Research Statement

Jae Ho Chang

Vision. My long-term goal is to develop a unified statistical theory explaining how *dependence*, *heterogeneity*, and *decentralization* shape the limits of learning and inference. Modern data violate classical independence assumptions: observations interact through networks, arise from distinct sensing modalities, and reside across distributed systems. These structural constraints challenge identifiability, asymptotic validity, and reproducibility. My research integrates tools from statistics, optimization, and information theory to construct learning principles that remain reliable and interpretable in such environments. Ultimately, I aim to lead a group developing rigorous guarantees for structured learning systems, clarifying the trade-offs among accuracy, privacy, and generalizability.

Past and Current Research. My work develops statistical frameworks that quantify how structural dependence and missing information distort inference, and establishes principles that correct such distortions. Across three research programs—network autoregression, heterogeneous transfer learning, and federated empirical Bayes—I build methodological foundations for trustworthy inference under modern data complexity.

(1) Embedding Network Autoregression (ENAR) provides a unified framework for network time-series analysis and causal peer-effect inference. Peer effects measure how individual behaviors propagate through social or biological networks, but such inference is often confounded by unobserved homophily—latent similarities that influence both tie formation and outcomes. ENAR addresses this issue by embedding these latent factors directly into the autoregressive dynamics through latent-space random-graph models. We prove asymptotic normality of the regression estimators under both finite- and infinite-time regimes; the first corresponds to peer-effect estimation and the second to networked time-series forecasting. Our theoretical results show that omitting or under-specifying the latent dimension K introduces asymptotic bias in peer-effect estimation, while over-selection is asymptotically negligible. These findings provide concrete guidance for dimension selection and quantify the inferential risk of ignoring latent homophily. Empirical studies—including behavioral contagion among 7,000 students and spatial wind forecasting—demonstrate substantial causal and predictive gains. This work is accepted (conditioned on minor revision) at the Journal of Machine Learning Research. This paper is accepted (conditioned on minor revision) at the *Journal of Machine Learning Research*.

(2) Heterogeneous Transfer Learning (HTL) investigates how to transfer information across datasets with non-overlapping feature spaces. In many scientific studies, two related datasets share outcome and covariate structures but differ in their measured features—creating feature-level heterogeneity that prevents direct model transfer. By learning feature mappings through proxy domains and correcting for cross-domain discrepancy, we obtain finite-sample bounds that reveal how transfer efficiency depends on mapping and model complexity. Applications to ovarian-cancer survival improve prediction by over 15% relative to approaches that ignore feature mismatch. Ongoing work develops nonlinear sieve-based mappings for multi-omics integration. Ongoing work extends this framework with data-integration pipelines in metabolomics

and multi-omics applications.

(3) Federated Learning with Personalization. Building on this foundation, I am developing a statistical theory for decentralized learning systems in which local data remain private and only model summaries are shared with a central server. Each client's estimator is modeled as a noisy sample from an unknown population prior G. The server estimates this prior using Nonparametric Maximum Likelihood (NPMLE) and Empirical Bayes principles. This produces a personalized update rule,

$$\hat{\theta}_k \leftarrow \hat{\theta}_k + \mathbb{E}_{s \sim \hat{G}}(s - \hat{\theta}_k \mid \hat{\theta}_k),$$

which shrinks each local estimator toward the empirical-Bayes population mean and effectively denoises individual clients through population-level calibration. I am establishing an oracle denoising inequality that provides non-asymptotic risk bounds for estimating the Bayes rule when client-level estimators are M-estimators. This result connects classical empirical-process theory with modern federated systems and formalizes how privacy-preserving information exchange can provably improve estimation efficiency.

Collectively, these projects build a coherent program on trustworthy statistical learning under dependence, heterogeneity, and decentralization.

Future Research Agenda. My current results establish consistency and finite-sample guarantees. The next phase of my research will focus on asymptotic efficiency, adaptive generalization, and computational scalability in dynamic, heterogeneous environments.

Aim 1: Dynamic Network Processes. I will extend ENAR to dynamic networks where edges evolve jointly with node-level outcomes. Real networks—social, economic, or biological—change as agents interact, creating time-varying dependence both across units and across time. I plan to develop asymptotic theory for nonlinear autoregressive models on evolving random graphs governed by dynamic latent-space processes,

$$y_{it} = f(y_{i,t-1}, z_{i,t-1}, A_{t-1}, u_{i,t-1}) + \epsilon_{it}.$$

I will characterize the stochastic evolution of latent variables u_{it}, including their temporal drift and diffusion. The goal is to quantify how this evolution propagates through the autoregressive mechanism, leading to time-varying measurement error and dynamic bias in peer-effect estimation. This framework will unify diffusion and contagion modeling under a rigorous statistical lens and connect temporal network statistics with causal inference.

Aim 2: Unified Transfer and Federated Learning for Heterogeneous Multi-Modal Data. Modern health and environmental systems collect multi-modal data—electronic health records (EHR), imaging, wearables, and genomics—across institutions, each with its own subset of modalities and population. This setting introduces two layers of heterogeneity: (i) *feature heterogeneity*, due to missing or mismatched modalities; and (ii) *distributional heterogeneity*, since each institution follows a distinct data-generating process. I will construct a statistical framework that jointly addresses both layers. At the feature level, I will extend HTL to align site-specific representations through proxy domains where multiple modalities co-occur, providing theoretical guarantees for feature imputation and mapping accuracy. At the distributional level, I will extend the federated empirical Bayes framework to model site-level estimators as noisy draws from a population distribution. Each site fits its local model and transmits summary estimates; the server infers the population prior via NPMLE and returns personalized shrinkage updates. This structure enables both privacy-preserving learning and adaptive calibration, ensuring that local estimators benefit from shared information while preserving local specificity. Integrating these layers yields a unified, nonparametric pipeline for improving target-domain prediction in multi-institutional, multi-modal systems.

Aim 3: Theory-to-Algorithm Translation and Applications. Finally, I will translate these theoretical results into scalable algorithms for large-scale biomedical and social network data. The goal is to design procedures that preserve provable guarantees—finite-sample risk control, asymptotic efficiency, and privacy protection—while remaining computationally tractable. These algorithms will serve as a statistical backbone for reproducible, privacy-aware machine learning. Potential applications include federated causal inference in precision medicine, decentralized model calibration across hospitals, and dynamic social contagion analysis.

Funding and Broader Impact. My proposed research aligns with several NSF programs, including the Division of Mathematical Sciences (DMS)–Statistics, the Mathematical Foundations of AI, and the Mathematics of Digital Twins initiatives. NIH programs such as Smart Health and Biomedical Research in the Era of Artificial Intelligence and Advanced Data Science (SCH) and the Joint DMS–National Institute of General Medical Sciences (NIGMS) initiative are also my potential funding sources. Lastly, Patient-Centered Outcomes Research Institute (PCORI) funding for EHR integration particularly aligns with my interest in federated and transfer learning. As a junior faculty member, I plan to lead a group advancing the theory-to-algorithm pipeline for the research agendas in this statement. I will mentor graduate students at the interface of statistics and data science, and foster collaborations that connect mathematical rigor with impactful applications. Ultimately, my goal is to help shape a modern statistical paradigm in which rigor, interpretability, and reproducibility remain central as data systems become increasingly complex, high-dimensional, and autonomous.