**Research Strategy**

**A. Significance**. Understanding the complexities of brain function requires advanced informatics tools that can analyze, visualize, and integrate different data at multiple scales. While there have been considerate efforts in generating genomic and imaging datasets, researchers lack a comprehensive and accessible framework to integrate these diverse modalities effectively. Current tools, though sometimes powerful in certain specific applications, remain fragmented, non-intuitive, or resource-intensive, limiting their widespread adoption. **In response, our proposal aims to develop advanced, user-friendly tools that incorporate genetics, imaging data, and artificial intelligence to provide a comprehensive multi-scale analytical framework.** These tools will support the analysis and visualization of integrated datasets from various sources, enabling the interpretation of brain functions and interactions without the need for local deployment. The development of these tools is crucial given the absence of a framework enabling researchers to easily handle the complexity and integrate multi-omics and imaging data. While some specific-purpose tools exist, they are not well-integrated or user-friendly, which hinders their wider applications. Thus, the proposed integrated and accessible framework may serve as an innovative new means for the research community or interested individuals to conduct comprehensive and insightful analyses.

Challenge 1: Multi-Omics Data Integration. Integrating multi-omics data is a complex task due to the significantly different data types involved, typically requiring customized strategies for integration. For example, combining imaging data with single-cell gene expression data requires specialized methods that can accurately link spatial structures with molecular details, indicating that strategies must be designed to deal with unique characteristics, including size, distribution, and variability.

Challenge 2: Limitations of Existing Analytical Models. Existing analytical models often struggle to differentiate complex biological data with the depth, clarity, and resolution needed for detailed analysis. Additionally, the computational resources required to process and analyze large-scale omics datasets are usually considerable, highlighting the need for a professional online platform optimized for performance and accessibility.

Challenge 3: Interpretability and User Adoption. The complexity and "black-box" nature of advanced analytical models, such as deep learning approaches, can obscure understanding of input-to-output transformations. Additionally, these sophisticated tools require effective training and support for users to fully utilize these capabilities, emphasizing the importance of simplicity in tool design for data sharing and collaboration.

In summary, addressing these challenges is essential for utilizing the potential of multi-omics data to uncover the complexities of brain function and disease. The development of dependable and trustworthy computational tools that can efficiently process and analyze complex datasets, offer detailed insights into cellular and molecular mechanisms, and provide clear, interpretable results is highly important. Such advancements will enable a deeper understanding of the complex relationships among genomics, transcriptomics, and brain structure, thereby enhancing understanding in neuroscience.

**B. Innovation**. We propose to build NeuroAI, a deep-learning embedded analysis tool, to integrate multi-omics and imaging data with cutting-edge artificial intelligence (AI) methodologies. This multi-layered tool transforms complex, heterogeneous data into interactive insights, enhancing both accessibility and interpretability. Currently, no tool exists that allows researchers to easily grapple with the integration of multi-scale biological and imaging data. Our proposed tool is unique in its ability to integrate various data types, providing a user-friendly platform that simplifies the complex process of data analysis. This distinctive capability positions NeuroAI at the forefront of innovation, filling in a gap in the field and offering indispensable functionalities that were previously unavailable.

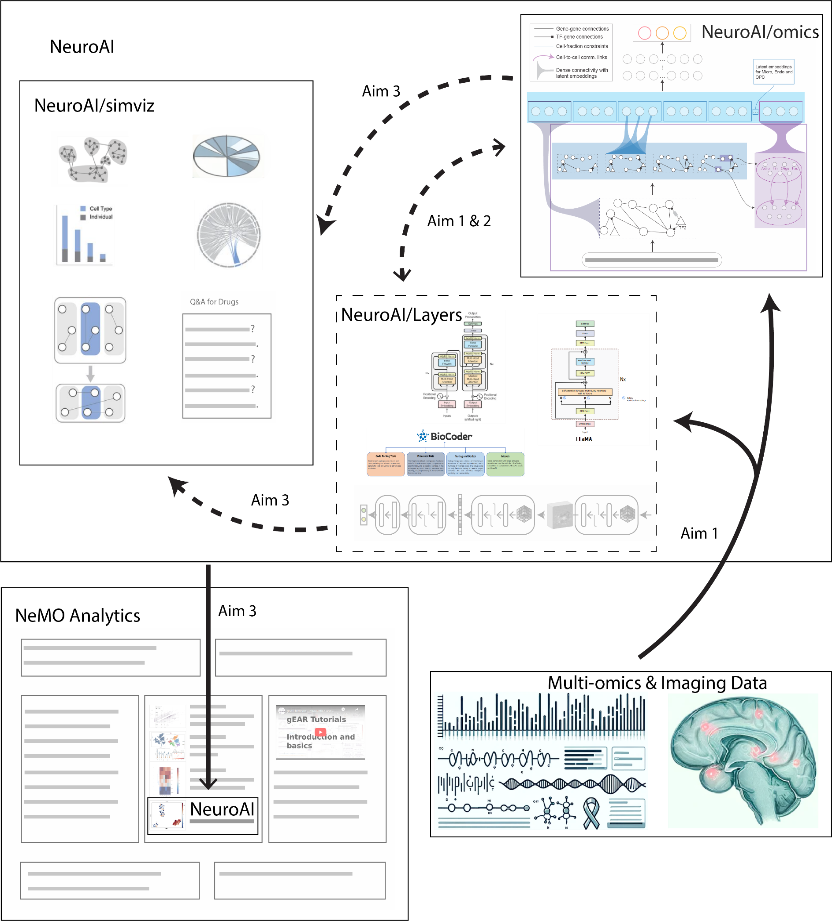
B.1 Comprehensive Integration via NeuroAI/omics. The purpose of NeuroAI/omics is to transform advanced omics-based deep learning methods into a user-friendly online interface while integrating graphical LASSO for network inference. Unlike conventional methods that handle data in isolation, NeuroAI/omics will utilize a Boltzmann-machine architecture coupled with hierarchical linear predictors. This setup will not only enhance the prediction accuracy of psychiatric phenotypes, but also enable users to input their own datasets, thus supporting a flexible, personalized analysis environment.

B.2 Advanced Machine Learning for Enhanced Interpretability with NeuroAI/layers. NeuroAI/layers will introduce a modular collection of advanced machine learning models, including large language models (LLMs), convolutional neural networks (CNNs), and more. NeuroAI/layers alleviates the interpretability challenges common in many neural network applications by enabling energy-based optimization to produce phenotype predictors and allowing detailed exploration of the connections between genomic and imaging data.

