Statistical models form the backbone of scientific reasoning, enabling new discoveries and decision-making across a wide range of disciplines, from assessing the health risks of air pollution to analyzing biodiversity loss in ecology. Behind these statistical models, however, are methodological frameworks that are often built on idealized assumptions and do not adapt well to large-scale, high-dimensional datasets. There is an ongoing need to enhance the robustness, interpretability, and scalability of statistical and machine learning methods for analyzing complex, high-dimensional data, which motivates the core question of my research:

> *"How can we develop statistical and machine learning methods that are* ***robust*** *to modeling assumptions,* ***informative*** *for scientific decision-making, and* ***scalable*** *to complex, high-dimensional data?"*

Here, I outline my research in three parts: theoretical and methodological developments, application-driven works addressing unique scientific questions across domains, and ongoing and future directions that build on these foundations. [1]

Theory and methods: Scalable, robust, and interpretable methods

Scalable and robust statistical models. All models are wrong in the sense that assumptions are never exactly met, but some models are particularly vulnerable to such deviations. Beta regression for modeling continuous proportional data $Y \in [0, 1]$ (e.g. rates, indices) is one such example, where inference and prediction are highly sensitive to outliers and model misspecification, and scalability to hierarchical model settings is limited.

In [1], I **developed a new class of generalized linear models** for continuous proportional data that addresses these limitations. The proposed models exhibit strong robustness properties, including consistency of MLE under potential model misspecification and resilience to outliers. A key innovation is a novel data augmentation strategy that transforms the model into a conditionally Gaussian form, enabling highly efficient computation via expectation-maximization (EM) and Markov chain Monte Carlo (MCMC) algorithms with theoretical guarantees. This framework scales naturally to large, spatially structured datasets, settings where beta regression typically fails. This work, also implemented as the R package `cobin` at CRAN [2], received the *Best Long Talk Award* at the 2025 Bayesian Young Statisticians Meeting (ISBA Junior Section) and is currently under revision at the *Journal of the American Statistical Association (Theory and Methods)*. Also, an ongoing line of work [3] extends this focus on scalability and robustness to the problem of **probabilistic graph clustering**. In particular, it aims to relax the strong generative assumptions of stochastic block models for clustering, while maintaining robustness to outliers and enabling efficient computation based on random spanning tree models.

Scalable and interpretable statistical models. Nonparametric methods are widely used to avoid restrictive parametric assumptions, but their flexibility often comes at the cost of interpretability, sensitivity to hyperparameter choices, and computational intractability for complex settings. A key challenge is to preserve the flexibility of nonparametric approaches while improving interpretability and ensuring computational efficiency.

Focusing on a density regression problem, where the conditional density of a response is modeled flexibly as a function of covariates, I developed a dependent Bayesian nonparametric model for density regression [4], recently accepted for publication in *Bayesian Analysis*. The proposed model enables **scalable density regression** with uncertainty quantification, while also mitigating sensitivity to hyperparameter choices and enhancing model interpretability.

Also, when a statistical model contains both parametric and nonparametric components, such as

[1] Research conducted as a Ph.D. candidate advised by Dr. Huiyan Sang (Texas A&M), graduate research assistant mentored by Dr. Eun Sug Park (Texas A&M Transportation Institute), postdoctoral associate advised by Dr. David Dunson (Duke), and collaborations with Dr. Elaine Symanski (Baylor College of Medicine, Epidemiology), Dr. Philip Hopke (U. of Rochester, Chemistry), Dr. Otso Ovaskainen (U. of Jyväskylä, Ecology) and many others.

mixed models with spatial random effects, careful specification of the nonparametric component can enhance the interpretability of the parametric part of the model. In [5], I developed a spatial logistic regression model that **enables both population-averaged and subject-specific odds ratio interpretations** of the regression coefficients, achieved by adopting a new class of non-Gaussian spatial processes. This is in contrast to traditional Gaussian process-based methods, which only provide subject-specific interpretations. The work also brings scalable computation strategies and is currently under revision at *Statistics in Medicine*.

Statistical machine learning methods with solid theoretical foundations. My research in statistical machine learning focuses on developing methods and algorithms that are theoretically grounded and broadly applicable across a wide range of tasks. One of the central goal is to improve interpretability alongside statistical rigor, ensuring that the resulting tools remain useful for domain-specific decision making.

