Research Statement

Minh Hoang
minhhoang@princeton.edu

Overview

Biological sequences lie at the foundation of life. DNA and RNA encode the genetic instructions that guide cellular processes, whereas proteins carry out the structural and functional work that sustains organisms. Advances in sequencing technologies have enabled us to collect these data at unprecedented rates. This wealth of sequence information has transformed biology into a data-driven science, offering opportunities to uncover principles of evolution, disease mechanisms, and novel treatments through computational approaches. However, the sheer scale and complexity of biological sequences also present formidable challenges. For example, single-nucleotide variants in a genome spanning billions of bases can sometimes lead to major differences in traits or disease. Proteins fold into complex three-dimensional shapes, where their functions arise from both local and long-range interactions. Extracting insights from such data requires computational representations that are both efficient enough to handle massive datasets, and expressive enough to capture the nuanced biological signals embedded in sequences.

I approach these challenges through the lens of **biological sequence representation learning**, drawing on techniques from machine learning, artificial intelligence, and algorithmic design. Working at the interface of **ML/AI and computational biology**, I develop methods that both push the boundaries of machine-learning research and open new avenues for biological discovery. Specifically, my work spans three interconnected thrusts: (1) **learnable sequence compression algorithms** that facilitate scalable pipelines that involve biological sequences; (2) **representation learning for biological foundation models (BFMs)** that improves structure/function understanding across proteins and genomes; and (3) **federated learning frameworks** that enable collaborative training of BFMs on distributed, privacy-sensitive datasets across institutions. My overarching goal is to build scalable, context-aware, and privacy-preserving models that advance fundamental biological discovery while serving as shared computational infrastructure for the life sciences.

Learnable Compression of Biological Sequences

Long DNA sequences pose both computational and statistical challenges for downstream biological analysis. Minimizers, a widely used technique for constructing k-mer sketches, compress long genomes into representative subsequences suitable for indexing, alignment, and assembly. However, classical minimizer design is driven by heuristics that are provably suboptimal for real genomic data.

My research introduced the first **learnable framework for minimizer construction**. In **DeepMinimizer** [5], I reconceptualized minimizer selection as a differentiable optimization problem, enabling the use of gradient-based methods to learn effective sketches. DeepMinimizer coordinates two neural networks that capture distinct structural properties of ideal sketches, optimizing them jointly to produce compact and sequence-specific minimizer schemes. This approach achieves **up to 40% reduction in sketch size** on human genomic benchmarks compared to state-of-the-art baselines. Beyond performance, this work opens a new paradigm in which classical bioinformatics algorithms are re-designed using continuous optimization, providing a flexible framework for incorporating biological context and downstream objectives.

Building on this foundation, I developed extensions that unify diverse sketching methods within a shared differentiable system and introduced new sketching capabilities, such as locality-sensitive minimizers [3]. These advances demonstrate that many sequence algorithms historically constrained by discrete design can be revisited and substantially improved through modern machine learning. In the long term, learnable compression can help BFMs scale to full-genome contexts by providing compact yet informative representations that preserve long-range dependencies, which is an active area of research in BFMs.

Representation Learning for Biological Foundation Models

While minimizers address efficiency, understanding the biochemical and evolutionary characteristics of biological sequences requires richer representations. BFMs trained on millions of DNA, RNA, and protein sequences have emerged as powerful tools for predicting structure, function, mutational effects, and other properties directly from sequence. However, fine-tuning BFMs for downstream applications remains challenging. A key bottleneck lies in converting residue-level embeddings into global sequence-level representations while preserving both local motifs and long-range interactions.

To address this, I developed **Bag-of-Mer (BoM) Pooling** [4], a locality-aware pooling strategy inspired by k-mer representations of biological sequences. BoM-Pooling combines local motif-level averaging with a global attention mechanism to generate representations that capture both fine-grained biochemical features and global context. This approach consistently outperforms conventional pooling methods across a wide range of biological tasks such as fluorescence prediction, enzyme activity, remote homology detection, and protein–protein interactions. The work was recognized with the **ISMB 2025 Lawson Van Toch Best Student Paper Award**, highlighting its impact and potential for widespread adoption.

In ongoing work, I am extending BFMs to long-context applications and designing representation-driven approaches to foundational bioinformatics tasks. One such direction is **multiple sequence alignment (MSA)**, a core problem underlying evolutionary analysis and protein structure prediction. Traditional MSA tools rely on heuristic substitution matrices that perform poorly on low-identity sequences. To this end, I developed a new MSA tool guided by a PLM-derived similarity measure. Our tool achieved substantially more accurate and biologically informed alignments than traditional methods, and is under review at the RECOMB 2026.

