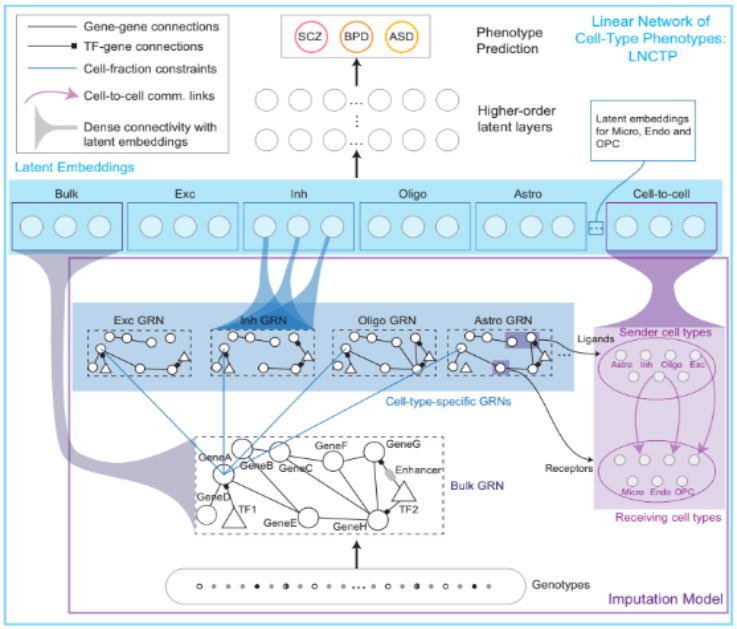
Tool # 1 - "Integrative tool to combine imaging and genomics to diagnose brain disease"

Previous work and preliminary results. 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[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. 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]. We have also 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.

Integrative analysis framework. The core of the module will be based on the deep-learning part of our most recent work, LNCTP. LNCTP is an integrative model that inputs gene expression and also 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 will include 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 will impute cell-type-specific gene expression from genotype with high cross-validated accuracy (Fig. 2).

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.As shown in Figure 2, bulk and cell-type gene expression levels were imputed from genotype using a conditional energy-based 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 linear architecture will be used for the trait-prediction portion of LNCTP, which has been demonstrated to perform comparably to or better than non-linear architectures. Moreover, the framework generates a model that is directly interpretable at multiple scales, avoiding many of the difficulties arising in the interpretation of deep neural networks, while maintaining a hierarchical structure. The linear architecture also enabled prioritization of intermediate phenotypes through gradient-based saliency and co-heritability.



Graphical-LASSO approach. We will enhance the network inference for the integrative model, using a graphical-LASSO training approach[64] instead of the maximum-likelihood approach used in the cornerstone paper[44]. The graphical-LASSO objective is efficient to optimize and flexibly allows multiple networks to be used in the expression imputation component of the LNCTP framework. Additionally, it permits each network to serve as a soft constraint when fine-tuning the model. Thus, besides removing edges from the prior networks provided, novel edges may be introduced, altering the sparsity structure and permitting the discovery of novel cell-type gene–gene interactions. We will provide a Bayesian optimization search method for setting optimal parameters.

Incorporate an imaging layer. The integration of imaging data into the LNCTP framework aims to bridge the gap between molecular and cellular processes and functional brain organization. By incorporating imaging modalities such as functional magnetic resonance imaging (MRI), we seek to enhance our understanding of spatial hierarchies and connectivity patterns within the brain. This integration will help predict psychiatric and brain-related phenotypes by combining genotypic, omic, and imaging data, potentially offering deeper insights into the intricate relationships among brain structure, function, and genetic factors. After appropriate covariate and batch correction, the resulting residual matrices can be fed as input to a deep learning layer. The subsequent integration with genotypes and omics data can be carried out in the following manner: (1) First, we will build purely deep learning models (such as feed-forward neural networks) that predict FC matrices conditional on matching genotypes. This analysis quantifies the heritable component of the functional connectivity signatures. (2) Second, we can integrate the omics-based imputation model from LNCTP on top of the genotypes to predict the FC matrices. This aims to determine the improvement in the heritability quantification of FC matrices by including inferred gene regulatory mechanisms and interactions. (3) Third, we will build a model to predict psychiatric and other brain-related phenotypes by incorporating FC matrices alongside the current LNCTP inputs. The result would be a model that quantifies the propensity for target phenotypes using both functional connectivity signatures (which correlate with certain conditions) and genetic/omic signatures.