In the context of probabilistic clustering, a widely used approach in various domains where the number of clusters is unknown, a common implicit assumption is the "rich-get-richer" property, which tends to favor imbalanced cluster sizes. Although often accepted by default, this assumption can be undesirable in many applications where more balanced clustering is preferred, yet it had remained unclear why such an assumption appears necessary and what principled alternatives might exist. In [6], presented and published at the *International Conference on Machine Learning (ICML)*, I introduced a theoretical framework that **identifies implicit assumptions driving the "rich-get-richer" behavior** in probabilistic clustering. I also proposed a general strategy for designing methods with **tunable balancedness properties**, allowing the clustering behavior to better match the needs of different applications.

Bayesian model selection is another area where I have contributed to the development of scalable and theoretically grounded machine learning algorithms. Despite its foundational role, high-dimensional Bayesian model selection problems remains computationally challenging, such as variable selection and structure learning, where the model space grows exponentially with the number of variables. While MCMC methods are popular, their inherently sequential nature limits scalability.

In [7], using the multiple-try Metropolis with N parallel trials and adopting a certain family of proposal kernels, we **proved that the mixing time bound can be improved by a factor of** N in the context of model selection problems. In other words, the number of MCMC iterations required for convergence is reduced by a factor of N, **overcoming the fundamental limitations of MCMC's inherently sequential nature by harnessing parallel computational tools** and offering significant practical gains. This work was presented at and published in *Advances in Neural Information Processing Systems (NeurIPS)*, where it was selected for an *oral* presentation (top 1.9% of all submissions).

Applications: Capturing complex spatial dependence structures

The growing availability of geocoded data, with advances in geographic information systems and remote sensing technologies, has made spatial and spatiotemporal analysis increasingly accessible, leading to important findings across scientific disciplines. I have developed novel statistical methods to address the unique challenges involved in analyzing complex spatial data across a broad range of applications.

Air pollution epidemiology. Exposure to air pollution is linked to approximately 7 million premature deaths annually, according to the World Health Organization. Ambient air pollutants (e.g. speciated PM2.5, volatile organic compounds) are typically measured at a limited number of monitoring stations, posing several statistical challenges due to sparse exposure data, complex spatial dependence structures, and high correlations between pollutants arising from a common source (e.g., vehicle exhaust). When analyzing the health effects of environmental exposures, one of the key challenges is accounting for measurement error caused by spatial/temporal misalignment between exposure and health outcome data, as well as incorporating the uncertainty in the exposure estimates into the health data model. Previous studies have shown that failure to properly address exposure measurement error can lead to biased health effect estimates as well as incorrect uncertainty quantification, but existing methods often scale poorly or ignore spatial dependence.

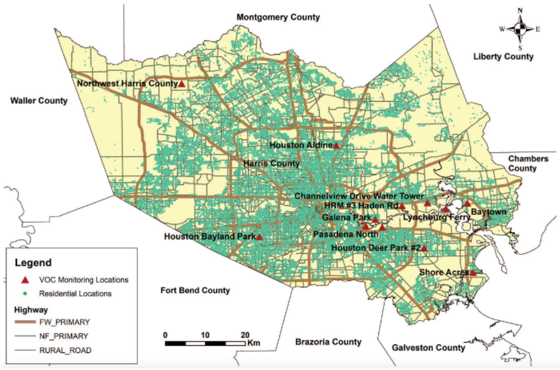

Fig.3 of [8]: Volatile organic compound monitoring stations (red) and residential locations (green) in Houston metropolitan area.

In [8], we **proposed an efficient uncertainty propagation method** to account for exposure measurement error that scales well to massive health data settings while maintaining minimal bias and reliable coverage probabilities. Applying this approach, we identified associations between source-specific/pollutant-specific exposures and reduced birth weights with better quantification of uncertainty in exposure estimates. This work is published at *Biostatistics* and was recognized with the *Early Career Award* from the American Statistical Association's Section on Statistics in Epidemiology.

More recently, motivated by the abundance of traffic and land-use data in contrast to spatial sparsity of monitoring stations, we have been developing a covariate-informed source apportionment model to **better quantify the apportionment of air pollutants to their sources**, which not only explains the association between traffic covariates and source-specific exposures, but also better characterizes small-scale spatial variation in exposure.

