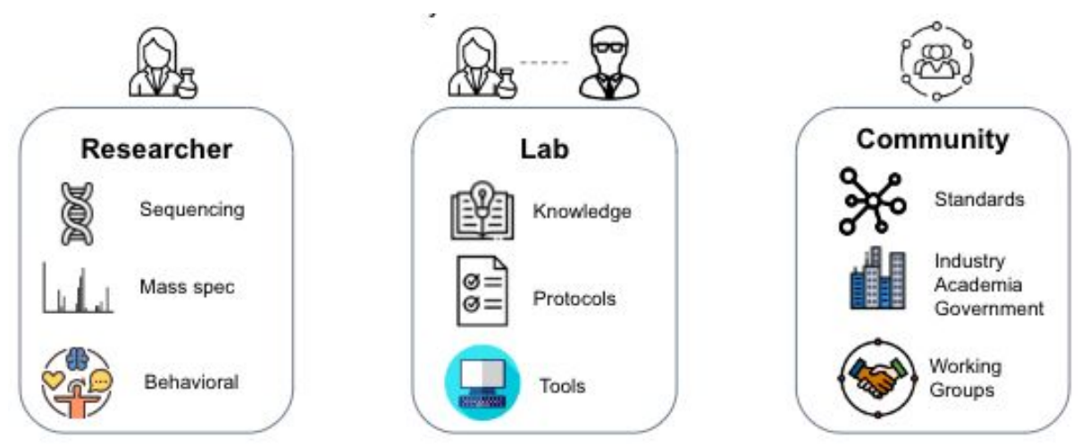
AI-Driven Harmonization and Curation of Data for Alcohol Researchers

Mohammed Eslami[1], Alex Verbitsky[1], Mark Weston[1], Nihal Salem[2], Laura Ferguson[2], Anna Warden[2], R. Dayne Mayfield[2]

[1]Netrias, LLC [2]Waggoner Center for Alcohol and Addiction Research, The University of Texas at Austin, Austin, Texas

INTRODUCTION: Data Curation for AI is Complex

- AI has the potential to revolutionize biological research and discovery to gain insights of organ/cell-type specific responses and associated mechanisms that lead to diseases such as alcohol use disorder (AUD)
- A central challenge of the application of AI for AUD is that data and metadata come in a variety of shapes, formats, representations, and scales



- Standardizing variable names,assessing quality, imputing values, and connecting datasets is a manual, laborious process hindering the development of large training corpi across labs for the application of AI
- Here, we present a computational toolkit for AUD researchers to help them get data AI-ready. AUD researchers will be able to use AI for AI.