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.

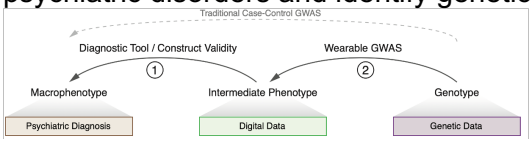
54. Liodis, Petros, et al. "Lhx6 activity is required for the normal migration and specification of cortical interneuron subtypes." Journal of Neuroscience 27.12 (2007): 3078-3089.

64. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso." Biostatistics 9.3 (2008): 432-441.

Tool #2 - "A tool to integrate wearable data and genomics to help predict disease risk"

We aim to construct variant impact models that incorporate wearable sensor data to ascertain the impact of variants on specific genes. We will build tools to leverage wearable and biosensor data to generate risk scores, as well as to identify and prioritize genetic loci of interest.

Previous Work: We have previous experience in analyzing sensor data in human health contexts. We leveraged the time-series nature of biosensor data to assess the impact of an intervention on individual health (Liu et al, 2021). We developed a tool that implements a Bayesian structural time-series framework, to forecast the impact that an intervention (e.g., drug treatment) may have on a time-series process (e.g., detection of glucose level change). This work exemplifies how analyzing time-series based biosensor data can aid in assessing personal health. In more recent work, we leverage AI and data from wearable devices to characterize psychiatric disorders and identify genetic associations (Liu et al, 2024; and see Figure below).



Proposed Work: Processing wearable biosensor data and generating biologically relevant features: The noisy nature of wearable time-series data requires rigorous preprocessing and feature engineering. Given our experience in time series signal processing, we aim to construct a tool to extract clinically relevant information from raw wearable data. We propose to combine different data modalities collected from wearable devices, as well as two distinct strategies for feature engineering: generating static and dynamic features. Static features are time-invariant and are created through summarization techniques, resulting in straightforward and efficient features commonly used in downstream modeling. Dynamic features retain the time-varying nature of the original digital signatures, preserving sequential and temporal patterns.
Characterizing brain diseases and disorders using wearable biosensors and AI: We will adopt a dual-architecture approach that employs XGBoost for static features and Xception for dynamic features. For dynamic features, we propose another module of our tool, which uses deep neural networks as the architecture for modeling. The neural network, which takes multichannel time-series data as input, will consist of a time-series encoder, pretraining decoder, disease classifier, and cognitive score predictor. We use the InceptionTime and XceptionTime architectures as integral parts of this tool.
Leveraging wearable-derived features to identify associations with genetic variants. Wearable-derived features may be used as GWAS phenotypes. As another component of our integrated tool, we plan to include a framework that performs a battery of multivariate GWAS leveraging sensor data to improve identification of disease-related variants. This framework within our tool will treat clusters of correlated features as multivariate digital phenotypes and implement a GWAS model use formula: Multivariate Digital Phenotype ~ Covariates + Disease + Genotype + Genotype:Disease, where “Genotype” corresponds to the genotype group of an individual at a particular genetic variant, “Disease” corresponds to the status of the individual (0 = control individual; 1 = individual with disease), and the interaction term Genotype:Disease allows us to identify a genetic effect on the multivariate digital phenotype that differs between cases and controls. For a given genome-wide significant locus, we will consider each feature in the significantly associated cluster and compare the feature value’s distributions among the three genotype groups. This approach will allow us to identify which specific wearable features are associated with the variant. Another way to harness the potential of wearable measurements as digital phenotypes for neuropsychiatric disorders is by using the wearable combination scores generated by our modeling framework. These scores combine wearable-derived features into a single continuous variable that summarizes an individual’s likelihood of having a particular disease.

