Gerstein Lab Experience in Leveraging AI and Biosensors for Personalized Medicine and Predictive Health Analytics

Drawing on a strong foundation of research, the Gerstein Lab has advanced the fields of personalized medicine and predictive health analytics by integrating artificial intelligence and biosensor technologies to tailor health interventions and predict disease outcomes:

We have established a strong foundation in developing practical tools for genomic analysis, highlighted by our contributions to major consortia such as PsychENCODE. We helped generate a comprehensive online resource for the functional genomics of the human brain, an initiative that has informed subsequent models and tools, including NeuroAI/omics[44]. This resource offers a detailed mapping of gene expression and regulatory networks across a large sample size, which aids in the understanding of the genomic basis of psychiatric disorders. We developed LNCTP, an innovative omics-based deep-learning approach designed to predict various psychiatric phenotypes from genotypes and detailed single-cell data. The LNCTP model utilizes a multi-level architecture incorporating a Boltzmann-machine gene expression imputation engine and hierarchical linear predictors (Fig. 2). This tool enabled us to explore the gene expression and chromatin states across a diverse cohort, including individuals diagnosed with various psychiatric disorders. The resulting insights have provided a robust foundation for our real-time analysis capabilities[45]. Moreover, we have been developing methods for genomic privacy and data anonymization, which are important given the sensitive nature of the data we handle. This work includes developing algorithms that prevent linkage attacks in genomic datasets and proposing novel data formats like the Mapped Read Format (MRF), which anonymizes sequence data while retaining useful information for analysis[46,47,48,49].

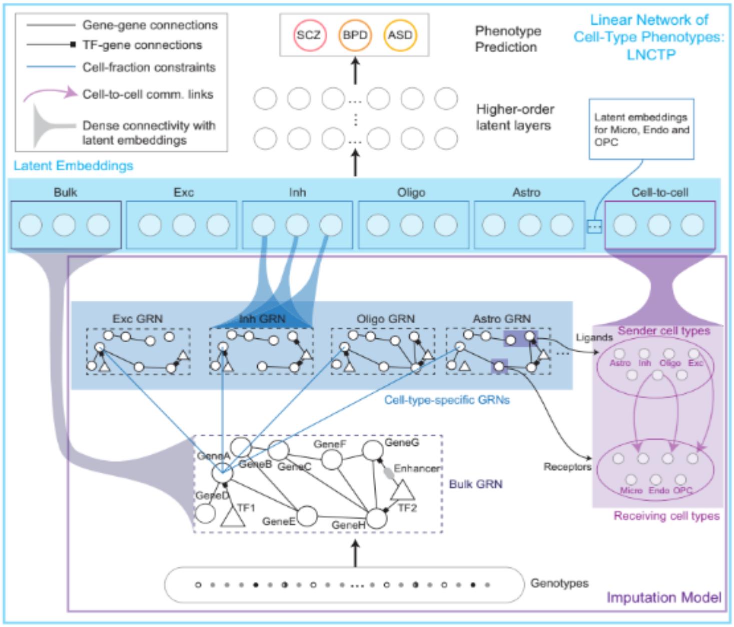

Figure 2: LNCTP Architecture. This figure presents the architecture of the LNCTP model, detailing its components and data flow. The diagram visualizes the integration of genotype data with cell-type-specific gene expression to predict psychiatric phenotypes. Key elements include the use of a conditional energy-based model for imputing gene expression and a hierarchical linear model for phenotype prediction.

We have developed various methods to analyze and integrate large-scale genomic data, including non-coding regions and their coding targets, to prioritize variants and understand their impacts on gene function and regulation[50,51,52,53,54]. Such genomic mapping efforts have informed the predictive models we are developing, enhancing accuracy and applicability. In our previous work, we successfully incorporated advanced techniques to enhance network inference capabilities in our analytical tools. We have developed various methods for processing datasets from the BICCN, demonstrating our capacity to handle and analyze genomic data from varied sources, as highlighted in our publications[55,56,57]. We are actively expanding our work to include more complex models of gene regulation and network dynamics, utilizing cutting-edge machine learning techniques to predict and simulate the effects of genetic variations on cellular and organismal phenotypes. These efforts not only improve our understanding of the human genome but also facilitate the translation of these findings into practical applications in medicine and healthcare. With many applicable tools and databases from our previous research, our work towards the goals of the NeuroAI project is realistic and likely to lead to improvements in understanding and treating brain-related issues.