B.3 Simulation and Visualization Innovations in NeuroAI/simviz. NeuroAI/simviz will offer a robust simulation platform for "what-if'' scenario analysis. This platform will implement an intuitive interface and leverage the Neuroscience Multi-omic (NeMO) archive’s cloud-native technologies and real-time processing capabilities. NeuroAI/simviz will also provide a means to observe changes in gene activity and how genes control each other, aiding in the understanding of genetic factors’ influence on various physical traits and conditions.

**C. Approach**.

***Overview.*** Our approach leverages advanced deep learning techniques to create NeuroAI, an integrative and user-friendly online platform for multi-omics and imaging data analysis. NeuroAI/omics will transform existing omics-based deep learning models into a flexible interface, incorporating graphical LASSO for enhanced network inference and allowing users to input their own datasets for personalized analysis. NeuroAI/layers will utilize a modular collection of advanced machine learning models for visual and omics data, such as LLMs and CNNs, to improve interpretability and explore the connections between genomic and imaging data. Finally, NeuroAI/simviz will provide a robust simulation platform for scenario analysis, such as drug effects on gene expression, using cloud-native technologies for real-time processing and integration with the NeMO archive. This comprehensive approach aims to enhance our understanding of brain phenotypes and diseases through advanced, accessible, and integrative computational tools. See Figure 1 for a summary of the aims.

*Figure 1: Comprehensive Workflow of the NeuroAI Project. This diagram illustrates the progression from data input through various analytical modules, including NeuroAI/omics, NeuroAI/layers, and NeuroAI/simviz. Each component demonstrates a key stage in the data processing pipeline, indicating the tool's multidimensional approach to understanding and predictions through advanced computational models and interactive tools.*

***Team.*** PI Gerstein, Co-I Holmes, and Co-I White bring extensive expertise in method development, network-based genomic analyses, and cloud-based software. Their collaborative work has produced practical tools for genomic analysis and contributed to large consortia such as PsychENCODE, yielding comprehensive resources for functional genomics of the human brain. They have processed millions of single-cell nuclei from the prefrontal cortex, studied gene regulation, and developed methods for genomic privacy. Additionally, they have launched multiple user-friendly bioinformatics platforms—such as MolMovDB.org, Pseudogene.org, exRNA.org, FunSeq, and AlleleDB—that exemplify their track record in building and maintaining cutting-edge informatics solutions.

***Datasets and data processing.*** We will use pipelines to uniformly process and integrate multi-omics data. We will use standardized preprocessing protocols for various genomic datasets, including single-cell RNA sequencing (scRNA-seq), bulk RNA sequencing, and whole-genome sequencing (WGS). Established pipelines will ensure precise, reliable data processing. To harmonize multimodal datasets, we will develop robust preprocessing pipelines to manage batch effects and sparsity issues, using tools like MetaCell-2 for scRNA-seq data and advanced scaling techniques for aligning genomic, transcriptomic, and imaging data.

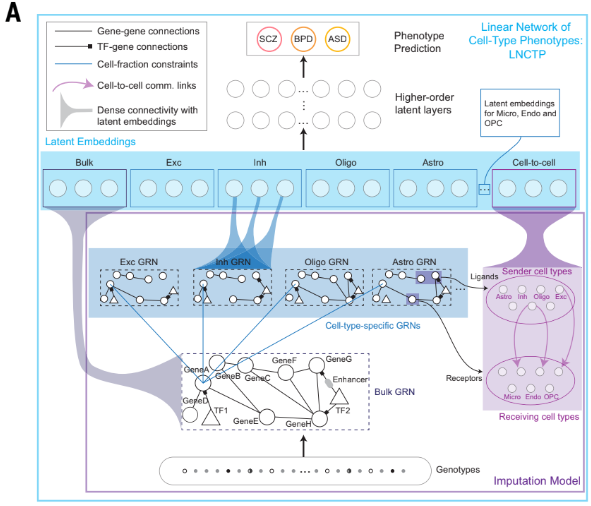
**C.1 Aim 1. NeuroAI/omics: Transformation of the Existing Software into a Comprehensive Online Tool.**

Rationale. Our first aim is to extend the accessibility of our sophisticated integrative analysis framework, designed to predict various brain phenotypes from genotypes and brain single-cell data, by creating a user-friendly online tool called NeuroAI/omics. This tool will incorporate graphical LASSO analysis for enhanced network inference and will be pre-trained on various established datasets. In addition, NeuroAI/omics will enable users to customize and fine-tune the model with their own datasets.

Aim 1a. Develop and release NeuroAI/omics as a user-friendly online tool. This tool will support real-time analysis and feature advanced data visualization capabilities to enhance the interpretation of complex omics data. Building upon our previous work with the Linear Network of Cell-Type Phenotypes (LNCTP) model, NeuroAI/omics will further integrate graphical LASSO analysis for improved network inference, in addition to being pre-trained on diverse datasets.

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, including NeuroAI/omics44. 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 capabilities45. 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 analysis46,47,48,49.

Integrative analysis framework. The core of the NeuroAI/omics module will be built 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 processes key genomic features including cell-type-specific GRNs, cell-type fractions, and gene co-expression modules to accurately impute cell-type-specific expression from genotype data, as shown in Figure. 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.*

Datasets for training. We will train LNCTP as a conditional energy-based model representing the joint distribution of the variables conditioned on genotype, with additional latent layers. To provide a robust core for analysis and fine-tuning, we will utilize as many related datasets as possible for training the model. Table 1 lists candidate datasets that we have examined and believe could be suitable for further training.

