Machine learning algorithms for spatial and network biology

Uthsav Chitra

High-throughput sequencing technologies are ubiquitous in modern biology, generating over a *zettabyte* (10^9 TB) of data annually [1]. The development of breakthrough technologies – including spatial sequencing, CRISPR-based gene editing, and deep mutational scanning – has enabled scientists to measure diverse molecular modalities (e.g., DNA, RNA, proteins, metabolites) in many biological systems at an unprecedented throughput and resolution. Machine learning (ML) methods are essential for analyzing and interpreting these large and high-dimensional sequencing datasets. However, standard "off-the-shelf" ML methods are severely challenged by the high noise, sparsity, heterogeneity, and other limitations of modern sequencing technologies. Thus, **my research goal is to develop specialized and principled machine learning methods that extract meaningful biological insights from high-dimensional and multi-modal biomedical data**.

My research focuses on developing ML methods to better infer, understand, and model the *spatial* and *network* structure and dynamics of complex biological systems. My work draws on techniques from *deep learning*, *statistical inference*, and *graph theory* and thus far has addressed three fundamental problems:

Modeling spatial gradients and geometry. Spatial transcriptomics (ST) technologies measure gene expression at spatial resolution but suffer from high levels of sparsity ($>$ 70% zeros). I introduced **gene expression topography**, a fundamentally new mathematical paradigm for modeling spatial gradients and tissue geometry in sparse ST data. I developed deep learning methods for learning *"topographic maps"* of 2-D tissue slices [2, 3]. This work builds on my previous model which **introduced complex analysis to spatial biology** [4].

Learning biological interactions. Biological networks are highly incomplete (missing $\approx$ 90% of edges) but are critical for understanding human health and disease. I developed a statistical framework for learning genetic interactions from noisy, high-throughput mutational data. I have used my framework to learn novel interactions between cancer driver mutations [5] (Best Paper Award, RECOMB-CCB) and to learn *epistatic* interactions in yeast and proteins, where **I corrected a major methodological issue in the epistasis literature and extended the known trigenic interaction network by 25%** [6].

Network anomaly detection. Anomalous interactions between genes/proteins underlie many complex diseases. I developed a unified theoretical framework for anomaly detection in networks and other structured data. I proved that the most widely-used algorithms for identifying disease modules from protein-protein interaction (PPI) networks are *statistically biased*, **resolving a nearly 20-year-old open problem in the field** [7]. I developed provably unbiased algorithms for network anomaly detection [8, 9] and structured anomaly detection [10] which achieve state-of-the-art performance in disease gene identification for cancer and other complex diseases.

My research has made **fundamental methodological contributions in machine learning and computational biology**, as recognized by my first-author publications in top machine learning venues (ICML, WSDM) and computational biology venues (Nature Methods, Cell Systems, RECOMB, ISMB) and my awards (Rising Stars in Data Science, Siebel Scholars Award, RECOMB-CCB Best Paper Award, NSF GRFP). Moreover, my work is highly interdisciplinary: in collaboration with biologists, I am using my methods to make **novel and impactful biomedical discoveries in diverse systems** including the brain, liver, and tumor microenvironment.

1 Spatial biology

Motivation. Modern advancements in spatial sequencing technologies have enabled the simultaneous collection of both high-throughput cellular measurements and the spatial location of the measured cells. For example, spatial transcriptomics (ST) technologies (named *"Method of the Year"* by Nature Methods in 2020 [11]) measure the expression of thousands of genes across thousands/millions of spatial locations in a tissue. ST and other spatial sequencing technologies allow researchers to analyze molecular measurements of human cells and tissues in a

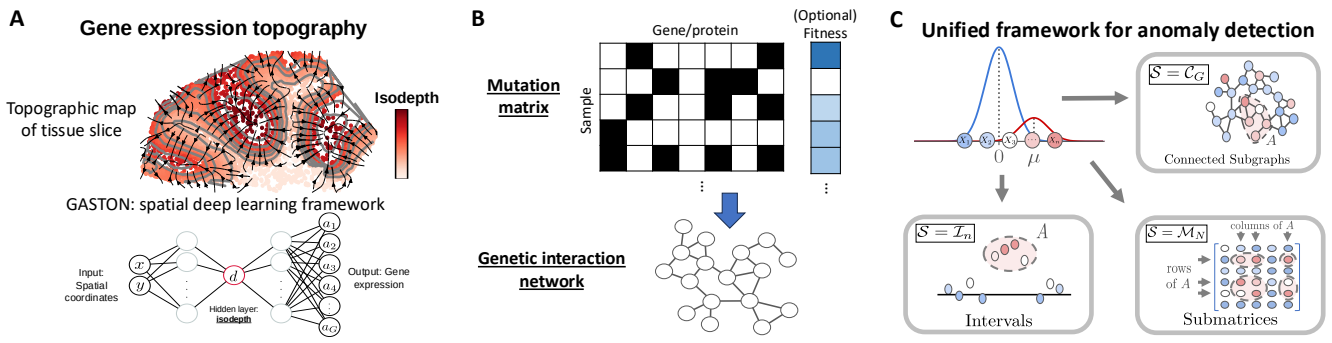

Figure 1: **Highlighted work. (A)** Gene expression topography and GASTON algorithm for learning topographic maps of tissue slices from spatial sequencing data [2]. **(B)** Using high-throughput mutation data ("mutation matrix") to learn functional interactions ("genetic interaction network") [5, 6]. **(C)** Unified statistical framework for anomaly detection in networks and other structured data [8, 9, 10].

