Strategies and Structures for Biological Datasets Integration and Knowledge Discovery

A Dissertation

Submitted to the Faculty

of

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment for the

Degree of Doctor of Philosophy

in

Data Science

By

Erin Teeple

July 24, 2023

APPROVED:

Professor Elke Rundensteiner Worcester Polytechnic Institute Advisor	Professor Randy Paffenroth Worcester Polytechnic Institute Committee Member
Professor Carolina Ruiz Worcester Polytechnic Institute Committee Member	Doctor Virginia Savova Sanofi External Committee Member

Abstract

Correlations between variables in biological data reflect underlying processes, but data science problems in this domain include how to perform dimension reduction or integrate data in ways that do not lose information and how to use such data for discovery and prediction tasks. In part 1, this dissertation details the generation of a dataset for machine learning from US air quality and cause-specific mortality records. An innovative CCA-derived epidemiological analysis is then presented for novel quantification of exposure-outcome associations, achieving stronger and significant quantification of air quality and health outcome association through covariation vs more commonly used multiple linear regression. Conceptual understanding of covariation modeling then guides alignment of single-nucleus RNAseq datasets by CCA-based features to extract new insights into relationships between regional cell states and disease. In part 2, the problem of extracting information from a knowledge graph is considered for the task of predicting drug indication status for target-disease pairs using as input features aggregated association evidence scores from the Open Targets platform. In part 2, first, an innovative new approach for the task leveraging local network topology is shown to achieve improved prediction performance over previously published works. The second work in part 2 is another novel approach for the same classification task which achieves further improved performance by integration of external biological data resources via a feature engineering informed by collaborative filtering and network embedding concepts. In part 3, the problem of transforming raw biological data with correlated features into data structures for knowledge discovery is further explored with illustration of how such preparations can generate custom data structures which are suited for feature generation as shown in part 2. Innovations in this work include the development of new integration strategies for biological data (part 1, 2, 3), development of multiple novel indication status prediction methods for use in the Open Targets platform (part 2), and generation of new data-derived networks for knowledge discovery (part 3).

Acknowledgments

I am truly grateful to my advisor Professor Elke Rundensteiner and PhD committee members Professor Randy Paffenroth, Professor Carolina Ruiz, and Doctor Virginia Savova. Their insightful guidance and generosity in sharing their knowledge have helped me to grow as a scientist, and this dissertation would not be possible without their support. I also thank the other faculty and students at WPI and my colleagues at Sanofi from whom I have learned so much. Lastly, I thank my husband Aaron, children Noah and Lailah, and my entire family for their support and encouragement.

Publications

PUBLICATIONS FEATURED IN THIS DISSERTATION

This dissertation describes work done in the following five papers Together, these offer solutions enacting strategies and structures for biological datasets integration and knowledge discovery.

[1] **Teeple E.** Joshi PN. Pande R. Huang Y. Karambe A. Latta-Mahieu M. et al. Integrated label transfer for oligodendrocyte population profiling in Parkinson's disease and multiple system atrophy. *Proceedings of the 15th International Joint Conference on Biomedical and Health Informatics* 2022.

[2] Han Y. Klinger K. Rajpal DK. Zhu C. **Teeple E. (senior author)** Empowering the discovery of novel target-disease associations via machine learning approaches in the Open Targets platform. *BMC Bioinformatics* 2022; 23(1): 1-19.

[3] **Teeple E.** Chang Y.C. Rajpal D.K. A target-specific evidence function for indication expansion queries in the Open Targets platform. Published in Proceedings IEEE BHI-BSN 2021.

[4] **Teeple E.** Kuhlman C. Werner B. Paffenroth R. Rundensteiner E. Air quality and cause-specific mortality in the United States: association analysis by regression and CCA for 1980-2014. *HEALTHINF 2020.*

[5] Teeple E. Jindal K. Kiragasi B. Annaldasula S. Byrne A. Chai L. Sadeghi M. Kayatekin C. Shankara S. Klinger K.W. Sardi S.P. Madden S.L. Kumar D. Network analysis and human single cell brain transcriptomics reveal novel aspects of alpha-synuclein (SNCA) biology. *https://www.biorxiv.org/content/10.1101/2020.06.05.137166v1*.

OTHER PUBLICATIONS

[6] Ryan SK. Zelic M. Han Y. **Teeple E.** Chen L. Sadeghi M. Shankara S. Guo L. Li C. Pontarelli F. Jensen EH. Comer AL. Kumar D. Zhang M. Gans J. Zhang B. Proto J. Saleh J. Dodge JC. Savova V. Rajpal DK. Ofengeim D. Hammond TR. Microglia ferroptosis is regulated by SEC24B and contributes to neurodegeneration. *Nat Neurosci* **26**, 12–26 (2023). https://doi.org/10.1038/s41593-022-01221-3

[7] Boddupalli CS. Nair S. Belinsky G. Gans J. **Teeple E.** Nguyen T-H. Mehta S. Guo L. Kramer ML. Ruan J. Wang H. Davison M. Kumar D. Vidyadhara DJ. Zhang B. Klinger K. Mistry PK. Neuroinflammation in neuronopathic Gaucher disease: Role of microglia and NK cells, biomarkers, and response to substrate reduction therapy. *eLife* 2022; 11: e79830.