*Table 1. The data we will use include scRNA-seq data integrated from multiple databases to capture cell-type-specific regulatory mechanisms. The sample counts are categorized by data modality, source status, and disease, ensuring a comprehensive and diverse dataset for our training purposes.*A table of numbers and letters

Description automatically generated

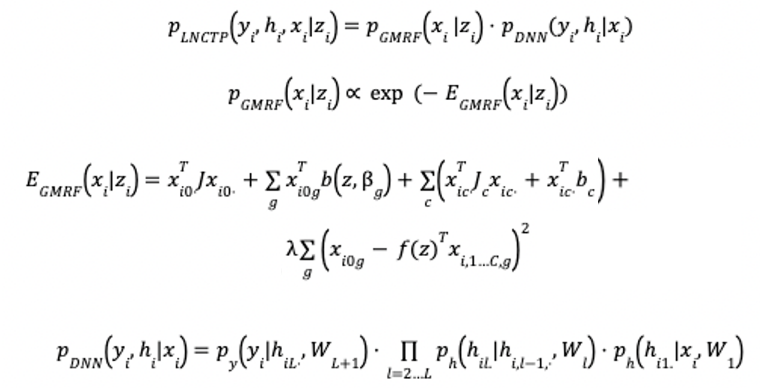
Pipeline for processing data. For pre-training or fine-tuning, the data needs to be pre-processed to handle various potential problems, such as batch effects, sparsity issues, and biases. We will use the processed scRNA-seq cohorts to quantify priors on the expression values for each gene in each cell type to constrain the cell-type expression imputation process. Additionally: (1) For each cell type in each individual in every cohort, we will use the program MetaCell-2 to generate “metacells," which aggregate counts across several cells in close proximity to each other; (2) We will convert the gene expression values in each individual into CPM units, averaged across all meta-cells in that individual and cell type, and subsequently converted into z-scores; (3) We will employ the z-score distribution for each cell type across all individuals and study cohorts as a prior in the deep learning framework. We will calculate the mean and standard deviation (SD) in the z-scores and incorporate them into the corresponding Gaussian Markov random field (GMRF) terms. For each sample, we filter genes (≥50 UMIs, ≥3 cells), process cell types with Metacell-2 or compute medians if a cell type has fewer than 30 cells or 30,000 UMIs, and then label and merge the resulting metacells into one consolidated dataframe; next, we convert their raw UMI counts to CPM, perform z-scoring by cell type (calculating mean expression, normalizing, merging z-scores, and determining overall means and SDs), and thus obtain normalized metacell gene expression values for downstream analyses.

Transformation of LNCTP into an online-ready version for integration into NeuroAI/omics. To enhance accessibility, usability, and functionality, it is necessary to transform the LNCTP model into an online-ready tool, we will implement several steps. Initially, we will develop a user-friendly web interface using HTML5, CSS, and JavaScript frameworks such as React or Angular. Concurrently, we will set up a server with Python packages like Django or Flask to manage computational processes. To ensure efficiency and scalability, NeuroAI/omics may undergo further optimization of the algorithms and containerization using Docker to simplify deployment and ensure consistency across different environments. We will implement security measures, including HTTPS, robust encryption, and authentication protocols, to safeguard user data.

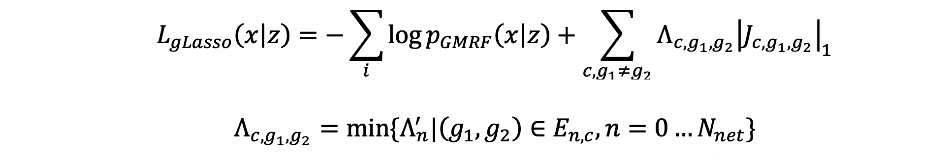
Aim 1b. Optimize the network inference capabilities of NeuroAI/omics using the graphical LASSO technique. NeuroAI/omics will enable users to fine-tune models with their own datasets, facilitating personalized research and uncovering novel insights into cell-type-specific regulations. NeuroAI/omics will support a customizable analysis environment, expanding its applicability for personalized research inquiries. Moreover, users will have the flexibility to choose from alternative processing pipelines based on their preferences, ensuring a customizable experience that meets diverse research needs.

Previous work and preliminary results. 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 regulation50,51,52,53,54. 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, demonstrating our capacity to handle and analyze genomic data from varied sources, as highlighted in our publications55,56,57. 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.

Graphical-LASSO approach. We will enhance the network inference for the integrative model, using a graphical-LASSO training approach64 instead of the maximum-likelihood approach used in the cornerstone paper44. 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. Below, we briefly summarize how LNCTP’s variables and energy function are set up before introducing the graphical-LASSO approach. The variables of LNCTP are partitioned into genotypes , intermediate phenotypes , hidden (latent) factors , and high-level/complex traits . These are further indexed, with intermediate phenotypes associated with cell types denoted as , and denoting phenotypes associated with bulk measurements. A set of variables represent the estimated cell fractions in the bulk observations. For an individual , represents the alternative allele dosage at a common single-nucleotide polymorphism (SNP) (where ), and represents a normalized summary gene expression in cell type for individual . The probabilistic model for LNCTP is defined as:

(1) The parameters of the model are represented by , and acts as a hyperparameter. As suggested by the notation, takes the form of a GMRF conditioned on , while is a stochastic deep neural network. Furthermore, the parameters and reflect the sparsity structure arising from the eQTLs and GRN linkages, respectively.

(2) We will use the following graphical-LASSO-based loss as a training objective to enhance LNCTP training. This takes the following form (writing for the -norm),

whererepresents the penalty associated with an edge between genes in cell type . This is determined by defining a set of penalties for diverse prior cell-type-specific networks, where is the edge set for prior network in cell type , and is the penalty associated with network type . The prior networks may include cell-type-specific GRNs, cell-type-specific protein–protein interaction networks, and other chosen network types. Additionally, denotes the fully connected network for all cell types, and is the number of network types; hence, an edge receives penalty if it is not in any prior network. The terms and in may be set as hyperparameters. We will provide a Bayesian optimization search method for best parameters.

Datasets, Evaluation, and Deliveries. The tool will utilize a comprehensive array of datasets, ensuring coverage across a wide range of brain cell types and conditions. To evaluate the tool's performance, we will implement benchmarking strategies against both simulated data and real-world case studies. The development process will adhere to standard software development practices, including version control (e.g., Git) for managing the codebase and collaborating with team members, unit testing to ensure individual components function as expected, and integration testing to verify seamless interaction between different modules. The tool will feature a well-documented application programming interface (API) to facilitate easy integration with other systems. We will perform regular refactoring of the codebase to improve code quality, maintainability, and performance. We will use branches to develop new features and fix bugs without affecting the main codebase, set up a build process to compile and package the tool for deployment, and establish a deployment pipeline to automate the process of releasing new versions. The deliverables will include a fully operational online tool with complete documentation, tutorials, and use examples, as well as source code and documentation via a version-controlled repository (e.g., GitHub). A README file with instructions for setting up, running, and contributing to the project, along with a user manual and API reference, will be provided to assist users and developers in leveraging the tool effectively. By incorporating these software development best practices and leveraging the power of d3.js, the resulting web-based tool will provide researchers with a robust and user-friendly platform for analyzing diverse brain cell datasets.

