Research Statement
Marinka Zitnik

My research investigates machine learning for biomedical sciences, focusing on large networks of interactions between biomedical entities—e.g., proteins, drugs, diseases, and patients. I leverage these networks at the scale of billions of interactions among millions of entities and develop new methods blending network science with statistical methods and machine learning. I use my methods to answer important scientific questions, such as how Darwinian evolution changes molecular networks, and how data-driven algorithms accelerate scientific discovery; and I use the methods to solve high-impact problems, such as what drugs and combinations of drugs are safe for patients, what molecules will treat what diseases, and how newborns are transferred between hospitals and how these transfers influence outcomes.

Large-scale biomedical data present multiple fundamental challenges: (1) Data involve rich multimodal and heterogeneous interactions that span from the molecular scale to whole ecosystems—the challenge is how to computationally operationalize these data to make them amenable to analytics; (2) Data are of many different types, including experimental readouts, curated annotations and metadata—no single data type can capture all the factors necessary to understand a phenomenon such as a disease; (3) Biomedical data are noisy due to the inherent natural variations and the limitations of measurement technologies, and include missing data, repeated measurements, and contradictory observations, which plague the analysis.

My methods address these challenges, uniquely advance data science and deep learning, and infuse biomedical knowledge into deep models. Networks, or graphs, pervade biomedical data—from the molecular level to the level of connections between diseases in a person, and all the way to the societal level encompassing all human interactions (Figure 1). These interactions at different levels give rise to a bewildering degree of complexity which is only likely to be fully understood through a holistic and integrated systems view and the study of combined, multilevel networks. This premise is a key principle of my research: We need to integrate the diverse data and knowledge into network representations and invent accurate machine learning and data science methods that use these networks to advance science and solve open problems. My research proves that this premise not only opens up new avenues for understanding nature, analyzing health, and developing new medicines to help people but can impact on the way predictive modeling is performed today at the fundamental level.

My research realizes an end-to-end scientific approach that consists of: (i) Finding ways to combine rich, heterogeneous network data in their broadest sense to reduce redundancy and uncertainty and to make them amenable to comprehensive analyses; (ii) Developing deep learning and statistical methods to reason over these complex networks and learn useful network representations; and (iii) Translating machine learning and data science research into innovative biomedical applications of high impact by fully harnessing large-scale and rich interaction data. The table below summarizes the structure of my research with the annotation symbols used in this statement.

My work is highly interdisciplinary and has made contributions to both computer science and to biomedicine. This interdisciplinary nature of my research is exemplified by publications in top-tier machine learning and data mining venues (JMLR, NIPS, IEEE TPAMI, KDD, Information Fusion), top-tier bioinformatics and computational biology venues (Bioinformatics, PLoS Computational Biology, ISMB, RECOMB, PSB), interdisciplinary journals (three publications in Nature Communications, paper in revision in Proceedings of the National Academy of Sciences (PNAS)), several Best Paper, Poster, and Research Awards from the International Society for Computational Biology, and two Rising Stars awards from the MIT EECS and The Broad Institute of MIT and Harvard, being the only young scientist who received such recognition in both EECS and Biomedicine.

My methods have had a tangible real-world impact, for which has garnered interests of government, academic, and industry researchers and has put new tools in the hands of practitioners. For example, my learning framework allows scientists at Baylor College of Medicine to identify new bacterial response genes forty times faster and better than standard genetic screens [44]; this data fusion methodology is now a core data science platform of several startups (e.g., Genialis). Through interdisciplinary collaborations, my machine learning work has given scientists at The Helleday Foundation, SciLifeLab, and Karolinska Institute insights on new therapeutic targets [50] and has fueled a clinical trial on a new cancer target. My data science work powers new drug discovery tools at Stanford Medical School [3], and gives scientists at Newton-Wellesley and Massachusetts General hospitals guidance on prescribing safe combinations of drugs [54].

<table>
<thead>
<tr>
<th>Methods &amp; algorithms</th>
<th>Analysis &amp; models</th>
<th>Discovery &amp; insights</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Coalescing knowledge by fusing heterogeneous networks</td>
<td>◊</td>
<td>○</td>
</tr>
<tr>
<td>(ii) Statistical methodology and deep learning on graphs</td>
<td>◊</td>
<td>◊</td>
</tr>
<tr>
<td>(iii) Enabling data-driven solutions to new biomedical problems</td>
<td>◊</td>
<td>○</td>
</tr>
</tbody>
</table>

Fusing heterogeneous networks

Multimodal, or heterogeneous, graphs, in which nodes represent different types of entities (modes) and edges capture various types of interactions between the entities, are a foundational mechanism for expressing complex biomedical systems. I believe that mining of massive multimodal networks is essential to close a gap between early single-data-type studies and the new biomedical era that strives for comprehensive views of complex systems, such as the human body. My research develops methods for creating and analyzing large-scale multimodal networks by fusing data of different types and from different sources, compensating for missing or unreliable information, and reducing false positives and negatives. These methods allow us to create holistic systems views, enabling discoveries at the frontier.