spatial context, and present exciting opportunities to understand how cells organize and interact in healthy and diseased tissues (e.g. tumors). However, there are two key challenges in studying spatial biology. First, current spatial technologies make *highly sparse* measurements (e.g. > 70% zeros). Second, spatial variation in gene expression and other molecular measurements results from *multiple, unknown spatial processes* with different intensities and scales (e.g. cell-cell interactions or oxygen gradients) which are difficult to model and infer.

Research contributions. I introduced **gene expression topography** (Figure 1A), a new approach for modeling spatial variation using sparse ST data [2]. I derive a topographic map of a tissue using the *isodepth*, a 1-dimensional coordinate that describes both the geometric arrangement of spatial domains (i.e. clusters) in a tissue slice as well as the relative position of a single location within a domain. Just as a topographic map of a landscape demarcates mountains and valleys by their elevation, our topographic map of gene expression delineates spatial domains within a tissue by their isodepth. Moreover, like elevation of a landscape, the isodepth varies continuously over a tissue slice, providing a coordinate that describes continuous gradients of gene expression. Gene expression topography relies on a novel model of spatial gradients parametrized with a *conservative* gradient field, which is applicable to many areas of spatial statistics. This work generalizes my previous model of spatial gradients in layered tissues, where we parametrize the layer depth of a tissue layer (analogous to isodepth) as the real part of a conformal map [4].

I developed two deep learning algorithms, GASTON [2] and GASTON-Mix [3], for learning topographic maps of 2-D tissue slices using *neural field* and *mixture-of-expert* models, respectively. My methods identify novel spatial gradients and reveal **spatial dynamics of neuronal differentiation/migration and tumor metastasis which were missed by existing approaches.** GASTON was accepted to RECOMB 2024, a top computational biology conference, and is in press at the journal ***Nature Methods***.

Gene expression topography and GASTON have received much interest from the community. I have been invited to present GASTON at seminars at the Broad Institute and University of Toronto. Further, I am currently using GASTON in collaboration with several experimentalists: with Princeton chemists, we are using GASTON to study spatial gradients of metabolites in the liver and intestine; and with Princeton neuroscientists, we are using GASTON-Mix to identify differential spatial gradients across the brains of different primate species.

2 Network biology

Motivation. Biological function is organized by physical and functional interactions between genetic variants, genes, proteins, and other components of biological systems. These interactions form the *biological interaction networks* that underlie human health and tissue function, with aberrations in these networks hypothesized to lead to disease [12]. However, current interaction networks are largely incomplete, e.g. current protein-protein

interaction networks contain $\approx 2-11\%$ of all interactions [13]. To this end, there is a need for methods that (1) **learn biological interaction networks** and (2) **identify anomalous network interactions that lead to disease.**

Research contributions. I have developed novel mathematical frameworks and algorithms to address both of these problems. First, **I developed a statistical framework for learning genetic interactions**, or interactions in which the function of a genetic mutation is altered by the presence or absence of other genetic mutations, from high-throughput mutation data (Figure 1B). Specifically, genetic interactions are described by the so-called "natural" parameters of a *multivariate Bernoulli* distribution, which describes any distribution on binary strings [14].

I applied my multivariate Bernoulli framework to two important types of genetic interactions. First, I used my framework to learn higher-order (> 3-way) genetic interactions (*epistasis*) in yeast and proteins [6]. I showed that many papers in the literature – including two recent high-profile papers in *Science* [15, 16] – incorrectly measure higher-order epistasis using a *"chimeric"* formula that erroneously conflates additive and multiplicative scales. I showed that this widely-used but incorrect epistasis formula corresponds to a parametrization of the multivariate Bernoulli that does *not* measure epistatic interactions, and that using the correct epistasis formula greatly changes reported findings of epistasis. In particular, using the correct epistasis formula results in a 25% increase in known trigenic interactions, **thus extending the known trigenic interaction network by 25%.** Second, I used my statistical interaction framework to infer novel pairwise interactions between cancer driver mutations by parametrizing these interactions using a bivariate Bernoulli. In particular, our model addresses a major deficiency in nearly a decade of existing methods – the conflation of driver mutations with passenger mutations that do not contribute to cancer progression – and **earned a best paper award at the RECOMB satellite workshop on cancer genomics**. These two lines of work build on my earlier theoretical framework for modeling higher-order interactions using hypergraphs which is actively used and cited by other machine learning researchers [17].

