Sophia Tang

sophtang@seas.upenn.edu | Google Scholar | LinkedIn | sophtang.github.io

EDUCATION

University of Pennsylvania — Philadelphia, PA
Dual Degree with Bachelor of Science in Engineering and Bachelor of Science in Economics — Sept 2023 - May 2027
Jerome Fisher Program in Management & Technology; Major in Computer Science & Statistics
Relevant Coursework: Algorithms, Programming Languages and Techniques I and II, Mathematics of Computer Science, Multivariable Calculus, Linear Algebra, Machine Learning, Computer Systems, Automata Computability and Complexity, Corporate Finance, Database & Information Systems, Statistical Inference

RESEARCH EXPERIENCE

Chatterjee Lab, University of Pennsylvania. — Philadelphia, PA
Undergraduate Researcher. **Advised by Dr. Pranam Chatterjee**. — Sept 2024 - Present
- Developing theoretical ML frameworks for generative modeling with applications in biological design and simulation.
- **First-author paper presented at main ICML 2025** and **multiple first-author papers under review at top ML venues** (**see below**).

Mitchell Lab, University of Pennsylvania. — Philadelphia, PA
Undergraduate Researcher. **Advised by Dr. Michael Mitchell.** — Sept 2023 - Present
- Developed novel targeted lipid nanoparticle (LNP)-encapsulated mRNA therapeutics with high transport across the blood-brain barrier after systemic administration and transfection into brain cell types.
- Synthesized and characterized ionizable LNPs with targeted peptide groups for in vitro and in vivo experiments. Evaluated relevant literature on targeted brain delivery techniques and planned experiments.
- **Several co-authored papers** and a **first-author review paper published (see below)**

MACHINE LEARNING PUBLICATIONS (*denotes equal contribution; † denotes advising role)

[1] PepTune: De Novo Generation of Therapeutic Peptides with Multi-Objective-Guided Discrete Diffusion.
Sophia Tang*, *Yinuo Zhang*, Pranam Chatterjee*†.
42nd International Conference on Machine Learning (ICML 2025), Dec 2024. **https://openreview.net/forum?id=FQoy1Y1Hd8**
Description: We introduce PepTune, a multi-objective discrete diffusion model for simultaneous generation and optimization of therapeutic peptide SMILES. Built on the Masked Discrete Language Model (MDLM) framework, PepTune ensures valid peptide structures with a novel bond-dependent masking schedule and invalid loss function. To guide the diffusion process, we introduce Monte Carlo Tree Guidance (MCTG), an inference-time multi-objective guidance algorithm that balances exploration and exploitation to iteratively refine Pareto-optimal sequences. MCTG integrates classifier-based rewards with search-tree expansion, overcoming gradient estimation challenges and data sparsity.

[2] Branched Schrödinger Bridge Matching.
Sophia Tang, *Yinuo Zhang, Alexander Tong*†, Pranam Chatterjee*†.
Preprint. Jun 2025. **https://doi.org/10.48550/arXiv.2506.09007**
Description: We introduce Branched Schrödinger Bridge Matching (BranchSBM), a novel framework that learns branched Schrödinger bridges. BranchSBM parameterizes multiple time-dependent velocity fields and growth processes, enabling the representation of population-level divergence into multiple terminal distributions. We show that BranchSBM is not only more expressive but also essential for tasks involving multi-path surface navigation, modeling cell fate bifurcations from homogeneous progenitor states, and simulating diverging cellular responses to perturbations.

[3] TR2-D2: Tree Search Guided Trajectory-Aware Fine-Tuning for Discrete Diffusion.
Sophia Tang*, *Yuchen Zhu*, Molei Tao*†, Pranam Chatterjee*†.
Preprint. Sept 2025. **https://doi.org/10.48550/arXiv.2509.25171**
Description: We introduce a novel framework that optimizes reward-guided discrete diffusion trajectories with tree search to construct replay buffers for trajectory-aware fine-tuning. These buffers are generated using Monte Carlo Tree Search (MCTS) and subsequently used to fine-tune a pre-trained discrete diffusion model under a stochastic optimal control objective. We validate our framework on single- and multi-objective fine-tuning of biological sequence diffusion models, highlighting the overall effectiveness of TR2-D2 for reliable reward-guided fine-tuning in discrete sequence generation.

[4] Entangled Schrödinger Bridge Matching.
Sophia Tang, *Yinuo Zhang, Pranam Chatterjee*†.
Preprint. Oct 2025. **https://doi.org/10.48550/arXiv.2511.07406**
Description: We introduce a framework that learns the first- and second-order stochastic dynamics of interacting, multi-particle systems where the direction and magnitude of each particle's path depend dynamically on the paths of the other particles. We define the Entangled Schrödinger Bridge problem as solving a coupled system of bias forces that entangle particle velocities. We show that our framework accurately simulates heterogeneous cell populations under perturbations and rare transitions in high-dimensional biomolecular systems.