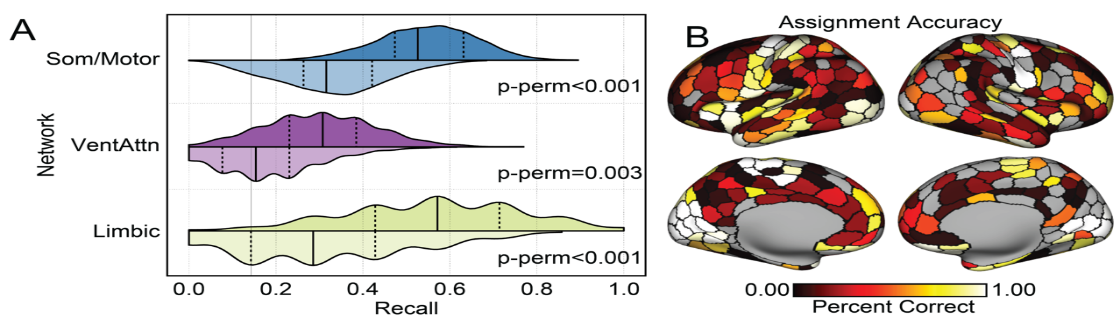
Pitfalls and Alternative Approaches. (1) Integration complexities between different omic layers and (2) the computational demands of real-time analysis present challenges. For (1), we may explore other processing methods for integration. For (2), further optimization for speed could enhance practical software deployment. Alternatively, we could split off the high-demand computational models from basic functions.

**C.2 Aim 2. NeuroAI/layers: Integrating Advanced Machine Learning for Enhanced Interpretability.**

Rationale. We aim to expand NeuroAI’s functionality by incorporating additional data types, including imaging and other single-cell omics modalities. Our goal is to enhance the framework based on previous research results that show promise in combining genetic and imaging data. This will empower the scientific community to gain a comprehensive understanding of various brain phenotypes using both pre-trained and custom models. To achieve this, we plan to integrate LLMs with NeuroAI’s core models to improve interpretability. Additionally, we will use advanced analytical methods like CNNs, graph transformers, and generative diffusion models. NeuroAI/layers will offer a flexible framework with a collection of foundational models for various data types that can be selectively applied to new datasets. Using energy-based optimization techniques, these models will generate predictors for phenotype properties, such as polygenic risk scores, heritability estimates, prioritized gene networks, and the effects of GWAS variants across multiple data types.

Aim 2a. Incorporate an imaging layer into LNCTP. We will expand LNCTP’s functionality by first integrating multi-modal imaging data through a sophisticated deep learning framework. 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. The framework will support a joint embedding space that aligns imaging features with genomic data through for instance attention mechanisms, which provides direct mapping between brain structure and molecular characteristics. This integration could help predict psychiatric and brain-related phenotypes by combining genotypic, omic, and imaging data, offering insights into the intricate relationships among brain structure, function, and genetic factors.

Previous work and preliminary results. Based on our previous experience in linking structural and functional connectivity data with diverse psychiatric traits, we can conveniently optimize the existing LNCTP architecture to include imaging biomarkers. Our team has developed methods that can jointly analyze this disparate information, applying brain-imaging-based predictive models to study cognition, personality, and mental health across the lifespan13-15. Recent projects have established how the choice of algorithms16, phenotypes3,17, and covariates may impact biological interpretability, model performance, and generalizability, particularly across demographic groups18-20. Our pilot analyses demonstrate the feasibility of integrating imaging and genetic data to uncover features of brain organization with cellular bases21. We also examined the association between functional networks and the spatial distribution of cell types, inferred from patterns of gene transcription in data from the Allen Human Brain Atlas. In doing so, we demonstrated that imputed cell distributions spatially track the macro-scale organization of the cortex, both at the level of individual cell types and multicellular profiles (Fig. 4)98. These pilot analyses highlight the feasibility of Aim 2 and the potential to uncover features of connectome organization with stereotyped cellular bases. For instance, genetic factors that differentially impact certain molecular cascades or associated cell classes could preferentially alter the function of distributed brain networks. For example, PVALB interneurons help tune excitatory and inhibitory drives in the cortex22 and are implicated in working memory23, goal-directed attention24, and psychotic illness25.

*Figure 4. Large-scale functional network assignment can be predicted by cell-type abundance in post-mortem tissue. (A) The probability of correctly classifying a parcel to the associated network (SVM recall). Darker distributions reflect 1,000 classifiers trained on real network labels. Lighter distributions represent classifiers trained on network labels shuffled by a Hungarian spinning method that controls for spatial autocorrelation. (B) Accuracy for network assignments across the cortex.*

The proposed NeuroAI/layers will utilize our recent network parcellation21, refining prior work by Yeo et al.26 We have developed a core pipeline based on this parcellation to robustly estimate a participant’s functional connectome in the form of a matrix of normalized Pearson correlation coefficients between the normalized time courses at each pair of network nodes. We have validated core features of this pipeline in multiple published reports7,8,11,12,27, including our open-access Genomic SuperStruct Project (GSP) data release28. We also established that connectomes can be used to predict the presence of active psychotic symptoms in previously unseen individuals diagnosed with schizophrenia and bipolar illness20. Further, utilizing GSP and Human Connectome Project data, we recently established that inter-subject differences in connectome organization can be used to predict individuals’ cognition, personality, and emotion15. These results demonstrate that connectivity-based imaging features are strongly predictive of high-level traits, suggesting that their use in integrated polygenic models will be effective. Furthermore, we have established appreciable twin-based and SNP-based heritability of functional connectome features8,11, implying the feasibility of polygenic imputation models of functional imaging features. Our team has also used human and non-human primate transcriptional atlases to establish techniques for mapping spatial (i.e., imaging data) profiles of gene expression (mRNA transcription) among anatomical regions10 and functionally defined networks7,11,12. Through this, we discovered synchronized genetic signatures of distributed functional networks and revealed hundreds of potential molecular mechanisms supporting brain functions in health and disease. Collectively, these discoveries demonstrate the possibility of linking *in vivo* estimates of network function and gene expression, prioritizing tractable sets of genes that may contribute to the onset of psychiatric illness.

Imaging Layer for NeuroAI/layers. By leveraging these advancements and our established methodologies, we will integrate a layer that analyzes imaging data into the LNCTP framework. We will particularly consider resting-state functional MRI datasets, which can be processed to produce correlation matrices of regional activity (i.e., functional connective [FC] matrices). 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 conditions20) and genetic/omic signatures.

