Gerstein lab experience with construction of cell-type specific gene regulatory networks (GRNs): We have previously utilized deep neural networks (e.g., DSPN[1], LNCTP[2]) to accurately impute cell-type-specific expression and phenotype from genotype in the context of neuropsychiatric disorders. This effort identified over 250 risk genes and potential drug targets for brain-related disorders, along with associated cell types. Specifically, our LNCTP model has already achieved useful performance. Currently, our model is partially linear, however, it successfully imputes gene expression—specifically, cell- type-specific gene expression from genotype—with high cross-validated accuracy: the mean correlation between the imputed and experimentally observed expression profiles is 69% across major cell types and approximately ~ 78% in excitatory and inhibitory neurons (**Fig. 1A**). This corresponds to explaining 38% of the variance in cell-type gene expression (i.e., estimating the heritability of cell-type gene expression, h^2), compared to a 34% baseline achieved by combining prior methods for bulk imputation and cell-type deconvolution (**Fig. 1B**). LNCTP was also able to sensibly prioritize disease genes across cell types. **Fig. 1C** provides an overview of key prioritized genes, cell types, and cell- to-cell interactions in various disorders. We identified 64, 51, and 108 gene/cell-type pairs for schizophrenia (SCZ), bipolar disorder (BPD), and autism spectrum disorder (ASD), respectively.

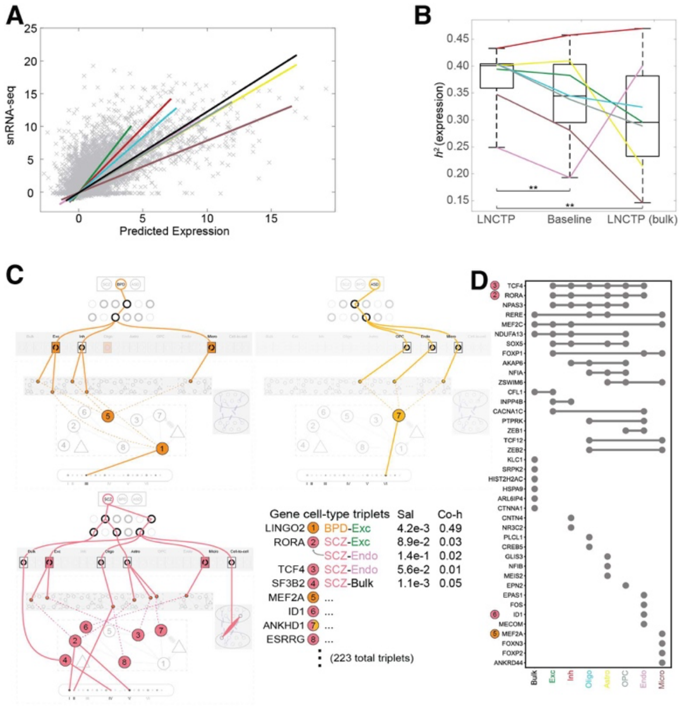

Figure 1. Imputing gene expression and prioritizing disease genes. (A) Imputed expression values from LNCTP compared with observed values (B) Comparison of explained variance in gene expression from the LNCTP model with a baseline model. (C) Schematic for LNCTP interpretation, illustrating relationships between prioritized intermediate phenotypes. (D) UpSet plot for SCZ showing overlap between genes with the highest saliency per cell type (colored circles).

Gerstein lab experience with prediction of ligand-protein binding affinities by meta-modeling: We showed that integrating outputs from multiple complementary models outperforms traditional methods) in both accuracy and robustness, enabling fast and more reliable predictions. Our approach[3] combines molecular docking scores, machine learning-based affinity predictors, and physics-informed calculations, leveraging each method's strengths while mitigating their limitations.

Gerstein lab experience with studying causative molecular mechanisms underlying diseases using attention mechanism of large language models (LLMs): In our publication[4] we demonstrated that the attention mechanism of the LLM effectively captures biologically meaningful signals related to the underlying mechanism of the Alzheimer's disease by highlighting regions associated with protein aggregation, implying for potential drug targets.

References:

1. D. Wang et al., Comprehensive functional genomic resource and integrative model for the human brain. *Science*, 362(6420), 2018.

2. P.S. Emani et at., Encode Consortium Psych. Single-cell genomics and regulatory networks for 388 human brains. *Science*, 384(6698), 2024.

3. Ho-Joon Lee, Prashant S Emani, and Mark B Gerstein. Improved prediction of ligand–protein binding affinities by meta-modeling. *Journal of Chemical Information and Modeling*, 64(23):8684–8704, 2024.

4. M. Frank, P. Ni, M. Jensen, and M. B. Gerstein. Leveraging a large language model to predict pro- tein phase transition: A physical, multiscale, and interpretable approach. *Proc Natl Acad Sci U S A*, 121(33), 2024.