**Gerstein lab experience in building a machine learning model for neurogenomics (LNCTP)**

Our LNCTP model has two key components: (i) **imputing gene expression levels** $\left(x\right)$ from genotype $\left(z\right)$ using a conditional energy-based model that integrates bulk and single-cell eQTL, GRNs, and cell-cell communication networks (CCC); and (ii) linking cell type-specific gene expression to clinical phenotype $\left(y\right)$ through a deep linear model with hidden layers $\left(h\right)$, thereby enabling **phenotype prediction** and **prioritization** of genetic risk factors as well as cellular and molecular pathways that underly phenotype.

* *Step 1:* Gene expression $x\_{i}$ of individual $i$ can be predicted directly from genotype $z\_{i}$ by decomposing the effects into four components: eQTL ($β\_{1,\cdots ,G^{'}}$), GRN in bulk tissue ($J\_{0}$), GRN for each cell type ($J\_{1,\cdots ,C}$), and CCC ($J^{C2C}$). We then model the conditional probability $p\_{GMRF}\left(z\_{i}\right)$ using a Gaussian Markov Random Field (GMRF) as: $p\_{GMRF}\left(z\_{i}\right)∝exp\left(-E\_{GMRF}\left(z\_{i}\right)\right)$

where the energy function $E\_{GMRF}$ is defined as ($λ$, hyperparameter; $f$, estimated cell fractions in bulk):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| $$E\_{GMRF}\left(z\_{i}\right)=$$ | $$xi\_{i0}^{T}Jx\_{i0}$$ | $$+ \sum\_{g}^{}x\_{i0}^{T}b\left(z,β\_{g}\right)$$ | $$+ \sum\_{c}^{}\left(x\_{ic}^{T}Jx\_{ic}+x\_{ic}^{T}b\_{c}+x\_{ic}^{T}J^{C2C}x\_{i,L,R,c}\right)$$ |  |
|  | **bulk GRN** | **bulk eQTL** | **cell type-specific eQTL, GRN, self-communication** |  |
|  | $$+\sum\_{c1,c2}^{}J\_{c1,c2}^{C2C}x\_{i,L,c1}^{C2C}x\_{i,R,c2}^{C2C}$$ | $$+ λ\sum\_{g}^{}\left(x\_{iog}-f\left(z\right)^{T}x\_{i,1,…C,g}\right)^{2}$$ | (1) |
|  | **CCC** | **Regularization** |  |

* *Step 2:* We model the hidden layers $\left(h\_{i}\right)$ and the phenotype ($y\_{i}$) of individual $i$ from gene expression $x\_{i}$ using an $L$ layer stochastic deep neural network (DNN), such that the conditional probability $p\_{DNN}\left(x\_{i}\right)$ is:

$p\_{DNN}\left(x\_{i}\right)=p\_{y}\left(h\_{iL},W\_{L+1}\right)\prod\_{l=2,\cdots ,L}^{}p\_{h}\left(h\_{iL-1},W\_{l}\right)p\_{h}\left(x\_{i},W\_{1}\right)$ (2)

By connecting these two steps, our LNCTP model (Emani et al, 2024) computes:

$p\_{LNCTP}\left(z\_{i}\right)$=$p\_{GMRF}\left(z\_{i}\right)p\_{DNN}\left(x\_{i}\right)$

In our recent work, we trained LNCTP using 388 brain samples, demonstrating its ability to link genetic variation to cellular and molecular mechanisms across brain disorders, including ASD (Emani et al, 2024).

**Gerstein lab experience in consortia work and disseminating software tools**

The Gerstein lab has a longstanding commitment to open science through the development, dissemination, and sharing of genomic data and computational tools. Dr. Gerstein has championed collaborative research through his leadership roles in ENCODE (Rozowsky et al, 2023), GTEx, IGVF, SCORCH (Ament et al, 2024), and PEC (Emani et al, 2024, Wang et al, 2018), consistently releasing widely adopted ML tools, data resources, and analysis pipelines. We have routinely deposited multiomic and high-throughput screening datasets in public repositories, develop new repositories for integrative resources (e.g., brainSCOPE), and maintain open-source tools on GitHub with version control and comprehensive user documentation. Collectively, we have developed a comprehensive suite of software packages for multmodal data analysis and brain disorder research. Many of these tools have been adopted and integrated into the standard workflows of major consortia (e.g., ENCODE, PEC, SCORCH), thus underscoring their robustness and utility.

**REFERENCES**

Ament, S. A. et al. Mol Psychiatry, (2024). PMC11609103

Emani, P. S. et al. Science, (2024). PMC11365579

Rozowsky, J. et al. Cell, (2023). PMC10074325

Wang, D. et al. Science, (2018). PMC6413328