Aim 2b. Layering on other architectures and extensions for enhanced performance, scalability, sustainability and interpretability. To advance the capabilities of NeuroAI/layers, we propose to integrate additional state-of-the-art deep learning architectures and extensions. The goal is to implement a modular system that can incorporate various models tailored for different types of omics data. We aim to enhance the precision and interpretability of phenotype predictions based on this modular system and our rich history with deep learning methods, such as graph transformers and LLMs. For instance, LLMs could provide human-readable explanations of model predictions, reasonings, and biological mechanisms, while attention mechanisms can reveal the association between different data. NeuroAI/layers will organize the expanded framework as a modular collection of foundation models for various data modalities, which can be applied selectively to new data. To ensure the scalability, sustainability, and reproducibility of NeuroAI/layers, the platform will be developed following sophisticated software engineering principles. Besides the modular design, version control tools like Git, alongside automated testing and continuous integration pipelines, will be employed to maintain software reliability and ensure that updates or extensions do not compromise functionality. Sufficient documentation such as user guides and example workflows, can encourage a broader user base, while open-source licensing will encourage community engagement and collaboration. We aim to provide a space for feedback, troubleshooting, and user-driven improvements. These strategies will position NeuroAI/layers as a sustainable, scalable, and widely accessible platform for multi-omics research.

Previous work and preliminary results. Our previous work has demonstrated significant 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 formats58. We developed MolLM, a pre-trained model that captures biomedical text and molecular information, enhancing performance65. Preliminary studies revealed the potential of LLMs and chain-of-thought reasoning to enhance complex reasoning tasks and develop autonomous agents66. Our Multi-disciplinary Collaboration framework significantly improved LLM reasoning in medicine67, and ML-Bench demonstrated LLMs' ability to utilize open-source libraries68. Additionally, our structure-aware fine-tuning improved LLMs' capability to generate complex structured data69, and the BioCoder benchmark illustrated our proficiency in bioinformatics coding and domain-specific challenges70. Finally, we fine-tuned an LLM to predict protein phase transitions, showing superior performance and interpretability, particularly for Alzheimer’s disease-related proteins71. 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 accuracy62. 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 landscape59. 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 disorders60. We also have successful experience in using CNNs to 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 enrichments72. ThermoNet, a 3D-CNN that accurately predicts mutation-induced changes in protein stability, has demonstrated its utility in clinical and biophysical applications73. In addition, attention mechanisms have emerged as an important approach for extracting interpretable insights from transformer models and have been successfully applied in various biological contexts81, 82. We have also leveraged attention extensively to interpret genomic variants in our recent paper83. In an earlier precursor model to LNCTP called DSPN, which used a fully connected neural network in lieu of a linear top layer, we developed a specialized technique known as rank projection trees 84,85). This methodology highlights key genes and pathways implicated in brain disorders and has been pivotal for elucidating complex disease mechanisms. Furthermore, we employ widely used, relatively straightforward methods for interpretability—such as SHAP and ablation—to evaluate feature importance in large-scale genomic or imaging datasets. Because these methods are highly accessible, they serve as a natural entry point for many prospective users of NeuroAI, who may be more comfortable with them than advanced statistical or neural network approaches.

Layers to be implemented for NeuroAI/layers. Our framework is designed to integrate a variety of advanced machine learning methods, including (but not limited to) large language models (LLMs). In recent years, deep learning technologies have shown tremendous success in handling large datasets, and not including them here would risk overlooking their potential benefits. By enabling users to incorporate diverse models—ranging from LLMs like Llama-3 or GPT to convolutional neural networks (CNNs) and transformer-based architectures—our tool can provide flexibility for mixing and matching methods depending on a user’s specific research objectives. For example, an LLM could be fine-tuned on the outputs of NeuroAI's models to allow the LLM to generate natural language explanations of the factors influencing the predictions with more accessible and actionable results. LLMs could also guide feature selection and model development by identifying relevant biological concepts and relationships from the literature, help prioritize the most informative features. Beyond LLMs, CNNs can capture spatial hierarchies in imaging data for analyzing functional and structural patterns across different brain regions. Transformer models can serve as a layer to manage and interpret the vast sequences of genomic and transcriptomic data. With self-attention mechanisms to weigh the significance of different genes or transcripts in relation to each other, NeuroAI/layers could identify critical biomarkers and gene interactions that predict complex phenotypes. Transformers may also be embedded as layers in NeuroAI/layers to decipher gene-regulatory sequences or transcription factors. These pre-trained models, adapted from vast datasets, can be utilized to specific genomic data to help explain complex regulatory mechanisms and provide real-time interactions. Aiming to provide a flexible and scalable framework for integrating various machine learning methods the modular system allows users to use specific strengths from a selection of tools while omitting unnecessary modules to meet performance and interpretability needs.

Datasets, Evaluation, and Deliverables. OpenNeuro is a platform for the aggregation of neuroimaging data. It contains several MRI datasets of relevance, allowing the exploration of many target phenotypes. We will leverage rich psychiatric collections, such as the UCLA Consortium for Psychiatric Phenomics41. Resting-state data exist on 81 schizophrenia patients and controls42 and a sibling-based study (n=99) of working memory deficits in schizophrenia43. The UK Biobank is a major international health resource29. As of the January 2020 data release, 37,848 participants are available with complete and usable imaging data. In our pilot analysis, we processed these data through the proposed connectivity pipeline, passing in-house MRI quality control. We will utilize our network parcellation21. Preprocessing steps will adhere to those described previously30: slice-time correction, motion correction, regression of motion parameters, ventricular signal, white matter signal, whole brain signal, linear trend, and low-pass temporal filtering retaining frequencies below 0.08 Hz. Detailed quality control procedures have been developed and will be used for the present project28. These include measures of slice- and voxel-based signal-to-noise (SNR), estimated absolute movement and instances of micro-movements (<0.2 mm). Runs with low SNR or excessive motion will be excluded. Participant time courses for each node will be calculated by averaging signal time courses of all constituent voxels/vertices. We will examine 400 cortical regions of interest21 and 19 non-cortical regions31. Pairwise Pearson correlation coefficients will be computed between the time courses of each pair of nodes and Fisher-normalized. The resulting correlation matrix represents the participant’s functional connectome. We have validated this pipeline in multiple published reports and open-access data-releases28. Additionally, the acquired imaging protocol includes high-resolution anatomical sequences optimized to have sufficient contrast-to-noise for automated analysis. We will process structural MRI scans using the FreeSurfer software, providing automated algorithms for volumetric segmentation, cortical reconstruction, and subject registration to a common spherical coordinate system. The accuracy of structural estimates derived from FreeSurfer has been validated against histological analysis and manual measurements.