Our previous work has also demonstrated advancements in the analysis and interpretation of multi-omics data, providing a solid foundation for integrating advanced deep learning architectures. In the context of enhancing the interpretability and application of machine learning models in neuroimaging and genomics, we have integrated LLMs and other advanced techniques into biomedical research. For instance, the BIOCODER project showcased the effectiveness of LLMs in managing and interpreting diverse biological data formats[58]. We developed MolLM, a pre-trained model that captures biomedical text and molecular information, enhancing performance[65]. Preliminary studies revealed the potential of LLMs and chain-of-thought reasoning to enhance complex reasoning tasks and develop autonomous agents[66]. Our Multi-disciplinary Collaboration framework significantly improved LLM reasoning in medicine[67], and ML-Bench demonstrated LLMs' ability to utilize open-source libraries[68]. Additionally, our structure-aware fine-tuning improved LLMs' capability to generate complex structured data[69], and the BioCoder benchmark illustrated our proficiency in bioinformatics coding and domain-specific challenges[70]. Finally, we fine-tuned an LLM to predict protein phase transitions, showing superior performance and interpretability, particularly for Alzheimer's disease-related proteins[71]. In the EN-TEx study, we developed a predictive multi-omics transformer model for evaluating the impact of genetic variants. The cross-tissue, cross-individual, and cross-assay aggregation strategies enhanced the detection power of allele-specific events, enabling the generation of a sizable catalog of such events that can be used to predict variant impact with high accuracy[62]. Moreover, we also have experience in developing integrated regulatory networks using high-throughput sequencing data. These networks provide a view of gene regulation by merging data from different omics layers, thus aiding our understanding of the transcriptional and post-transcriptional landscape[59]. Another area of our expertise is in the application of various sophisticated tools to map intricate relationships in biological systems. These models have proven particularly effective in analyzing microbial communities and their metabolic pathways, demonstrating our team's capability to correlate environmental factors with biological data, which can help delineate metabolic impacts on brain functions and disorders[60]. We also have successful experience in using CNNs interpret machine learning and deep learning models. For example, our DECODE framework, which outperforms state-of-the-art methods in enhancer prediction and precise boundary detection, significantly enhances the accuracy and resolution of regulatory element mapping, thus improving downstream analyses and variant enrichments[72]. ThermoNet, a 3D-convolutional neural network that accurately predicts mutation-induced changes in protein stability (ΔΔG), has demonstrated its utility in clinical and biophysical applications[73].

We have a considerable history of conducting simulation and perturbation calculations. For instance, we developed Forest Fire Clustering, a method that efficiently identifies and evaluates cell-type transitions, aiding in robust simulation and perturbation calculations in large-scale single-cell data.[74] Additionally, we developed VarSim, a comprehensive framework for simulating and validating genetic variants, which supports the simulation and evaluation of perturbations in next-generation sequencing data[75]. We also developed Paired-End Mapper, an analysis pipeline for processing genomic structural variants, featuring simulation-based error models that support the evaluation of perturbations in next-generation sequencing data[76]. In the DREAM3 Challenges, we performed computational reconstruction of *in silico* GRNs, effectively integrating heterogeneous data from deletion strains and perturbation time series to enhance network prediction accuracy[77]. Furthermore, we introduced SCAN-ATAC-Sim, an efficient and scalable method for simulating scATAC-seq experiments with known cell-type labels, enhancing the benchmarking and evaluation of scATAC-seq analysis techniques[78]. We also embedded the regulatory network into a deep-learning model, the precursor to LNCTP, to predict psychiatric phenotypes from genotype and expression. The model has improved prediction accuracy over traditional additive models[44]. It can highlight key genes and pathways associated with disorder prediction, including immunological, synaptic, and metabolic pathways, recapitulating *de novo* results from more targeted analyses.

We have also sufficient experience in developing tools that support interpretation purposes, as well as on a cloud-based platform for real-time processing ability. For instance, we developed "Gene Tracer," an innovative voice-controlled tool designed to enhance the interactive querying and visualization of genomic information. This cloud-based approach not only meets the computational demands of processing large genomic datasets but also guarantees that the system remains responsive and accessible to users from any location[61].

44. Wang, Daifeng, et al. "Comprehensive functional genomic resource and integrative model for the human brain." Science 362.6420 (2018): eaat8464.
45. Emani, Prashant S., et al. "Single-cell genomics and regulatory networks for 388 human brains." bioRxiv (2024): 2024-03.
46. Harmanci, Arif, and Mark Gerstein. "Quantification of private information leakage from phenotype-genotype data: linking attacks." Nature methods 13.3 (2016): 251-256.
47. Habegger, Lukas, et al. "RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries." Bioinformatics 27.2 (2011): 281-283.