1 Citations refer to the numbered publication list in my CV.
New collective latent factor models enable large-scale biomedical data fusion

My research developed a novel class of collective latent factor models founded on nonlinear dimensionality reduction to conduct powerful data fusion [36,39,43,46,47]. Previously, data integration research was fundamentally limited in its ability to scale due to customized solutions designed for very specific studies. My approach represented the first general framework that could fuse any set of biomedical networks (Figure 2) and infer an accurate latent representation of the networks. The learned representation enabled us to make predictions outside of an existing knowledge domain, e.g., extrapolate a drug response from an animal model to human patients. Here, predictions led to several discoveries, including the discovery of eight new bacterial response genes in eukaryotes [44]. Validation of predictions for this specific question required only a few days of work in the wet laboratory; however, it would take a year to discover the same number of genes with standard genetic screens. In another study, published in Nature Communications [50], I used my approach to fuse genomic, structural, and biochemical networks, which led to the first-ever characterization of the Nudix protein family that fueled a clinical trial on Nudix proteins as cancer targets.

Network enhancement as a general method to denoise networks

To address the challenge of noisy and incomplete network data, I developed methodology to enhance networks using spectral network theory and higher-order structures. The method takes as input a noisy network and outputs a network on the same set of nodes but with a new set of enhanced, more accurate edges. This work, published in Nature Communications [55], gives theoretical advancements, including a closed-form solution with mathematical guarantees on spectral structure of the enhanced network. These advancements provide new insights into how network denoising can remove weak edges, enhance real interactions, and improve performance by up to 41% on various prediction tasks and diverse types of networks, from gene interaction networks to noisy Hi-C contact maps and similarity networks.

New network science paradigm reveals evolution of interactomes across the tree of life

Using protein–protein interaction data that have only recently become available, I composed and analyzed interactome networks from 1,840 species across the tree of life, expanding the number of species from about 5 in previous studies to 1,840 (Figure 3). This unique dataset allowed me to conduct the largest ever study of protein interactomes and quantify the resilience of interactomes—a critical property as the breakdown of proteins may lead to cell death or disease [57]. I quantified a new interactome resilience measure based on Shannon’s information theory, which had never been described before, as even the most extensive studies were unable to compare interactomes globally. This study revealed that evolution leads to more resilient interactomes, providing evidence for a longstanding hypothesis that interactomes evolve favoring robustness against protein failures. Designing stratification experiments and statistical procedures to test for several dozen confounders, I showed that biases in network data could not explain our key findings. We showed that a highly resilient interactome has an astonishingly beneficial impact on the organism to survive in complex, variable, and competitive habitats, a finding that draws attention to a previously unknown critical role of evolution in mediating the effects of the interactome on the ability of a species to thrive in specific habitats. This work, now in revision in Proceedings of the National Academy of Sciences (PNAS) [57], is the first to reveal how interactomes change during evolution, and how these changes impact their response to environmental unpredictability.

Statistical methodology and deep learning on graphs

Deep learning is emerging as a leading candidate to learn and reason over complex data, such as multimodal networks. The challenge, however, is that the prevailing deep learning algorithms are designed for data with a regular, grid-like structure (e.g., images have a 2D grid structure and sequences have a linear 1D structure). These algorithms are unable to truly exploit complex, irregular interactions between entities, i.e., edges, the essence of graphs. Furthermore, they ignore that graphs are of arbitrary size with complex topology (i.e., no spatial locality like grids), that a node’s neighbors have no natural ordering, and that graphs are multimodal with rich information on the nodes and the edges. Developing deep learning algorithms for graphs and building a bridge between network science and deep learning thus represents an important scientific challenge. To address this challenge, I have been developing deep models for biomedical graphs.

Graph neural networks for learning deep embeddings of biomedical entities

At the technical core of my work is the notion of vector space embeddings. The idea is to learn how to represent nodes in a graph as points in a low-dimensional space, where the geometry of this space is optimized to reflect the structure of interactions between the nodes.
Building off my work in classic machine learning problems and dimensionality reduction [39], my research formalizes this idea by specifying deep transformation functions, or graph neural networks, that map nodes, or larger graph structures, to points in a low-dimensional space, termed embeddings (Figure 4). Importantly, these functions are optimized to embed the input graph so that performing algebraic operations in this learned space reflects the topology of the graph. For example, I developed a method for learning node embeddings based on structural similarity in graphs and showed mathematically that the method accurately recovers structurally similar and equivalent nodes [32]. I also designed a graph neural network that embeds not only nodes but also higher-order structures like subgraphs, and I demonstrated that such capability is crucial for problems, such as drug discovery [3]. To go beyond simple edge prediction, we introduced embedding frameworks that can handle more complex logical queries on graphs [33], allowing us to make predictions about logical queries potentially involving multiple unobserved edges, entities, and variables.