To address the second problem, **I developed a unified theoretical framework for anomaly detection in networks and other structured data** (Figure 1C). My model unifies many existing problems in biology, statistics, and epidemiology, including the problems of identifying *disease modules* (groups of interacting disease genes) from protein-protein interaction (PPI) networks. I proved that the two most widely-used algorithms for identifying disease modules (subnetworks) from PPI networks are *statistically biased*, in that they identify much larger subnetworks than expected by chance. My result **resolved a nearly 20-year-old open problem in the field** on why these methods consistently returned large subnetworks [7]. I also derived a mathematical characterization of statistical bias in other structured anomaly detection problems [10]; for example, I showed that the "graph scan statistic", a standard tool for network anomaly detection in epidemiology, is also statistically biased. I addressed these deficiencies by developing several **provably unbiased algorithms for network anomaly detection which achieve state-of-the-art performance in disease gene identification** for complex diseases including cancer and schizophrenia [8, 9]. Separately, I developed a mathematical model for anomalous behavior (polarization) in social network interactions which is well-cited in the social media literature (≈ 200 citations) [18].

Our algorithms for learning and analyzing biological interactions are currently being used for large-scale analysis of multiple lung cancer samples as a part of The Genomic Data Analysis Network (GDAN) project, demonstrating the impact of my work.

References

[1] Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big data: astronomical or genomical? *PLoS biology*, 13(7):e1002195, 2015.

[2] **Uthsav Chitra**, Brian J Arnold, Hirak Sarkar, Cong Ma, Sereno Lopez-Darwin, Kohei Sanno, and Benjamin J Raphael. Mapping the topography of spatial gene expression with interpretable deep learning. ***Nature Methods, in press***, 2024.

[3] **Uthsav Chitra**, Dan Shu, Fenna Krienen, and Benjamin J Raphael. Gaston-mix: a unified model of spatial gradients and domains using spatial mixture-of-experts. *In preparation.*

[4] Cong Ma*, **Uthsav Chitra***, Shirley Zhang, and Benjamin J Raphael. Belayer: Modeling discrete and continuous spatial variation in gene expression from spatially resolved transcriptomics. *Cell Systems*, 13(10):786–797, 2022.

[5] Ahmed Shuaibi*, **Uthsav Chitra***, and Benjamin J Raphael. A latent variable model for evaluating mutual exclusivity and co-occurrence between driver mutations in cancer. *RECOMB Satellite Workshop on Computational Cancer Biology,* Best Paper Award, 2024.

[6] **Uthsav Chitra**, Brian J Arnold*, and Benjamin Raphael. Quantifying higher-order epistasis: beware the chimera. *In press, Nature Communications*, 2024.

[7] Iryna Nikolayeva, Oriol Guitart Pla, and Benno Schwikowski. Network module identification—a widespread theoretical bias and best practices. *Methods*, 132:19–25, 2018.

[8] Matthew A Reyna*, **Uthsav Chitra***, Rebecca Elyanow, and Benjamin J Raphael. Netmix: a network-structured mixture model for reduced-bias estimation of altered subnetworks. *Journal of Computational Biology (RECOMB 2020)*, 28(5):469–484, 2021.

[9] **Uthsav Chitra***, Tae Yoon Park*, and Benjamin J Raphael. Netmix2: A principled network propagation algorithm for identifying altered subnetworks. *Journal of Computational Biology (RECOMB 2022)*, 29(12):1305–1323, 2022.

[10] **Uthsav Chitra***, Kimberly Ding, Jasper CH Lee, and Benjamin J Raphael. Quantifying and reducing bias in maximum likelihood estimation of structured anomalies. In *International Conference on Machine Learning*, pages 1908–1919. PMLR, 2021.

[11] Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature methods*, 18(1):9–14, 2021.

[12] Jessica Xin Hu, Cecilia Engel Thomas, and Søren Brunak. Network biology concepts in complex disease comorbidities. *Nature Reviews Genetics*, 17(10):615–629, 2016.

[13] Katja Luck, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E Begg, Wenting Bian, Ruth Brignall, Tiziana Cafarelli, Francisco J Campos-Laborie, Benoit Charloteaux, et al. A reference map of the human binary protein interactome. *Nature*, 580(7803):402–408, 2020.

[14] Bin Dai, Shilin Ding, and Grace Wahba. Multivariate Bernoulli distribution. *Bernoulli*, 19(4):1465 – 1483, 2013.

[15] Elena Kuzmin, Benjamin VanderSluis, Wen Wang, Guihong Tan, Raamesh Deshpande, Yiqun Chen, Matej Usaj, Attila Balint, Mojca Mattiazzi Usaj, Jolanda Van Leeuwen, et al. Systematic analysis of complex genetic interactions. *Science*, 360(6386):eaao1729, 2018.

[16] Elena Kuzmin, Benjamin VanderSluis, Alex N Nguyen Ba, Wen Wang, Elizabeth N Koch, Matej Usaj, Anton Khmelinskii, Mojca Mattiazzi Usaj, Jolanda Van Leeuwen, Oren Kraus, et al. Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. *Science*, 368(6498):eaaz5667, 2020.

[17] **Uthsav Chitra** and Benjamin Raphael. Random walks on hypergraphs with edge-dependent vertex weights. In *International conference on machine learning*, pages 1172–1181. PMLR, 2019.

[18] **Uthsav Chitra** and Christopher Musco. Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 115–123, 2020.