modeling of gene regulatory networks.

Alin Dobra received the BS degree in computer engineering from the Technical University of Cluj-Napoca, Cluj-Napoca, Romania, in 1998, and the PhD degree in computer science from Cornell University, Ithaca, New York, in 2003. He is currently an associate professor with Computer and Information Science and Engineering Department, University of Florida. His main research interests include large-scale database systems, approximate query processing, and probabilistic models. He was the recipient of the National Science Foundation (NSF) CAREER Award in 2005 and SIGMOD 2007 Best Paper Award.

Tamer Kahveci received the PhD degree in computer science from the University of California, Santa Barbara, CA, USA, in 2004. He has been working as a faculty member with Computer and Information Science and Engineering, University of Florida, since Fall 2004. His main research interests include bioinformatics, computational biology, and databases. He has worked on indexing large databases, sequence alignment, and computational analysis of biological pathways. He was the recipient of the Ralph E. Powe Junior Faculty Enhancement Award in 2006, CSB Best Paper Award in 2008, NSF Career award in 2009, ACM BCB Best Student Paper Award in 2010, and BICOB Best Paper Award in 2018.