[5] Gumbel-Softmax Flow Matching with Straight-Through Guidance for Controllable Biological Sequence Generation.
Sophia Tang, *Yinuo Zhang, Alexander Tong†, Pranam Chatterjee†.*
Preprint. Mar 2025. **https://doi.org/10.48550/arXiv.2503.17361**
Description: We introduce a generative framework on the simplex based on a novel Gumbel-Softmax interpolant with a time-dependent temperature. Our framework enables high-quality, diverse generation and scales efficiently to higher-dimensional simplices.

[6] Multi-Objective-Guided Generative Design of mRNA with Therapeutic Properties.
Sawan Patel, **Sophia Tang***, Yinuo Zhang, Pranam Chatterjee†, and Sherwood Yao†.*
Preprint. Jul 2025. **https://openreview.net/pdf?id=Jiv4B97NAn**
Description: We introduce the first multi-objective guided generative model for simultaneous mRNA codon optimization and de novo design of untranslated region (UTR) sequences.

[7] Multi-Objective-Guided Discrete Flow Matching for Controllable Biological Sequence Design.
*Tong Chen, Yinuo Zhang, **Sophia Tang**, and Pranam Chatterjee†.*
Preprint. May 2025. **https://doi.org/10.48550/arXiv.2505.07086**
Description: We introduce a multi-objective guidance method for discrete flow matching that leverages rank-based directional scoring and hypercone filtering to guide pre-trained discrete flow velocities to Pareto-optimal sequences.

BIOENGINEERING PUBLICATIONS

[8] Peptide-functionalized nanoparticles for brain-targeted therapeutics.
Sophia Tang, *Emily L. Han, and Michael J. Mitchell†.*
Drug Delivery and Translational Research, Springer Nature. Mar 2025. **https://doi.org/10.1007/s13346-025-01840-w**

[9] Peptide-Functionalized Lipid Nanoparticles for Targeted Systemic mRNA Delivery to the Brain.
*Emily L. Han, **Sophia Tang**, Dongyoon Kim, Amanda M. Murray, Kelsey L. Swingle, Alex G. Hamilton, Kaitlin Mrksich, Marshall S. Padilla, Rohan Palanki, Jacqueline J. Li, and Michael J. Mitchell†.*
ACS Nano Letters. Dec 2024. **https://doi.org/10.1021/acs.nanolett.4c05186**

[10] Predictive High-Throughput Platform for Dual Screening of mRNA Lipid Nanoparticle Blood–Brain Barrier Transfection and Crossing.
*Emily L. Han, Marshall S. Padilla, Rohan Palanki, Dongyoon Kim, Kaitlin Mrksich, Jacqueline J. Li, **Sophia Tang**, Il-Chul Yoon, and Michael J. Mitchell†.*
ACS Nano Letters. Jan 2024. **https://doi.org/10.1021/acs.nanolett.3c03509**

[11] Optimized microfluidic formulation and organic excipients for improved lipid nanoparticle-mediated genome editing.
*Rohan Palanki, Emily L. Han, Amanda M. Murray, Rohin Maganti, **Sophia Tang**, Kelsey L. Swingle, Dongyoon Kim, Hannah Yamagata, Hannah C. Safford, Kaitlin Mrksich, William H. Peranteau, and Michael J. Mitchell†.*
Lab on a Chip, Royal Society of Chemistry. Jan 2024. **https://doi.org/10.1039/D4LC00283K**

TECHNICAL ARTICLES

A Complete Guide to Spherical Equivariant Graph Transformers
https://alchemybio.substack.com/p/spherical-equivariant-graph-transformer
A 2.5-hour breakdown of the theory behind spherical equivariant (SE(3)) graph neural networks (EGNNs) and the SE(3)-Transformer architecture.

A Complete Guide to Protein Folding Prediction
https://alchemybio.substack.com/p/a-complete-guide-to-protein-folding
A 3-hour breakdown of the three-track RoseTTAFold protein prediction model that leverages multi-modal deep learning architectures to transform single sequences into dynamic 3D structures.

INVITED TALKS

Discrete Diffusion Reading Group. Nov 2025.
https://youtu.be/NX2Mth_Epvk?si=LsVS_bhzyD-RV03m

Starkly Speaking Reading Group. Jul 2025.
https://youtu.be/inVYA0pQ4Wg?si=mdwhTE3wu0vXJ4ZX

Learning on Graphs and Geometry Reading Group. Jan 2025.
https://youtu.be/KVr8ryclwdA?si=RdRPdqNqvmblk3Og

WORKSHOP PRESENTATIONS

FPI, SPIGM (NeurIPS 2025)
GenBio, ExAIT, FM4LS, SIM (ICML 2025)
DeLTa, FPI, GEM, AI4NA (ICLR 2025)