Network science and medical imaging. Network science is a rapidly growing field that provides a flexible framework for capturing structure and relationships in complex datasets through graphs. For example, traffic patterns can be encoded in spatial road networks, while brain activity measured by electroencephalography (EEG) or functional MRI (fMRI) can be represented as networks, where nodes correspond to electrodes or brain voxels. Analyzing such data requires innovative statistical approaches that account for their unique, nontrivial geometries.

In [9], we developed a **probabilistic clustering and variable selection method for graph-structured data** based on a Bayesian hierarchical model involving random spanning trees, applicable to a variety of settings while providing interpretable results through clustering. We focused on the problem of anomaly detection in road networks caused by an event, where the proposed method identified a previously undetected affected area along with uncertainty quantification. This work was presented at and published in *NeurIPS*. The proposed method is also applicable to **scalar-on-image regression tasks, which are commonly encountered in medical imaging analysis**. Recent studies involving EEG and fMRI data have extensively used the proposed method as a benchmark [10] [11], highlighting its fundamental role in advancing statistical modeling for biomedical imaging studies.

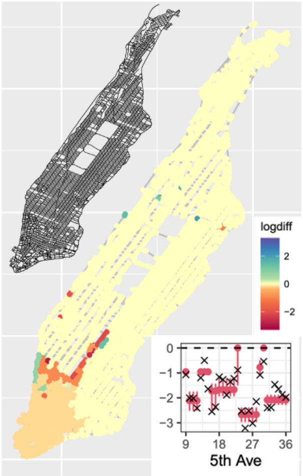

Fig.5 of [9]: Estimated anomaly signal on road network of Manhattan.

Community ecology. Identifying natural and human-driven factors of biodiversity loss is a central goal in community ecology, where spatial statistics plays a key role by **accounting for spatial correlation** in geographically indexed data, such as species diversity of aquatic macroinvertebrates across U.S. lakes. Ecological indices are typically bounded (e.g., between 0 and 1), motivating the use of appropriate regression models. We applied our proposed new class of generalized linear models [1] to a large-scale spatial dataset and successfully identified both natural and anthropogenic drivers of lake biodiversity, **providing more robust and reliable** results than beta regression.

Ongoing works and future directions

While my research interests span a broad range of methodological and applied areas, I plan to focus on two directions over the next few years: developing probabilistic machine learning methods for scalable probabilistic learning and prediction, and building novel statistical models for complex data arising in neuroscience and molecular biology.

Amortized inference and in-context scalable Bayesian prediction. Large-scale pretrained models, often referred to as foundation models, are becoming increasingly popular in modern machine learning. A particularly promising recent development is **prior-data fitted networks** [12, 13], which introduces a new paradigm for Bayesian prediction. These models are pretrained on synthetic data generated from a wide variety of plausible data-generating mechanisms, including different statistical models, noise levels, and parameter configurations, and then perform probabilistic prediction via a single forward pass in a few-shot setting, **without updating model parameters**. This form of **amortized inference** shifts the computational burden to the pretraining stage, enabling fast and scalable prediction. It significantly broadens the scope of probabilistic modeling, especially in settings where specifying an explicit statistical model or evaluating likelihoods is infeasible or computationally intractable. By reframing statistical model design as a problem of generating diverse, plausible synthetic training data, this approach offers a new perspective on how to perform flexible and scalable Bayesian prediction.

One area where I am actively exploring this idea is **probabilistic sound source localization**, as part of a collaborative effort with Finnish ecologists to develop an **autonomous bird biodiversity monitoring system**. By generating synthetic training data from a range of plausible sound propagation models and microphone array configurations, we develop models that perform localization with uncertainty quantification from synchronized bird vocalization recordings, enabling fast, automated inference of bird locations at scale.

Brain structural and functional connectomics. Understanding how the brain is organized and connected, both structurally and functionally, provides critical insight into human cognition, behavior, and disease. Neuroimaging studies often involve high-dimensional structured data, such as diffusion MRI or fMRI, where modeling connectivity across brain regions requires accounting for complex spatial dependence structures. One of my ongoing collaborative projects with epidemiologists at the University of North Carolina at Chapel Hill examines **how early life exposure to phthalates is associated with brain development**. Brain voxels are parcellated into distinct regions of interest, and a key statistical challenge lies in inferring associations between chemical exposures and developmental outcomes while **accounting for dependence structure across brain regions**, in a way that yields results that are both robust and interpretable.