Liu, J. et al. Bayesian structural time series for biomedical sensor data: A flexible modeling framework for evaluating interventions. PLoS Comput. Biol. 17, e1009303 (2021).

Liu, J. et al. Digital phenotyping from wearables using AI characterizes psychiatric disorders and identifies genetic associations. (in press at Cell; 2024).

Tool #3 - "Using LLMs with protein structure to find aberrant proteins "

Understanding gene regulatory networks (GRNs) helps delineate the complex genetic and epigenetic orchestration, offering insights into how specific genes and pathways contribute to diseases. This knowledge is crucial for identifying potential biomarkers for early detection and

monitoring disease stages. To reveal the interactions within affected cell-specific GRNs, we propose our **scBayesDiffNet module** (which applies **LatentDAG** followed by **DiffusionNet)** and LLM (Fig. 3)**:**

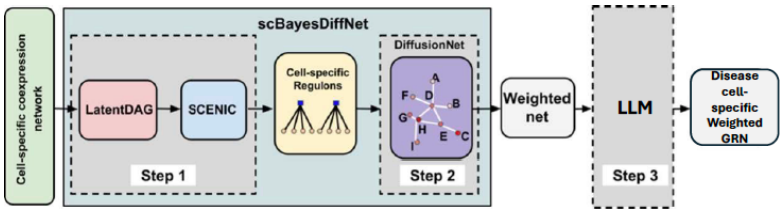


Fig. 3 Schematic of scBayesDiffNet and LLM for constructing disease-specific GRNs.

Step 1: Refining cell-specific co-expression networks using the **LatentDAG** Bayesian approach

1.1. We will build cell-specific co-expression networks from scRNA-seq data and simplify them with our LatentDAG Bayesian algorithm, which uses a directed acyclic graph (DAG) structure to capture conditional independence among variables{PMID: 26554085}. LatentDAG refines gene activity relationships in co-expression and ChIP-seq networks, creating clearer clusters and boundaries between modules, and enhancing connections in transcriptional and RNA-binding protein networks. The refined co-expression network is then input to SCENIC{PMID: 28991892} to identify regulons, with LatentDAG pre-filtering strengthening regulon motifs and reducing false positives.

Step 2: Extracting cell-specific weighted GRNs from regulons using **DiffusionNet**

Given the cell-type-specific regulon structure, we will apply a network diffusion method to better relate a set of input genes to their upstream regulators{PMID: 25078397,PMID: 15087500}. This approach allows us to integrate larger networks involving multiple transcription factors (TFs), surpassing simple combined regulons. The resulting GRNs will have weighted edges between TFs and their target genes, suggesting high-confidence interactions. This method identifies key regulators for a given target gene by providing the aggregate regulation score of each TF for that target.

Step 3: Prioritizing and weighting key nodes and edges within the GRNs related to diseases using **LLM**

We will highlight GRN nodes and edges based on associations with protein aggregation-related diseases such as Alzheimer's and cardiovascular, and cancer. First, we weight nodes by their differential expression in diseased vs. control samples and adjusting for fold changes. Then, additional weights will be added to emphasize pathways linked to protein aggregation, a key mechanism in diseases like Alzheimer's and cardiovascular diseases. Using fine-tuned large-language models (LLMs), we will predict protein aggregation likelihood and trace the TF-target gene subnetworks to update their weights. We note that LLMs are especially useful for their sensitivity to point mutations affecting aggregation as diseases progress.

Tool #4 - "Using LLMs or ML to highlight key variants for precision medicine"