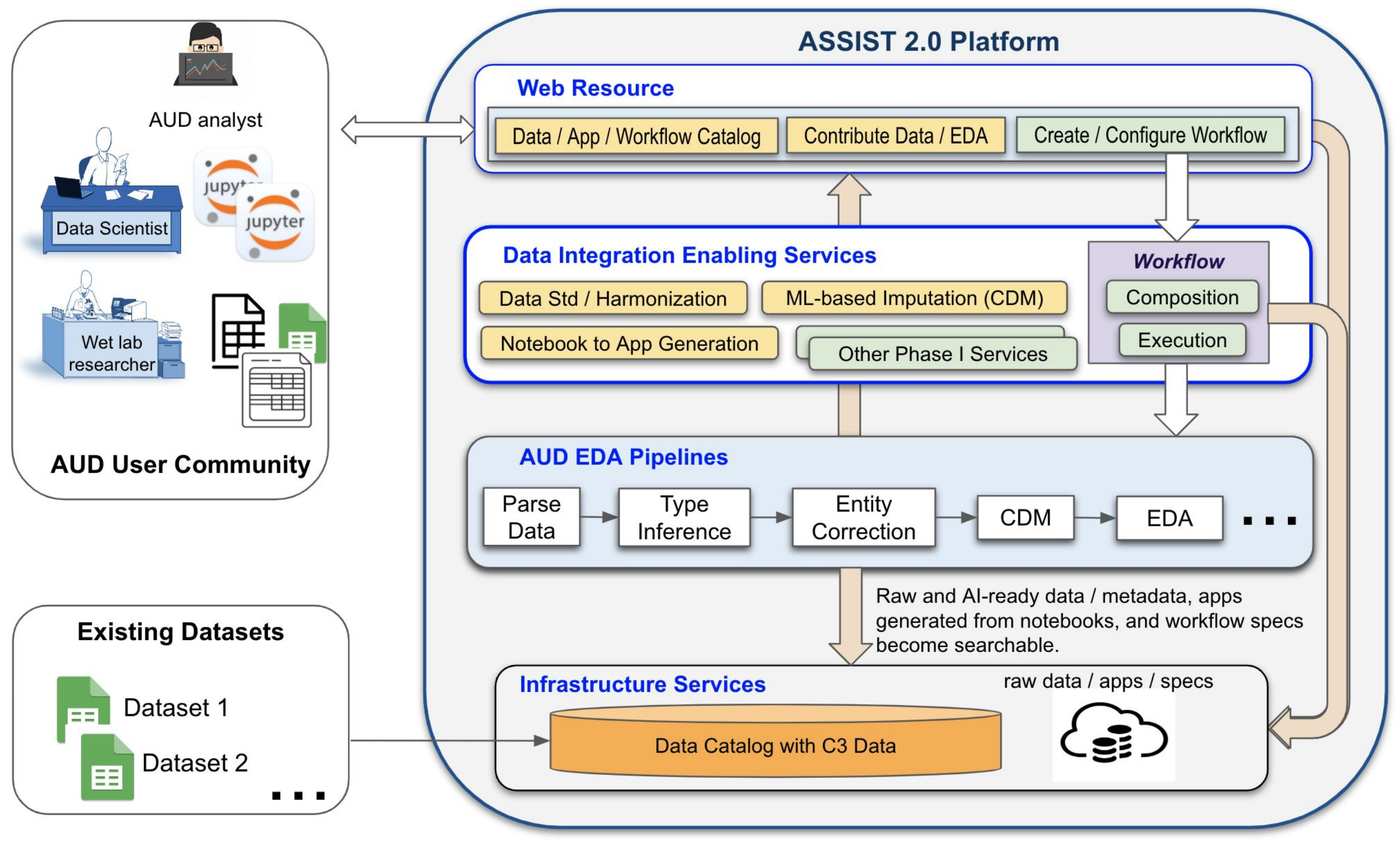
AIM: Simplifying the Collection of C^3 Data

- **Consistent** annotations of data across researchers and labs with semi-automation
- **Complete** datasets through inference of condition outcomes not executed in multifactorial experiments
- **Connected** datasets and applications for users with little software experience

SYSTEM OVERVIEW: AI-Enabled Harmonization

Alcoholism Solutions: Synthesizing Information to Support Treatments (ASSIST) 2.0

