Matt Ebrahim, Ph.D.

Senior Data Scientist — Generative AI — Drug Discovery — Biotech
m.ebrahimkhani1993@gmail.com — (631) 275-5369 — U.S. Permanent Resident
LinkedIn GitHub Google Scholar

Professional Experience

Senior Data Scientist, Formation Bio – *Jun 2025 – Present*

- Lead multiple high-impact drug repurposing initiatives, owning the full ML lifecycle from data curation and preparation through stakeholder communication and strategic decision support.
- Architected scalable graph neural network pipelines trained on large-scale biomedical knowledge graphs to estimate probability of success between drug assets and disease indications, directly informing portfolio investment decisions.
- Collaborated with data engineering and platform teams to build robust knowledge graph infrastructure and implement comprehensive data processing pipelines for GNN training at scale.
- Developed MRI imaging as a surrogate endpoint for drug efficacy studies, potentially reducing clinical trial duration. Coordinated with external vendors to acquire a dataset of tens of thousands of patients with multiple MRI series and acquisitions.
- Built scalable data curation pipelines for large-scale MRI processing on GCP, including DINOv2 embedding generation, clustering-based quality assessment, and outlier removal to prepare datasets for model training.
- Designed and trained Vision Transformer (ViT) architectures for time-to-event prediction from MRI, bridging computer vision with clinical drug development applications.
- Fine-tuned domain-specific LLMs (BioMistral, BioMegatron, SapBERT) for ontology mapping framed as named-entity recognition, improving data standardization across therapeutic areas.

AI Scientist II, 1910 Genetics – *Nov 2024 – May 2025*

- Led design and deployment of generative AI systems for de novo molecular generation, overseeing the full research lifecycle from literature review through production deployment.
- Developed novel multimodal AI architecture integrating molecular graph representations with molecular dynamics features for blood-brain barrier permeability prediction in CNS drug discovery (manuscript submitted to NeurIPS 2025).
- Delivered computationally designed molecules that progressed through in vitro and in vivo validation, demonstrating high target engagement with nanomolar inhibition coefficients.
- Managed cross-functional teams of data engineers; architected and deployed models on Azure ML, AWS Bedrock, and SageMaker.
- Served as regular participant in investor meetings, demonstrating AI capabilities and communicating strategic vision to stakeholders.

AI Scientist I, 1910 Genetics – *2023 – Nov 2024*

- Built and deployed Graph Neural Networks, chemical language models (LSTM and Transformer architectures), and Denoising Diffusion Probabilistic Models with GNNs for de novo small molecule generation.
- Integrated multi-parameter optimization techniques (Pareto front analysis, Bayesian optimization) to identify drug candidates optimizing for novelty, potency, and metabolic stability simultaneously.
- Integrated quantum mechanics-derived features into graph models for improved ADMET prediction; benchmarked against Graphormer, AttentiveFP, and QM cross-attention embeddings.
- Developed SMILES-based RNNs with reinforcement learning fine-tuning for CNS-targeted drug discovery applications.
- Curated and standardized datasets (CrossDocked, ChEMBL, MOSES) for reproducible model development and benchmarking.

Adjunct Teaching Professor, Northeastern University – *Spring 2025 – Present*

- Designed and delivered CSYE 7374: *Deep Learning and Generative AI in Healthcare*, a graduate course providing hands-on experience with state-of-the-art AI for healthcare applications.
- Developed comprehensive curriculum covering foundation models (MedGemma, BioMedCLIP), DDPMs for medical image synthesis in rare disease settings, and chemical language models for de novo drug generation.
- Created challenging programming assignments and capstone projects mirroring real-world healthcare AI problems; conducted office hours and contributed to broader university AI initiatives.
- Course site: matt-ebrahim.github.io/CSYE7374

Clinical Research Associate, Northwestern University – *2022 – 2023*

- Led two major research initiatives at the intersection of medical imaging and deep learning, leveraging Northwestern's high-performance computing cluster for scalable processing.
- Developed CycleGAN architecture to synthesize MRI flow imaging from CT angiographic data, enabling hemodynamic assessment from anatomy-only scans. Implemented medical image registration algorithms for multi-modal spatial alignment.
- Developed deep learning pipelines combined with seismocardiography as a cost-effective alternative to 4D flow MRI for cardiovascular assessment, resulting in a first-author publication and U.S. patent.
- Built end-to-end CNN-MLP systems for non-invasive estimation of aortic blood flow from wearable SCG signals.

Ph.D. Research, Stony Brook University – *2017 – 2022*

- Developed signal- and image-processing ML models for terahertz imaging and burn injury diagnostics.
- Published 8+ first-author papers and co-authored 30+ peer-reviewed publications.

Technical Skills

Languages: Python, C++, MATLAB, Bash
AI/ML Frameworks: PyTorch, TensorFlow, Scikit-learn, Optuna, PyCaret, Hugging Face
Generative Modeling: Diffusion (DDPM), Transformers (GPT, T5), VAE, GAN, RNN/LSTM
Molecular Modeling: RDKit, DeepChem, AutoDock Vina, Open Babel, ESMFold, AlphaFold
Graph ML: Graphormer, GCN, GAT, MPNN, DGL, PyG
Foundation Models: BioMistral, BioMegatron, SapBERT, MedGemma, BioMedCLIP
Cloud & MLOps: Azure ML, AWS (Bedrock, SageMaker), GCP, Snowflake, Docker, SLURM, Git, GitHub
Project Management & Collaboration: Jira, Confluence, Agile/Scrum
Domains: Drug discovery, ADMET modeling, protein-ligand design, medical imaging, knowledge graphs

Education

Ph.D., Biomedical Engineering – Stony Brook University, 2022
B.Sc., Electrical Engineering – Amirkabir University of Technology, 2016

Select Publications & Patents

- **Ebrahim M.** et al. Multimodal Graph-Attention Networks with QM-Guided Cross-Attention for ADMET Prediction. *Submitted to NeurIPS 2025.*
- **Ebrahim M.** et al. Deep Learning for Aortic Flow Estimation from SCG. *Annals of Biomedical Engineering*, 2023. Link
- **Ebrahim M.** et al. Deep Learning for Triage of *in vivo* Burn Injuries. *Biomed. Opt. Express*, 2022. Link
- **U.S. Patent** (2025): Personalized Chest Acceleration Using Deep Learning. Link
- 30+ additional co-authored publications in medical imaging, spectroscopy, and biomedical signal processing.