The Gerstein lab has experience with building bioinformatic tools to predict phenotypes from genotypes and brain single-cell data. For example, we have developed an integrated omics-based deep-learning approach (the Linear Network of Cell-Type Phenotypes; LNCTP for short) for predicting various neurodegenerative disease phenotypes from genotypes and brain single-cell data (Emani et al, 2024). This incorporates a multi-level architecture with a Boltzmann-machine gene-expression-imputation engine coupled to a set of hierarchical linear predictors for phenotypes (with both visible and latent nodes).

LNCTP is an integrative model that inputs gene expression and prioritizes disease genes across different cell types. Here, the core handles the following tasks: (1) imputing cell-type-specific and bulk tissue gene expression from genotype; (2) predicting the risk of disorders based on input genotypes; and (3) highlighting genes and pathways contributing to particular phenotypes in their specific cell type of action. The framework includes visible options, including genotypes at scQTL and bulk eQTL sites, cell-type-specific and bulk tissue-based GRNs, cell-type fractions, cell-to-cell communication networks, gene co-expression modules, and sample covariates. It imputes cell-type-specific gene expression from genotype with high cross-validated accuracy (Fig. 1).

***A diagram of a cell type

Description automatically generatedFigure 1: 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 disease phenotypes. Key elements include the use of a conditional energy-based model for imputing gene expression and a hierarchical linear model for phenotype prediction.***

As shown in Figure 1, bulk and cell-type gene expression levels were imputed from genotype using a model incorporating GRNs and cell-to-cell networks. Cell-type-specific nodes with dense connectivity were then incorporated into a deep linear model to predict phenotypes in each sample and prioritize cell types and genes for each trait. A hierarchical architecture is used for trait-prediction, which has been demonstrated to perform comparably to or better than non-linear architectures. Moreover, the framework generates a model that is directly at multiple scales, avoiding many of the difficulties arising in the interpretation of neural networks. The linear architecture also enabled prioritization of intermediate phenotypes through gradient-based saliency and co-heritability. Figure 2 outlines key prioritized genes, cell types, and cell-to-cell interactions in disorders such as Alzheimer's disease, schizophrenia, bipolar disorder, and autism spectrum disorder.

In addition, the Gerstein lab has experience with launching tools of this nature in large consortia, such as PsychENCODE. Here, we helped build a comprehensive resource for functional genomics of the human brain, which has informed subsequent models and tools (Wang et al, 2018). It was as part of our work within PsychENCODE in which we developed LNCTP, in order to predict disease phenotypes from genotypes and single-cell data (Emani et al, 2024).

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 more generally. Our previous work incorporated network inference, and we have developed various methods for processing datasets from large consortia, demonstrating our capacity to handle and analyze genomic data from varied sources, as those highlighted in our publications (GTEx Consortium, 2015; Khurana et al, 2013). We have previous experience in linking structural and functional data with diverse disease traits (Wang et al, 2018; Emani et al, 2024). Our previous work has demonstrated significant advancements in the analysis and interpretation of multi-omics data (Emani et al, 2024). In addition, we have experience with conducting simulation and perturbation-based calculations using this tool, as demonstrated by our work published earlier this year (Emani et al, 2024).

***A diagram of a cell line

Description automatically generated***

***Figure 2: The linear activations within LNCTP, showing how genes and pathways are prioritized and their contributions to predicting neurodegenerative and psychiatric disorders. The visualization indicates the system’s ability to identify key genetic interactions and their impact on various phenotypes.***

Finally, the Gerstein lab has experience in developing tools that support interpretation purposes, as well as on a cloud-based platform for real-time processing ability (Clarke et al, 2016).

**References**

Clarke D, et al. “Identifying allosteric hotspots with dynamics: application to inter-and intra-species conservation”. Structure. 2016 May 3;24(5):826-37.

Emani PS, et al. “Single-cell genomics and regulatory networks for 388 human brains”. Science. 2024 May 24;384(6698):eadi5199.

GTEx Consortium, et al. "The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans." Science 348.6235 (2015): 648-660.

Khurana, Ekta, et al. "Integrative annotation of variants from 1092 humans: application to cancer genomics." Science 342.6154 (2013): 1235587.

Wang, D, et al. “Comprehensive functional genomic resource and integrative model for the human brain”. Science 362.6420 (2018): eaat8464.