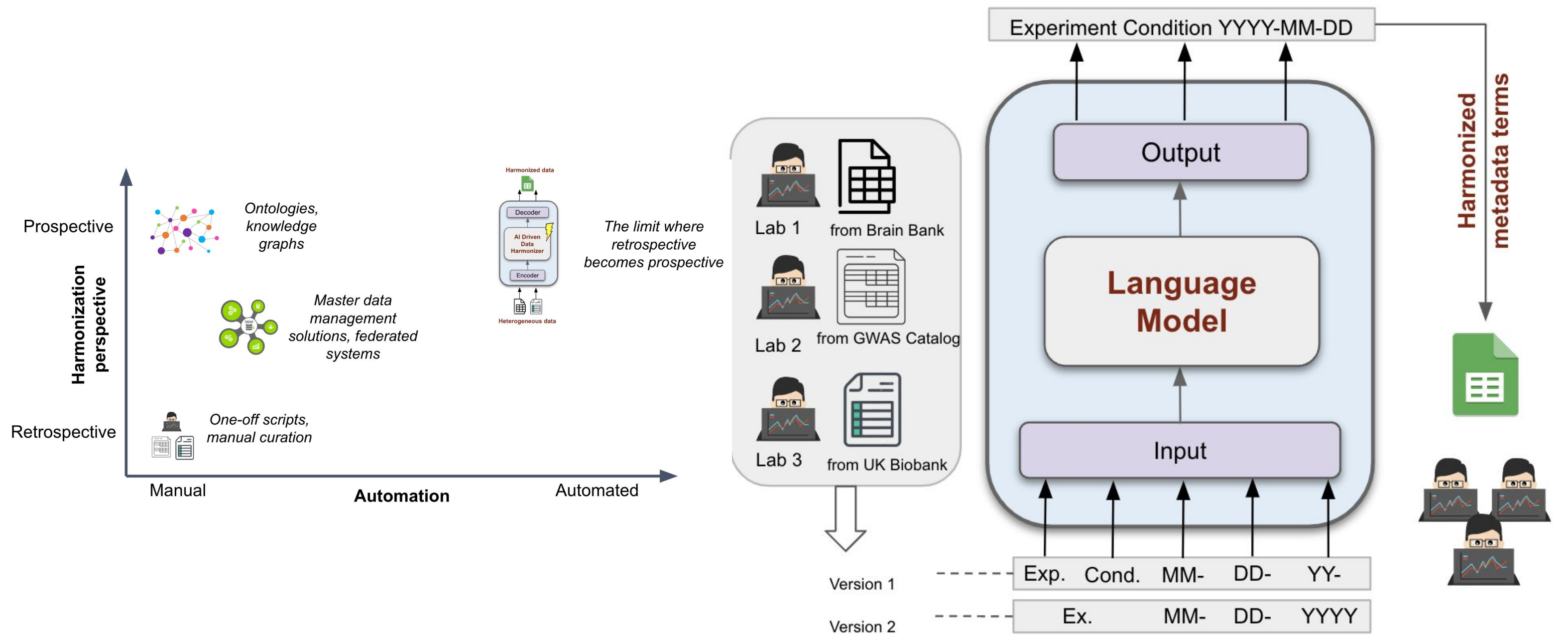
- **Web Resource** enables users to access tools and data
- **Data Integration Enabling Services** provide access to tools that enable the collection of C^3 data as well as tools from ASSIST 1.0 (Critical Gene Identifier)