Spatial transcriptomics and proteomics. Spatial transcriptomics and spatial proteomics have opened up a new area of research by providing molecular expression from tissue along with spatial information, offering a new approach to uncover the relationships between structural organization and function in biological systems. However, such spatial omics data come with **high dimensionality and strong spatial autocorrelation**, making it challenging to perform interpretable and reliable statistical inference. Building on my research in spatial statistics and probabilistic clustering, I plan to develop new statistical methodologies for analyzing spatial omics data in collaboration with domain scientists. One particularly promising direction is **spatial clustering of transcriptomic and proteomic profiles**, which can reveal spatially contiguous molecular patterns that are clinically meaningful and interpretable.

References

[1] Changwoo J Lee, Benjamin K Dahl, Otso Ovaskainen, and David B Dunson. “Scalable and robust regression models for continuous proportional data”. In: *arXiv preprint arXiv:2504.15269* (2025). URL: https://arxiv.org/abs/2504.15269.

[2] Changwoo Lee, Benjamin Dahl, Otso Ovaskainen, and David Dunson. *cobin: Cobin and Micobin Regression Models for Continuous Proportional Data*. R package version 1.0.1.3. 2025. URL: https://CRAN.R-project.org/package=cobin.

[3] Changwoo J. Lee. “Chapter 5: Graph product partition models for robust probabilistic graph clustering”. In: *Doctoral dissertation, Texas A&M University* (2024). URL: https://oaktrust.library.tamu.edu/items/bd682fba-79d7-4b2d-b895-728444431122.

[4] Changwoo J Lee, Alessandro Zito, Huiyan Sang, and David B Dunson. “Logistic-beta processes for dependent random probabilities with beta marginals”. In: *Bayesian Analaysis (in press)* (2025). URL: https://doi.org/10.1214/25-ba1541.

[5] Changwoo J Lee and David B Dunson. “Marginally interpretable spatial logistic regression with bridge processes”. In: *arXiv preprint arXiv:2412.04744* (2024). URL: https://arxiv.org/abs/2412.04744.

[6] Changwoo J. Lee and Huiyan Sang. “Why the rich get richer? On the balancedness of random partition models”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 12521–12541. URL: https://proceedings.mlr.press/v162/lee22j.html.

[7] Hyunwoong Chang*, Changwoo J. Lee*, Zhao Tang Luo, Huiyan Sang, and Quan Zhou. “Rapidly mixing multiple-try Metropolis algorithms for model selection problems”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 25842–25855. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/a600cdf3a53f93bcb85cb37343a8d831-Abstract-Conference.html.

[8] Changwoo J Lee, Elaine Symanski, Amal Rammah, Dong Hun Kang, Philip K Hopke, and Eun Sug Park. “A scalable two-stage Bayesian approach accounting for exposure measurement error in environmental epidemiology”. In: *Biostatistics* 26.1 (2025), kxae038. URL: https://doi.org/10.1093/biostatistics/kxae038.

[9] Changwoo J. Lee, Zhao Tang Luo, and Huiyan Sang. “T-LoHo: A Bayesian Regularization Model for Structured Sparsity and Smoothness on Graphs”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 598–609. URL: https://proceedings.neurips.cc/paper/2021/hash/05a70454516ecd9194c293b0e415777f-Abstract.html.

[10] Zikai Lin, Junsouk Choi, Ruoxuan Mao, Bangyao Zhao, and Jian Kang. “Spatial Adaptive Selection using Binary Conditional Autoregressive Model with Application to Brain-Computer Interface”. In: *Journal of Computational and Graphical Statistics* 0.0 (2025), pp. 1–12. URL: https://doi.org/10.1080/10618600.2025.2495256.

[11] Yuliang Xu and Jian Kang. “Bayesian Image Regression with Soft-thresholded Conditional Autoregressive Prior”. In: *International Conference on Learning Representations* (2025). URL: https://openreview.net/forum?id=rnL3OafDdw.

[12] Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. “Transformers can do Bayesian inference”. In: *International Conference on Learning Representations* (2022). URL: https://openreview.net/forum?id=KSugKcbNf9.

[13] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. “Accurate predictions on small data with a tabular foundation model”. In: *Nature* 637.8045 (2025), pp. 319–326. URL: https://doi.org/10.1038/s41586-024-08328-6.