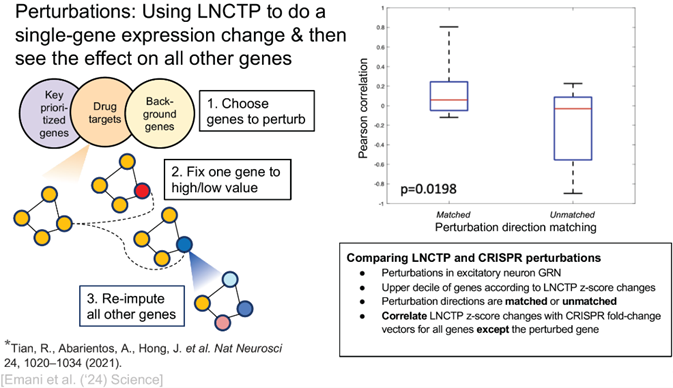
Pitfalls and Alternative Approaches. (1) A potential challenge is the integration complexity due to the high dimensionality of omics and imaging data. To address this, we can simplify the models using dimensionality reduction techniques and other robust algorithms that have been demonstrated to handle such complexities.

**C.3 Aim 3. NeuroAI/simviz: Simulating Drug Effects on Cellular Phenotypes and Providing Comprehensive Visualization.**

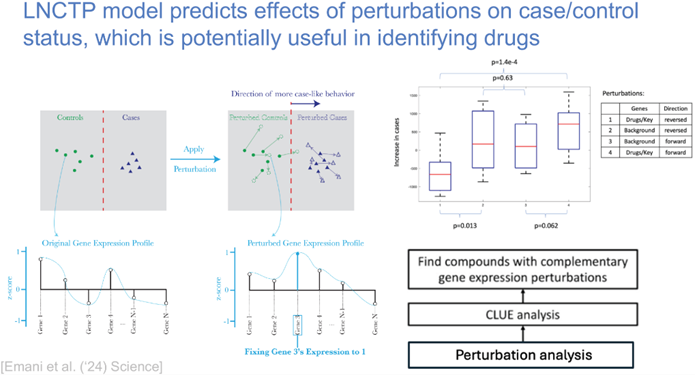
Rationale. We plan to create a flexible simulation module that builds on top of NeuroAI/omics and NeuroAI/layers to perform detailed "what-if" analyses. This module can simulate the effects of drugs on cell behaviors and gene networks, allowing users to ask questions like, "If I add this drug, how will the expression of gene X in cell type Y change?" Additionally, we will develop a flexible visualization module that combines real-time processing with cloud-based technology and an easy-to-use interface to integrate into the NeMO archive. This visualization tool will help users understand how the model works and observe interactions between different cell types or nodes. It will also enable the visualization of genetic risk and co-heritability, providing a clear way to interpret the genetic factors influencing various phenotypes.

Aim 3a. Develop a mechanism to address questions about potential drug effects on genes and cell types. This module will use the current framework to predict changes in gene expression and cellular behavior post-drug application, ensuring robustness and ease of interpretability for users from different research backgrounds.

Previous work and preliminary results. 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 data75. 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 data76. 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 accuracy77. 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 techniques78. 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 models44. 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 performed a series of perturbation analyses as outlined in Figure 5 and 6 to evaluate how well the LNCTP model can predict changes in gene activity. Figure 5 illustrates the way we pick specific genes to target, change their activity levels, and observe how this can affect other genes in the network. To make sure our predictions are accurate, we compared them with real lab experiments using CRISPR gene editing. The results showed our predictions matched well with actual biological changes (p=0.0198). In Figure 6, we show how this simulation method could help find new drugs. We observed what happens when we turn genes up or down, and whether these changes help move diseased cells back toward a healthy state. By testing different ways of changing genes, we could find patterns that could point us toward promising new treatments.

*Figure 5: Simulation of gene perturbations using the LNCTP framework. (Left) Schematic overview of the perturbation analysis workflow: (1) Target genes are selected from key prioritized genes, drug targets, or background genes; (2) Individual genes are perturbed by fixing their expression to high/low values; (3) Expression values for all other genes in the network are re-imputed to simulate network-wide effects. (Right) Comparison of LNCTP predictions against CRISPR perturbations. Box plots show Pearson correlations between matched vs unmatched directions.*

*Figure 6: LNCTP model’s potential use for perturbation analysis. (Left) Demonstration of the perturbation process: the model simulates the effect of targeted gene perturbations on overall expression patterns. The expression profiles before and after perturbation are shown as z-scores across genes. (Right) Box plots comparing the effects of different perturbation strategies (reversed vs. forward direction in drug/key genes and background genes) with statistical significance indicated by p-values.*

Module for simulating gene perturbations and drug effects. We will utilize this module to conduct *in silico* perturbation analysis, where we perturb the expression of a specific gene and observe the induced expression changes in other genes, as well as the ensuing changes in trait propensity. This serves two purposes: (1) We can apply the framework to predict how a drug will modify gene expression and cellular behavior, providing insights into how that drug may influence gene expression, for example, in brain cells. (2) Using the input genotype data, we can help identify candidate drugs that may reverse aberrant gene expression changes, restoring them to normal levels. We will collect drug target genes from public databases by cross-referencing drug-indication and drug-target connections from multiple sources. We will consider a drug to be associated with a particular disease/indication if it is listed for prescription for that indication, and then extract the target genes of these drugs. We have consulted DrugBank, Ruiz et al., and ChEMBL for drug-indication pairs. The drug target genes were sourced from the DGIdb, STITCH, phamGKB, and ChEMBL databases (see Table 2). Based on this information, we can easily connect a drug with its associated disease and target genes. Additionally, we can integrate two drug repurposing datasets, ConnectivityMap (also known as the Drug Repurposing Hub) and PRISM, which provide data on cellular responses to various drugs. We will use these resources to help predict how a drug will modify gene expression and cellular behavior.