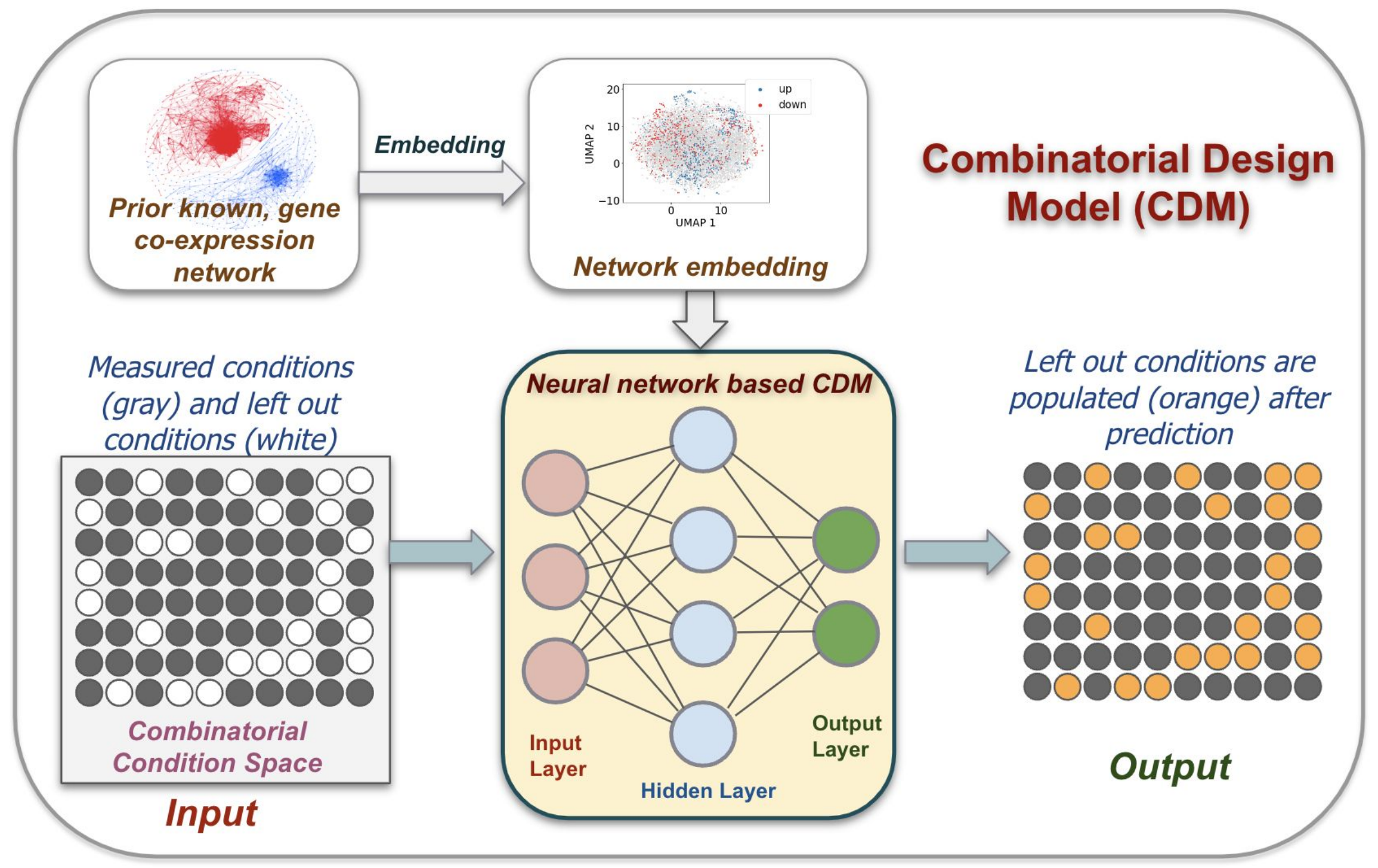
- **AUD EDA Pipelines** enable users to compose the tools they need together to setup their own data harmonization and processing pipelines
- **Infrastructure Services** is a backend database and app store that includes all datasets made publicly available as well as applications users would need to harmonize and analyze their data