My research has pioneered graph neural networks in bioinformatics and computational biology (Figure 4). This allowed me to apply neural networks much more broadly and set sights on new frontiers beyond classic applications of neural networks that learn on images and sequences. I demonstrated that graph neural networks advance the state-of-the-art on a wide range of tasks and remove the need for painstaking manual feature engineering. In my tutorial on Deep Learning for Network Biomedicine, which was the most visited tutorial at the ISMB, the leading bioinformatics conference, I covered a wide range of problems for which graph neural networks are excellent models, including drug discovery, disease gene discovery, and single-cell genomics.

Applications to new biomedical problems

Next, I give four concrete examples of important biomedical applications that were enabled for the first time by my methods.

Enabling uniquely precise insights into the roles of human proteins

With few exceptions, all cells in our body have the same DNA and genes. As cells divide and grow, different genes are expressed, resulting in different cell types. Those cells then produce a variety of proteins specific to the cell types and tissues they form. If cells have the same DNA, why are eyes and lungs so different? At the core of this notoriously challenging question is the understanding of different roles that the same protein can have in different tissues. Leveraging tissue-specific gene expression data and protein–protein interaction networks, I was able to tackle this question by conducting the first-ever study of tissue-specific protein function prediction analyzing 107 protein–protein interaction networks and a hierarchy describing relationships between 219 human tissues [49]. Through solving an optimization problem across 107 tissue networks involving over 36 million potential protein–protein interactions, the approach embedded proteins, represented as nodes, into an embedding space (Figure 5), and used the embeddings for prediction. The approach outperformed baselines by 27% and, remarkably, showed a unique transfer-learning ability to predict protein functions in never-before-seen tissues. The research was recognized as a spotlight at ISMB, and I gave an invited talk about this line of research at NetSci.

Figure 5: Two-dimensional projection of protein embeddings. Protein–protein interaction networks from 9 human brain tissues are embedded in a vector space such that proteins with similar graph neighborhoods in each tissue are embedded close together [49]. Proteins are colored by the active tissue.

Learning to predict safety and side effects of drug combinations

Many patients take multiple drugs to treat complex, co-existing diseases; however combinations of drugs lead to a much higher risk of adverse side effects that occur because of interacting effects that cannot be attributed to either drug alone; this serious healthcare problem costs more than $177 billion a year in the US alone. I developed a deep learning framework to address this pressing health care problem. My approach enabled the first-ever predictive study of side effects caused by drug combinations across the entire US. I designed a graph convolutional neural network for learning on highly multi-relational graphs. Applying my method to a multi-relational graph of over 1 thousand different relation (edge) types representing 5 million drug–drug, drug–protein, and protein–protein interactions, I showed that the graph neural network could predict the safety and side effects of drug combinations, outperforming non-neural baselines by up to 69%, with larger gains achieved on side effects with strong molecular origins. Previously, pharmacological research was fundamentally limited in its ability to scale due to the exponential number of higher-order combinations of drugs. My approach has solved this challenge, predicting side effects even for combinations not yet used in patients. I gave invited talks about this research at the US National Academy of Sciences, the European Bioinformatics Institute (EMBL–EBI), the Stanford AI in Medicine Symposium, the Chan Zuckerberg Biohub, and Stanford Clinical Excellence Research Center as a National Guest Scholar. While the paper is the 6th most read Bioinformatics paper [54], what drives and excites me most is the opportunity to develop methods that give doctors guidance on prescribing safe treatments and help patients recognize side effects. Predictions of my approach led to follow-up research on prostate cancer patients at Stanford Medical School, and I now work with the Massachusetts General Hospital to validate predictions in the clinic.

Novel statistical network methods can accelerate scientific discovery

For centuries, the scientific method—the fundamental practice of science that scientists use to explain the natural world systematically and logically—has remained largely the same. Machine learning and data science are helping to change that. To this end, I developed a rank aggregation model to prioritize network communities—hypotheses given by groups of putatively related nodes in the network—and
identify the most promising ones for further experimentation and downstream scientific inquiry. This work, published in *Nature Communications* [56], is the first to give a principled method for prioritizing communities in networks, yielding a nearly 50-fold improvement in performance over the traditional practice in which the ordering of communities is ignored.