Previous Work. We have made methodological contributions to analyzing and integrating large-scale genomic data, including those targeting non-coding regions and their coding targets, to prioritize variants and understand their impacts on gene function and regulation 173–177. Transformer models for predicting allele-specific behavior. We recently developed the EN-TEx resource, comprising >1,600 multi-tissue epigenetic assays mapped to personal genomes of four individuals, and analyzed the effects of non-coding variants on regulation178. As part of this work, we developed a large language model (LLM) to predict variant effects in allele-specific (AS) behavior. Traditionally, AS is measured by mapping datasets to diploid personal genomes and calculating read depth changes between haplotypes at heterozygous SNVs179. We trained a transformer model incorporating DNABERT180 to predict heterozygous SNVs that exhibit AS activity based on local sequence contexts. Attention layers within the model captured complex sequence interactions. Our model outperformed the prediction accuracy of several baselines, including for transcription factor (TF, e.g., CTCF) and histone modification activity (e.g., H3K4me3). Attention scores highlighted genomic sequences important for prediction, recapitulating known TF binding motifs and revealing potential new motifs. When combined with tissue-specific epigenetic signals, the scores accurately predicted differential variant effects across tissues. Thus, transformer models can learn dependencies between genomic sequences without prior knowledge to produce novel insights into the mechanisms underlying variant effects. Similarly, our work on the DECODE framework leveraged sophisticated deep neural networks to refine genomic annotations for precise enhancer prediction and localization181. Additionally, we developed advanced frameworks to integrate text with multimodal molecular representations, i.e., 1D sequences, 2D interactions, and 3D structures182. We have also fine-tuned the ESMFold LLM for predicting protein phases, demonstrating its superior performance compared to classical benchmarks such as random forest models183.

Proposed Work. Transformer model approach for prioritizing variants. We will extend our transformer model approach for predicting allele-specific behavior and apply it to all variants to filter those likely to exhibit functional impacts. We will also compare the effect of using the reference genome sequence to the personal genome sequence in the window surrounding the heterozygous SNV (hetSNV). We will investigate which DNA sequence features are important for the transformer model performance. By observing the attention score patterns around AS hetSNVs, we will determine whether the transformer model focuses on distinct sequence features in the local neighborhood of hetSNVs. For example, in predicting AS CTCF binding, the transformer model identifies not only CTCF motifs but also other motifs corresponding to associated TF cofactors that bind near CTCF, which are used as additional features. We will apply the transformer model approach to predict SNVs that exhibit allele-specific expression within genes and allele-specific binding for regulatory regions near genes, aiming to identify hetSNVs likely to impact the function of those target genes. Additionally, we will extend the transformer model approach to predict other types of genomic function, such as tissue specificity, using the local sequence context around a targeted variant. This approach can be further refined to prioritize variants that significantly impact the expression of specific genes expressed in subsets of tissues or under certain phenotypic conditions.

173.Saliba, A.-E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. 42, 8845–8860 (2014).
174.Fode, C. et al. A role for neural determination genes in specifying the dorsoventral identity of telencephalic neurons. Genes Dev. 14, 67–80 (2000).

175.Rasmussen, A. H., Rasmussen, H. B. & Silahtaroglu, A. The DLGAP family: neuronal expression, function and role in brain disorders. Mol. Brain 10, 43 (2017).
176.Erlander, M. G., Tillakaratne, N. J., Feldblum, S., Patel, N. & Tobin, A. J. Two genes encode distinct glutamate decarboxylases. Neuron 7, 91–100 (1991).
177.Liodis, P. et al. Lhx6 activity is required for the normal migration and specification of cortical interneuron subtypes. J. Neurosci. 27, 3078–3089 (2007).
178.Rozowsky, J. et al. The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models. Cell 186, 1493–1511.e40 (2023).
179.Rozowsky, J. et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol. Syst. Biol. 7, 522 (2011).
180.Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics 37, 2112–2120 (2021).
181.Chen, Z. et al. DECODE: a Deep-learning framework for Condensing enhancers and refining boundaries with large-scale functional assays. Bioinformatics 37, i280–i288 (2021).
182.Tang, X., Tran, A., Tan, J. & Gerstein, M. B. MolLM: A unified language model for integrating biomedical text with 2D and 3D molecular representations. bioRxiv (2023) doi:10.1101/2023.11.25.568656.
183.Frank, M., Ni, P., Jensen, M. & Gerstein, M. B. Leveraging a large language model to predict protein phase transition: a physical, multiscale and interpretable approach. bioRxiv (2023) doi:10.1101/2023.11.21.568125.