METHODS

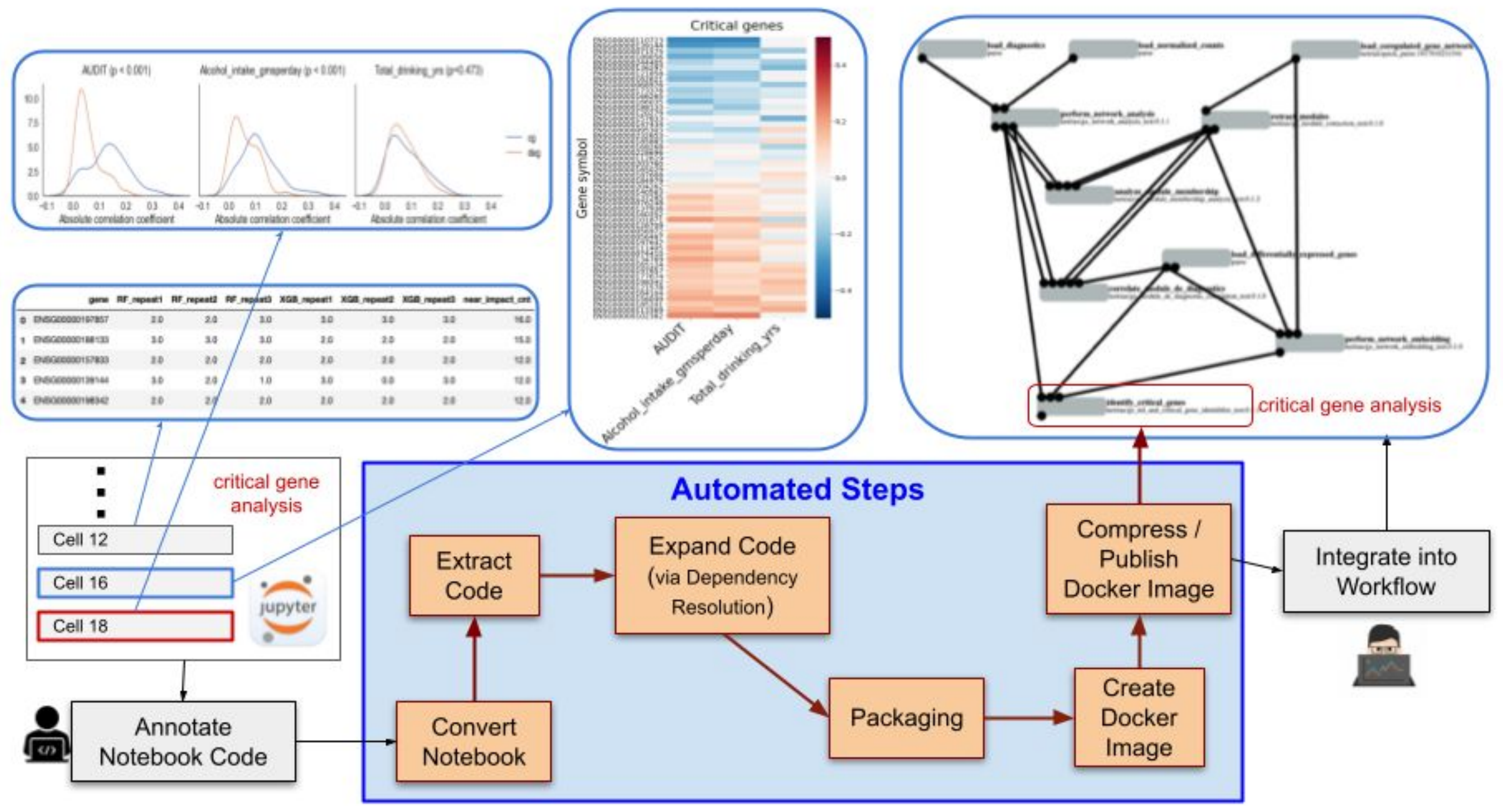
1. **Language models translate terms across labs**: Forcing every researcher to adhere to a specific ontology or set of terms is infeasible, as each has his or her own unique set of questions, experimental variables, and analyses of interest. Small and large language models can help translate terms across labs to provide each lab the flexibility they need to collect and share their data.


Prospective
Retrospective
Harmonization perspective
Manual
Automation
Automated
Ontologies, knowledge graphs
Master data management solutions, federated systems
One-off scripts, manual curation
Harmonized data
Decoder
AI Driven Data Harmonizer
Encoder
Heterogeneous data
The limit where retrospective becomes prospective
Lab 1
from Brain Bank
Lab 2
from GWAS Catalog
Lab 3
from UK Biobank
Version 1
Version 2
Exp. Cond. MM- DD- YY-
Ex. MM- DD- YYYY
Input
Language Model
Output
Experiment Condition YYYY-MM-DD
Harmonized metadata terms

2. **Combinatorial design model predicts left out conditions**: Predict whole transcriptome cell type specific responses to ethanol with prior knowledge, bulk RNASeq of control and ethanol, as well as controls of cell types


Prior known, gene co-expression network
Embedding
Network embedding
UMAP 1
UMAP 2
up
down
Combinatorial Design Model (CDM)
Measured conditions (gray) and left out conditions (white)
Combinatorial Condition Space
Input
Neural network based CDM
Input Layer
Hidden Layer
Output Layer
Left out conditions are populated (orange) after prediction
Output