|  |  |
| --- | --- |
| **Database** | **Description** |
| ConnectivityMap | Uses cellular responses to identify relationships among diseases, genes, and therapeutics |
| Open Targets | Uses human genomics data for systematic drug target identification and prioritization |
| DrugBank | Comprehensive and reliable drug data resource for pharmaceutical research |
| DGIdb | Open-source search engine for drug–gene interactions and the druggable genome |
| ChEMBL | Manually curated database of bioactive molecules with drug-like properties |
| STITCH | Database of known and predicted interactions between chemicals and proteins |
| PharmGKB | Curated knowledge about the impact of genetic variation on drug response |
| PRISM | Profiling relative inhibition simultaneously in mixtures |

*Table 2. Databases for Drug Target Identification and Gene Interaction Analysis. This table lists the potential databases for the NeuroAI project for identifying drug targets and analyzing gene interactions. Each entry includes details on the database's primary focus, such as linking drug indications to gene targets or exploring chemical*–*protein interactions.*

A group of math equations

Description automatically generatedIn addition, we will provide an interface to perturb target genes in each cell-type using a trained model, evaluating their predicted effects on cell-type specific gene expression, omics, and selected high-level phenotypes. To perform a perturbation, we will use the following conditional form of the LNCTP energy model: denotes the perturbed gene and cell type, whose expression is set to 1 or -1 (or another desired offset), and is a delta function whose value is 0 if expression is true and 1 otherwise.

The perturbations modeled using the above energy will form the primitive actions that may be used to perform more complex queries. For instance, in simulating the effect of a drug on gene expression, the drug target may be perturbed in a positive or negative direction, as predicted by the drug’s mechanism of action. Combinations of drugs may thus be screened for ‘rescuing’ a given phenotype, by finding those combinations of perturbations that reverse a given phenotype. Alternatively, given a particular drug’s observed expression profiles across cell types, additional drugs may be found that are predicted to minimize side effects by supplying a desired ‘target’ expression profile to match. Users will be able to manually conduct such simulations using our interface. We also will offer an LLM agent-based interface to generate desired sequences of atomic perturbation actions to address a user’s natural language query. Using a contextual learning approach, as in Gupta et al.32, the AI agent will generate appropriate sequences of perturbations in response to a user’s query. We will provide pre-specified contexts appropriate for drug discovery tasks, as described above, and allow users to augment these with novel contexts suitable for more complex queries.

Aim 3b. Develop analysis and visualization modules and integrate them into the NeMO archive. In this aim, we will introduce and implement various popular visualization and analysis tools for better interpretation of the omics data, either from our pre-processed data or potential new data from users. These visualization tools, with real-time processing and cloud-native capabilities, will be integrated into the NeMO archive.

Previous work and preliminary results. We have sufficient experience in developing tools that support interpretation purposes, as well as cloud-based platform for real-time processing ability. For instance, we developed "Gene Tracer," a cloud-based voice-controlled tool for interactive querying and visualization of genomic information, not only able to proces large genomic datasets but also remains responsive and accessible to users from any location61; a comprehensive practical guide that documents the ins and outs of generating brain visualizations and outlines the arguments supporting a transition from graphical user interfaces to code-based figure generation63. We also established NeMO Analytics, which supports hundreds of registered users and hosts over 1k datasets. Over the past several years, we have supported numerous genomics-based publications and data collections, organized into well-structured and accessible pages. During the project, new displays will be made available on NeMO Analytics to handle various data33. The NeMO Analytics infrastructure has following components: A dataset uploader and curator that allows users to upload data and choose from various displays for interactive visualizations of the data; a manager component to group and arrange datasets and to specify different levels of sharing among these datasets; a browser to visualize gene expression that provides valued resources such as annotation and link-outs to repositories; embedding with Epiviz34, a linear genome browser that enables concurrent visualization of genome accessibility and epigenetic information as well as the expression results; a tool to compare and visualize expressions under different conditions of the same dataset; and a single-cell workbench to perform *de novo* analysis from scRNA-seq data or start with existing analyzed data for analyses and visualizations. The workbench is based on the Scanpy35 pipeline and utilizes faster data processing than traditional pipelines. This web-based platform enables it to support multiple publications33,36-39.

Analysis and visualization modules with NeMO compatibility. The visualization module will build upon established analytical methods to provide reliable and accurate interpretations of biological data, ensuring that our tool remains grounded in scientifically validated techniques for trustworthy analyses. For instance, NeuroAI/simviz includes developing a web-based tool for analyzing brain cell datasets using d3.js, a powerful JavaScript library for data visualization. This visualization module will graphically represent outcomes from our trait prediction and imputation frameworks, allowing users to explore how genetic and cellular characteristics influence various brain phenotypes. Another feature of the module will be the ability to generate heatmaps that detail the effects of perturbations, such as drug interventions or genetic modifications, on cellular phenotypes. NeuroAI/simviz will provide heatmaps to visualize changes in gene expression or cellular activity under different conditions, highlighting significant responses and interactions related to disease mechanisms or therapeutic effects. NeuroAI/simviz will also feature the PsychScreen79 interface, a user-friendly portal that facilitates the exploration of data and results. PsychScreen can serve as a gateway for researchers to interact with the system, combining real-time data processing capabilities with powerful visualization tools. Then, NeuroAI/simviz, NeuroAI/omics, and NeuroAI/layers will be made compatible for integration with existing genomic databases and tools (i.e., NeMO Analytics) for enhanced functionality. We will perform testing on various aspects, such as functionality, performance, and user acceptance, to ensure the tool's reliability and user-friendliness. Our NeMO team will identify tools for inclusion into NeMO Analytics, review prototypes, generate specific use cases to drive user interface development, and incorporate analysis tools into the NeMO analytics site to ensure the broadest possible dissemination to the user community. Additionally, as many other useful tools require different data formatting, creating an issue of uniformity, we will follow the previously established Matrix and Analysis Metadata Standards (MAMS)40, a set of standards to uniformly format single-cell data. We have engaged tool developers to use this data model. MAMS formatting enables R and Python modules to deliver the data to users in easy-to-use, standard formats that can be directly imported into widely used single-cell analysis packages such as Seurat and Scanpy. Thus, utilizing MAMS in NeuroAI can help ensure FAIR use of data generated by this project. We anticipate that the adoption of a uniform data model will greatly facilitate the interoperability of the datasets and tools developed in this project. NeMO Analytics, our target integration platform, is a cloud-based web tool that has demonstrated success in visualizing functional genomics data across multiple experiments simultaneously. With over 840 registered users, 4,200 unique visitors, and more than 2,566 datasets, it provides a robust foundation for our implementation. The platform currently offers a gene expression viewer, single-cell analysis workbench, and transfer learning capabilities. To manage computational demands efficiently, we will implement a tiered access system: (1) pre-computed results for common analyses, (2) limited computational resources for basic model analysis, and (3) containerized packages for users requiring large-scale calculations. Users can either utilize the platform's resources within defined computational limits or opt for a paid cloud subscription for intensive computations.