**Graph convolutional drug discovery**

It can take 15 years and cost $1 billion for a new drug to reach patients as the question of identifying which diseases a new drug (compound) could treat is tremendously complex. Diseases are not independent of each other and a large number of genes are shared between often quite distinct diseases [30,31,36]. Similarly, the effects of drugs are not limited to the molecules to which they directly bind in the body; instead, these effects spread throughout biological networks in which they act. Therefore, the effect of a drug on a disease is inherently a network phenomenon. Leveraging this understanding, I have developed a graph convolutional neural model to predict what diseases a new compound could treat. Crucially, the model represents drugs and diseases as higher-order objects, or subgraphs (instead of single nodes), embedded within a larger genome-wide biological network and then uses the notion of stochastic message passing and graph convolutions for prediction [3]. The model outperforms baselines by up to 172% and has correctly predicted drugs that were repositioned at Stanford SPARK Translational Research Program over the last decade and ranked them in the top 2% of all drugs in the US. I currently work with Stanford School of Medicine and SPARK (Figure 6) to use my models for schizophrenia, a disease for which no good treatments exist.

**Vision for future research**

Looking forward, I aim to develop the next generation of machine learning and data science methods for biomedical data. My research will focus on the following complex technical challenges: designing algorithmic solutions to optimize and manipulate networked systems and to predict their behavior, such as how genomics—nature’s experiments on people—influences human traits in the context of a particular environment; harnessing richer types of interactions and network dynamics; and scaling up the analyses to see structure in massive amounts of data that are too complex to be detected with other methods.

*Algorithmic solutions for predicting behavior of networked systems and manipulating them for useful purposes.* In biomedical problems, achieving high predictive accuracy is often as important as understanding the prediction. An exciting phrase is not only ‘Eureka!’, but also ‘That’s weird—what’s going on?’ as the answer to this question can give cues as to how to re-design and manipulate the existing system for useful purposes. At the same time, new learning capabilities risk leading to biased, inappropriate, or unintended decision making. Here, I would like to continue collaborating with Stanford School of Medicine on drug discovery (Figure 6) and also initiate new collaborations where my methods could enable decision making in high-stakes problems. Through this line of research, I will design transparent and explainable deep models for graph-structured data and develop new means to transform black-box methods into white-box methods whose predictions are accurate and can be interpreted meaningfully.

*Designing contextually adaptive AI to advance health informatics.* In the current wave of statistical learning on graphs, scientists create models for specific problems and train them on large, clearly labeled datasets with millions of data points. These models have nuanced prediction capabilities, but little to no contextual capability and are sensitive to maladaptation. I plan to advance the algorithms to train more with less data, exploit the ability of models to apply prediction prowess acquired from one data type to another type, and design contextual models for classes of phenomena that can learn and reason about never-before-seen systems as they encounter new tasks and situations. Natural case studies here are electronic health record systems, internet-based phenotyping, dosage error reduction, clinical trial participation, and population health studies. The nature of this research will enable me to fund my research program through a variety of sources including science and technology (e.g., NSF, DARPA), biomedical (e.g., NIH), philanthropic (e.g., CZ Biohub), and industry organizations. Throughout my postdoctoral training, I have gained substantial grant writing experience in these domains.

*New architectures for deep learning on rich interaction data and computations over richer types of graphs.* Most algorithms and models today work on standard graph prediction tasks, i.e., node classification, link prediction, and node clustering. An open challenge in this area is developing techniques that can go beyond simple edge prediction and handle more complex prediction problems. My research has made initial steps towards this long-term goal; for example, through a framework that can handle more complex logical queries—a tractable subset of the first-order logic—on incomplete knowledge graphs [33]. Incorporating rich notions of entities, relationships, and combinational generalization into graph neural networks will allow for more accurate and theoretically principled models.

*Scalable data science to transform medicine to a computationally enabled discipline.* Soon, the state of a person will be characterized with increasing precision incorporating data modalities like genetic code, behavior, therapeutics, nutrients, and the environment. These high-dimensional characterizations will lead to far more complex diagnostic and prognostic categories than currently in use, requiring new data science methods for decision making. In this context, I want to define new algorithmic problems and provide solutions for them that can be harnessed by the medical and machine learning communities. To that end, my research has started developing open-source data resources (e.g., BioSNAP) that bring biomedical data closer to other computer scientists who can now use these datasets in algorithm development and benchmarking. I believe that biomedicine is among the most exciting areas for machine learning with many hard problems and applications of immense impact. The statistics community embraced biomedicine as a core application area decades ago and I believe it is time for the machine learning community to adopt it as well.

These directions capture my sense for the beginning of an evolving research program that will allow me to tackle these challenging problems in a unique way. My research on networks is theoretically grounded and spans several areas of computer science including machine learning, bioinformatics, and network science. Implications of my research have direct applications well beyond computer science—in life sciences, biology, medicine, and engineering. I am excited about the influence that my research has already had and look forward to continuing to forge ahead with both methodological foundations and real-world applications.