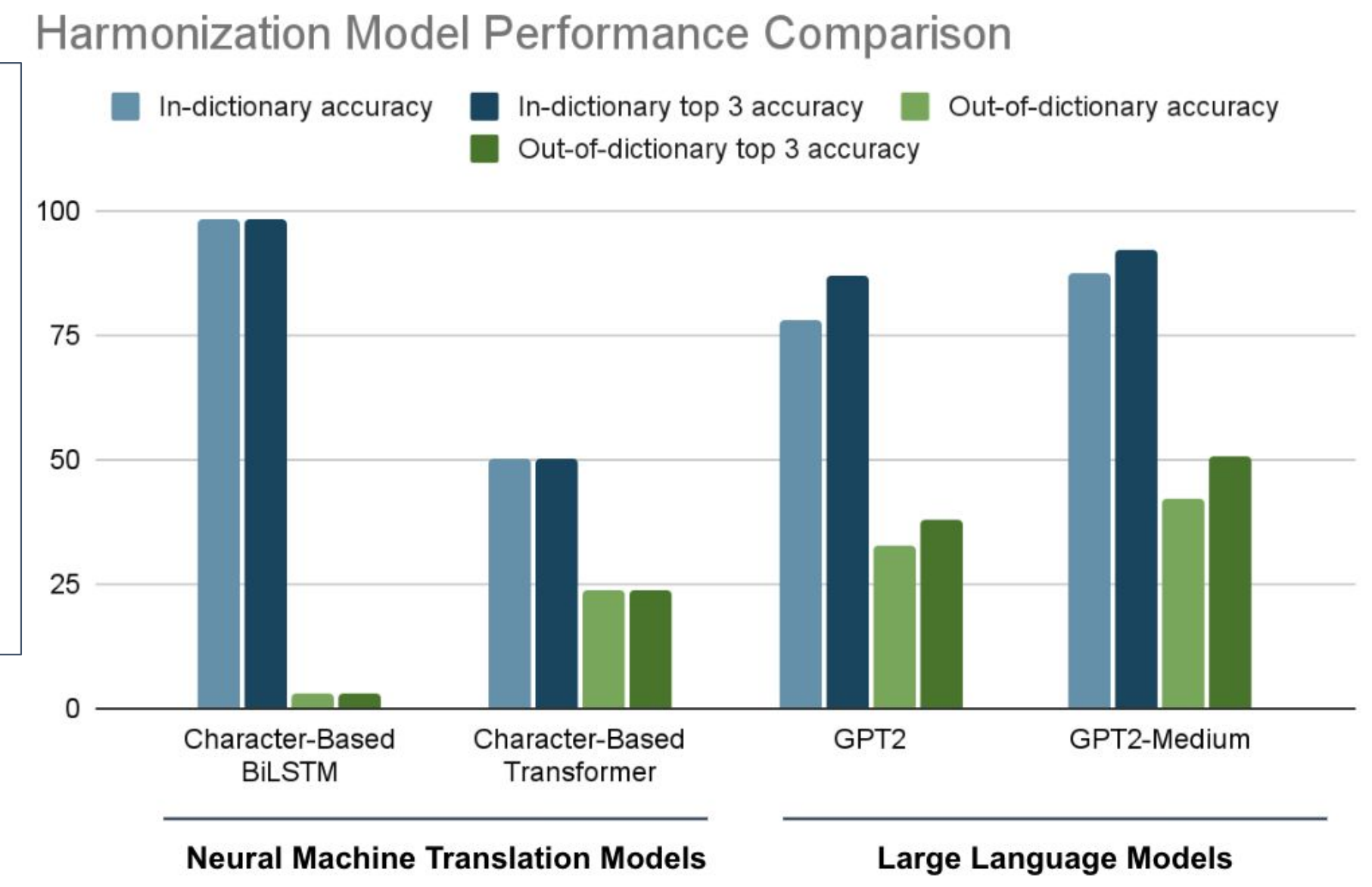
3. **Automated Generation of Applications from Custom Analyses:** Generate publishable, deployable applications from existing Python/R notebooks and scripts with minimal background in software engineering