Federated Learning for Distributed Biological Data

The full potential of BFMs depends on access to diverse datasets spanning populations, environments, and organisms. However, biological datasets are frequently siloed due to privacy constraints, proprietary restrictions, or regulatory barriers. Federated learning (FL) offers a solution to this conundrum by enabling institutions to collaboratively train models without sharing raw data, but its effectiveness is limited when data silos exhibit severe **distribution shift**.

Drawing from my expertise in distributed machine learning [6, 1] and probabilistic methods [2], I developed **Probabilistic Federated Prompt Tuning (PFPT)** [7], a framework for parameter-efficient and heterogeneity-aware fine-tuning of foundation models. PFPT represents prompts as probabilistic latent variables that capture both shared global structure and silo-specific variation. This design allows different institutions to adapt a pre-trained foundation model to their local data while communicating only small, privacy-preserving parameter updates. Our result shows that PFPT reduces communication cost by up to **90%** relative to full-model fine-tuning and substantially improves robustness to distribution shift.

Although initially developed outside biological domains, PFPT provides a blueprint for federated training of BFMs across hospitals, biobanks, and biotechnology companies. By combining PFPT with learnable compression and locality-aware pooling, my research envisions a new generation of biological foundation models that can be trained collaboratively at scale, respecting privacy and data governance while enabling shared scientific discovery. These models will respect privacy, regulatory constraints, and data-governance requirements while still benefiting from the diversity of globally distributed datasets. Such systems would allow institutions to contribute to and share in powerful models without relinquishing sensitive data, ultimately fostering more robust and equitable biological discovery.

Long-Term Vision

My long-term vision is to build the next generation of BFMs that can reason over genome-scale context, integrate diverse molecular modalities, and adapt continuously as new data and tasks emerge. Current models remain limited by restricted context windows and simplified representations of biological sequences. By developing principled compression schemes and representation architectures that preserve long-range dependencies, I aim to push BFMs beyond these barriers, enabling them to model entire genomic loci, multi-domain protein architectures, and complex evolutionary patterns directly from sequence.

A second pillar of my vision is to make these models **scalable and privacy-preserving**. As biological data continue to grow in volume and diversity, they will remain distributed across hospitals, biobanks, research institutes, and industry partners. Leveraging insights from my work on probabilistic federated learning, I seek to design frameworks that allow BFMs to learn collaboratively from such distributed data without compromising privacy or institutional boundaries. This will enable broader participation in foundational model development, reduce biases arising from data silos, and support the responsible deployment of large-scale biological AI systems.

Ultimately, I envision BFMs that act not merely as predictive tools, but as scientific systems capable of generating hypotheses, uncovering disease mechanisms, guiding protein and therapeutic design, and providing mechanistic insight across scales of biological organization. My goal is to establish the computational foundations for flexible, biologically faithful, and globally accessible models that accelerate discovery in genomics, proteomics, and precision medicine.

References

[1] Minh Hoang, Nghia Hoang, Bryan Kian Hsiang Low, and Carl Kingsford. Collective model fusion for multiple black-box experts. In *International Conference on Machine Learning*, pages 2742–2750. PMLR, 2019.

[2] Minh Hoang, Nghia Hoang, Hai Pham, and David Woodruff. Revisiting the sample complexity of sparse spectrum approximation of gaussian processes. *Advances in Neural Information Processing Systems*, 33:12710–12720, 2020.

[3] Minh Hoang, Guillaume Marçais, and Carl Kingsford. Density and conservation optimization of the generalized masked-minimizer sketching scheme. *Journal of Computational Biology*, 31(1):2–20, 2024.

[4] Minh Hoang and Mona Singh. Locality-aware pooling enhances protein language model performance across varied applications. *Bioinformatics*, 41(Supplement_1):i217–i226, 2025.

[5] Minh Hoang, Hongyu Zheng, and Carl Kingsford. Differentiable learning of sequence-specific minimizer schemes with deepminimizer. *Journal of Computational Biology*, 29(12):1288–1304, 2022.

[6] Trong Nghia Hoang, Quang Minh Hoang, and Bryan Kian Hsiang Low. A distributed variational inference framework for unifying parallel sparse gaussian process regression models. In *International Conference on Machine Learning*, pages 382–391. PMLR, 2016.

[7] Pei-Yau Weng, Minh Hoang, Lam Nguyen, My T Thai, Lily Weng, and Nghia Hoang. Probabilistic federated prompt-tuning with non-iid and imbalanced data. *Advances in Neural Information Processing Systems*, 37:81933–81958, 2024.