48. Harmanci, A., and M. Gerstein. "Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions. Nat. Commun. 2018; 9 (1): 2453."
49. Gürsoy, Gamze, et al. "FANCY: fast estimation of privacy risk in functional genomics data." Bioinformatics 36.21 (2020): 5145-5150.
50. Saliba, Antoine-Emmanuel, et al. "Single-cell RNA-seq: advances and future challenges." Nucleic acids research 42.14 (2014): 8845-8860.
51. Fode, Carol, et al. "A role for neural determination genes in specifying the dorsoventral identity of telencephalic neurons." Genes & development 14.1 (2000): 67-80.
52. Rasmussen, Andreas H., Hanne B. Rasmussen, and Asli Silahtaroglu. "The DLGAP family: neuronal expression, function and role in brain disorders." Molecular brain 10 (2017): 1-13.
53. Erlander, Mark G., et al. "Two genes encode distinct glutamate decarboxylases." Neuron 7.1 (1991): 91-100.
54. Liodis, Petros, et al. "Lhx6 activity is required for the normal migration and specification of cortical interneuron subtypes." Journal of Neuroscience 27.12 (2007): 3078-3089.
55. GTEx Consortium, et al. "The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans." Science 348.6235 (2015): 648-660.
56. Khurana, Ekta, et al. "Integrative annotation of variants from 1092 humans: application to cancer genomics." Science 342.6154 (2013): 1235587.
57. Fu, Yao, et al. "FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer." Genome biology 15 (2014): 1-15.
58. Tang, Xiangru, et al. "Biocoder: A benchmark for bioinformatics code generation with contextual pragmatic knowledge." arXiv preprint arXiv:2308.16458 (2023).
59. Cheng, Chao, et al. "Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data." PLoS computational biology 7.11 (2011): e1002190.
60. Gianoulis, Tara A., et al. "Quantifying environmental adaptation of metabolic pathways in metagenomics." Proceedings of the National Academy of Sciences 106.5 (2009): 1374-1379.
61. Lou, Shaoke, et al. "Gene Tracer: a smart, interactive, voice-controlled Alexa skill For gene information retrieval and browsing, mutation annotation and network visualization." Bioinformatics 37.18 (2021): 2998-3000.
62. Rozowsky, Joel, et al. "The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models." Cell 186.7 (2023): 1493-1511.
65. Tang, Xiangru, et al. "MolLM: A Unified Language Model to Integrate Biomedical Text with 2D and 3D Molecular Representations." bioRxiv (2023): 2023-11.
66. Zhang, Zhuosheng, et al. "Igniting Language Intelligence: The Hitchhiker's Guide From Chain-of-Thought Reasoning to Language Agents." arXiv preprint arXiv:2311.11797 (2023).
67. Tang, Xiangru, et al. "Medagents: Large language models as collaborators for zero-shot medical reasoning." arXiv preprint arXiv:2311.10537 (2023).
68. Liu, Yuliang, et al. "Ml-bench: Large language models leverage open-source libraries for machine learning tasks." arXiv preprint arXiv:2311.09835 (2023).
69. Tang, Xiangru, et al. "Struc-Bench: Are Large Language Models Really Good at Generating Complex Structured Data?." arXiv preprint arXiv:2309.08963 (2023).
72. Chen, Zhanlin, et al. "DECODE: A De ep-learning Framework for Co n de nsing Enhancers and Refining Boundaries with Large-scale Functional Assays." Bioinformatics 37.Supplement_1 (2021): i280-i288.
73. Li, Bian, et al. "Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks." PLoS computational biology 16.11 (2020): e1008291.
74. Chen, Zhanlin, et al. "Forest Fire Clustering for single-cell sequencing combines iterative label propagation with parallelized Monte Carlo simulations." Nature communications 13.1 (2022): 3538.
75. Mu, John C., et al. "VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications." Bioinformatics 31.9 (2015): 1469-1471.
76. Korbel, Jan O., et al. "PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data." Genome biology 10 (2009): 1-14.
77. Yip, Kevin Y., et al. "Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data." PloS one 5.1 (2010): e8121.
78. Chen, Zhanlin, et al. "SCAN-ATAC-Sim: a scalable and efficient method for simulating single-cell ATAC-seq data from bulk-tissue experiments." Bioinformatics 37.12 (2021): 1756-1758.