RESULTS

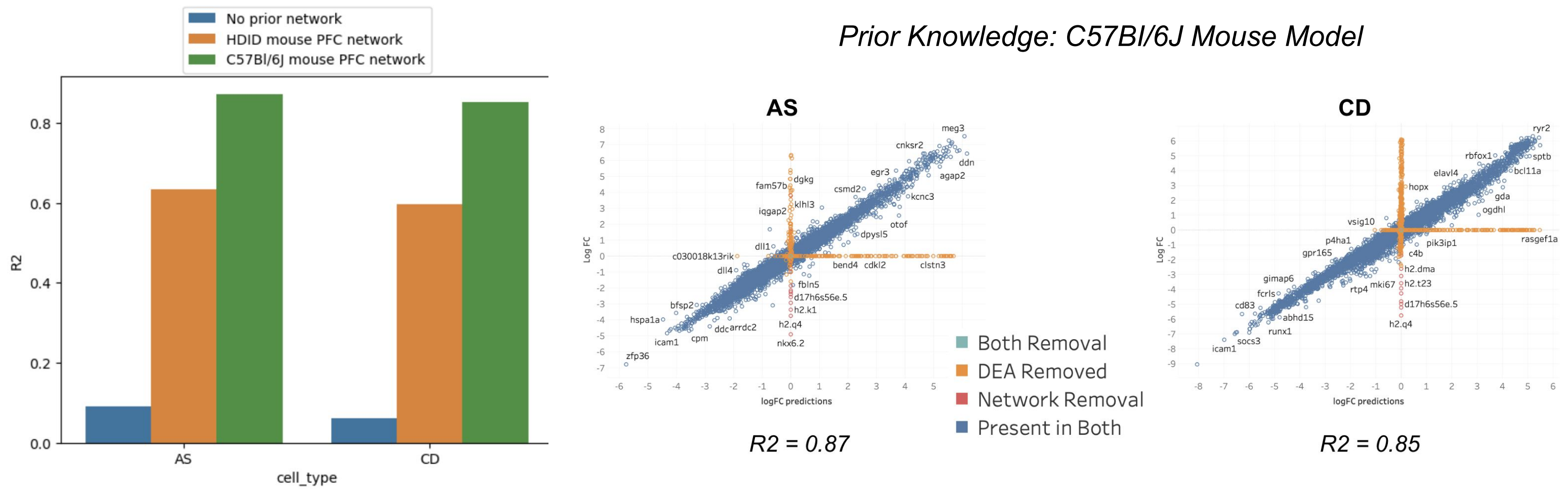
1. **Language Models Accurately Harmonize Terms:** We collected 24K standard terms from GEO, PhenX, WHO lexicon of alcohol and drug terms, and internal WCAAR data. We then generated acronyms, misspellings, abbreviations, and synonym substitutions of these terms which resulted in 1.5M representations of standard terms. We then trained (fine-tuned) a (large) language model and evaluated its ability to harmonize in-dictionary (standard terms used during train/fine-tuning) and out-of-dictionary (standard terms not used during fine-tuning) terms.

Generated **1.5M representations** of standards from alcohol related terms in these datasets. An example term and associated variations is "*Total number of Drinking Years*" could be represented as ["*Total # of Drinking Years*",...,"*Tot Num Drink Yrs*"]



Small language models are great for closed, slowly evolving ontologies while the LLMs general understanding of English makes it better for rapidly evolving ones.

2. **CDM Accurately Predicts Ethanol Response of Specific Cell Types:** We trained a machine learning model to predict whole transcriptome response to ethanol for specific cell-types in the prefrontal cortex for a C57Bl/6J mouse. Training data included Total homogenate - Ctrl, Total homogenate - ethanol, Astrocyte (AS) - Ctrl, Microglia (CD) - Ctrl, while testing data was: AS - Ethanol, CD - Ethanol. The model was also tested with three forms of prior knowledge: None, WGCNA network of an HDID mouse, and WGCNA network of C57Bl/6J mouse. We evaluated the model with an R^2 metric between predicted vs actual log(FoldChange).



CONCLUSIONS

- AI-based language models show significant promise in the ability to semi-automatically harmonize terms across labs. This will allow researchers to focus on the science and leave data curation to the tools!
- CDM provides researchers with the opportunity to identify the best conditions to run in the lab before running them. Inferences serve as *in-silico* predictions that shed light on conditions not yet tested in the lab saving time, money, and labor.
- We are actively looking for new use cases, data, and users to enhance the capabilities of our system. Please reach out to meslami@netrias.com if you are interested in using the ASSIST platform for your research.

DISCLAIMER AND ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health 75N94020C00003. Netrias, LLC has developed all technology and code relevant to this project in collaboration with the Waggoner Center of Alcohol and Addiction Research has provided the subject matter expertise, data, and annotations.