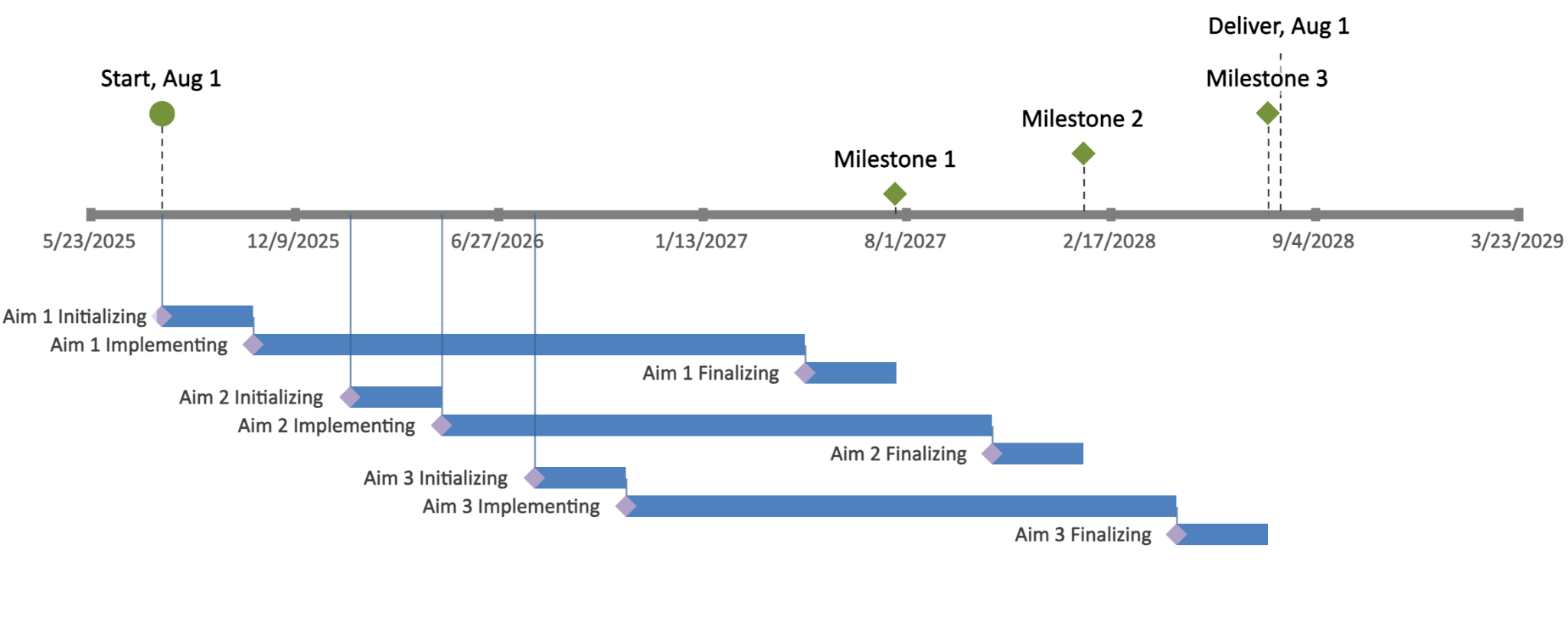
*A diagram of a cloud computing process

Description automatically generatedFigure 7: NeMO Archive Architecture Diagram. The diagram shows the overall architecture of the NeMO Archive Data Center, including the ingest system, the data processing system, the release processing system, and the data dissemination tools. NeMO also supports multiple APIs to support programmatic access to data and metadata at the data center.*

Security and privacy are crucial in our implementation strategy. Our system will leverage Google Cloud Platform (GCP) security infrastructure for both authentication and authorization of controlled access data. All data, whether open or controlled access, will be encrypted both in transit and at rest. We will implement OAuth authentication across all subsystems, including the Submission and Data portals, ensuring a unified and secure user experience. Our team brings extensive experience in data privacy and security, currently managing 1 PB of open data and 1.5 PB of controlled access data across NeMO and SCORCH (comprising 848,552 public files and 59,770 controlled access files). Our expertise in maintaining privacy and sanitizing multi-omics datasets is demonstrated through multiple publications in Cell and other prominent journals (Gursoy et al.86-91; Harmanci & Gerstein92). For technical implementation, we will integrate NeuroAI with NeMO's established infrastructure through interconnected subsystems. The integration will utilize NeMO's web resources for project information, data portals for discovery, and APIs for programmatic access. Our implementation will make pipelines cloud-compatible through Docker and Nextflow frameworks, with deep learning models deployed via Hugging Face for rapid inference. We will implement a tiered system where common queries are handled through cached results and pre-trained models, while complex simulations utilize scalable cloud computing solutions. The underlying data management will leverage MySQL databases for tracking simulation assets and Neo4J graph databases for modeling relationships between perturbations, all integrated with NeMO's existing processing subsystems for data ingest, analysis, and access control.

Datasets, Evaluation, and Deliverables. We will utilize a wider array of datasets, ensuring extensive coverage across various brain cell types and conditions. The evaluation will involve benchmarking strategies against both simulated data and real-world case studies to demonstrate the robustness and reliability of the visualization and analysis tools. Deliverables will include the deployment of a fully operational online tool integrated with NeMO Analytics, complete with user documentation, tutorials, and example use cases to facilitate ease of use and broad adoption within the research community.

Pitfalls and Alternative Approaches. Possible challenges include (1) ensuring the accuracy of simulation predictions and (2) the computational demands of real-time processing. Similarly, we may simplify the models for faster computation or use approximations for close predictive accuracy to reduce complexity.

**D. Timeline and milestones.** The project will begin in August 2025, and proceed through three main aims, each with initializing, implementing, and finalizing phases, overlapping for continuous progress. Key milestones include completing Aim 1 by Summer 2027, Aim 2 by Winter 2027, and Aim 3 by Summer 2028, followed by finalization and delivery. The project will use an iterative development cycle on GitHub, with minor releases, regular testing, and validation, and major releases timed with the milestones. Detailed guidelines for training and testing models will be provided for potential users.