[8] Tasdemir-Yilmaz O. Gans J. **Teeple E.** Escobedo J. Bu J. Madden S. Mueller C. Ramachandran S. Dorsal Root Ganglia Single-Nucleus Transcriptomics Reveal Cellular and Molecular Responses to High Dose AAV-Induced Toxicity. *Molecular Therapy* 2022; 30(4): 528-529.

[9] **Teeple E.** Hartvigsen T. Sen C. Claypool K. Rundensteiner, E. Clinical performance evaluation of a machine learning system for predicting hospital-acquired clostridium difficile infection. *HEALTHINF 2020.*

[10] Hartvigsen T. Sen C. Brownell S. **Teeple E.** Kong X. Rundensteiner E. Early prediction of MRSA infections using electronic health records. *HEALTHINF 2018.*

Contents

1 | Introduction

Biological research is often exploratory, with aims including identification of novel disease mechanisms or further understanding of fundamental functions or processes. The application of methods from data science are particularly suited for these aims, but a fundamental challenge in the application of data science in biomedical research is the necessity to align diverse approaches from mathematics, computer science, and statistics with equally complex knowledge sets and information sources for domain-specific application. This dissertation engages with this challenge across tasks including correlated variables analysis (Part I & III), data integration strategy (Part II), and network analysis (Part II & III). The unifying theme across the projects presented in the dissertation is the requirement to take a systematic approach in such works starting from the parallel formulation of a research aim as simultaneously a biological domain question and a data science problem. For each problem, then, a first step is to identify suitable data and appropriate methods. Methods for data analysis in this work are then applied with evaluation both for biological insights resulting from an analysis and their analytical validity.

1.1 Motivation

The central motivation for each of the works in this dissertation is to apply data science methods to extract new insights from biological datasets. In problem 1, the focus is on understanding and applying canonical correlation analysis (CCA) methods to datasets with intercorrelated features firstly to quantify associations between dataset covariance of air quality and mortality measures and secondly for preprocessing to identify corresponding cell states from RNAseq data. In problem 2, the works focus further on data sets integration approaches including network

embedding techniques applied for extraction of information from a public aggregated evidence data resource. In problem 3, work on network creation and analysis is presented that illustrates the diversity of network information which can be extracted from biological datasets for use in applications such as described in problem 2 and validation approaches.

1.2 Canonical Correlation Analysis (CCA)

Canonical correlation analysis was first described in 1936 by Hotelling as a method by which to examine latent (canonical) relationships between multi-dimensional vectors $X = (x_1, x_2, \ldots x_n)$ and $Y = (y_1, y_2, \ldots y_n)$ which have non-zero Pearson correlations (ρ) among variables such that $\rho(x_i, x_j)$, $\rho(y_q, y_r)$, $\rho(x_k, y_p)$ are non-zero for some variables [11, 12]. Non-zero intercorrelations imply that linear combinations of variables in the two sets may be predictable by or predictive of the others. The mathematical procedure underlying CCA seeks to find linear combinations of X and Y with maximal correlations with each other, and these linear combinations may then be used to examine and characterize relationships between multidimensional X and Y domains, with correlations between X and Y sets in the canonical dimensions taken to represent latent factors accounting for correlated set covariations [11, 12]. CCA is specifically suited for the situation where we have multiple intercorrelated exposure measures and multiple interrelated health effects with the relationship between variable sets being driven by complex, multidimensional latent phenomena, broadly, the effects and interactions of air pollutants on interdependent body systems in individuals with a wide range of susceptibilities and underlying health conditions. Unmeasured latent phenomena can be further explored through canonical weight comparisons as will be shown, by focusing on statistically significant high magnitude cross-set collaborations as will be further detailed in the following sections. CCA-based methods have found recent use in single-cell RNA seq analysis workflows, where this procedure is

applied as a processing step for multi-dataset integration to identify correlated cell states from the canonical weights – the application of this procedure is facilitated by the measurement of gene expression using standard nomenclature in these expression datasets.

1.3 Network Construction and Link Prediction in Biological Data Resources

Problem 2 of the dissertation concerns extracting information from aggregated knowledge resources and integration of data from knowledge base sources with orthogonal data resources. Work for problem 2 uses aggregated information available from the Open Targets Platform. Open Targets is a public-private research initiative that began with the formation of the Centre for Therapeutic Target Validation (CTTV), a collaboration between GSK, the Wellcome Sanger Trust, and the European Bioinformatics Institute (EMBL-EBI) (www.opentargets.org). CTTV was renamed to the Open Targets Initiative in 2016 and the aim of work by this organization is to advance the development of methods for exploring and integrating large volumes of scientific data for the support of target validation analyses. Open Targets makes available both a web-based search platform and an application-programming interface (API) where systematically aggregated association evidence may be searched and/or downloaded. The foundation of searches in Open Targets is its defined ‘target’ entity, which is a protein, protein complex, or RNA molecule. Targets in Open Targets are named by their official Human Gene Nomenclature (HGNC) gene name and annotation by ENSEMBL stable ID is also recorded. The use of these standardized identifiers facilitates aggregation of information from diverse scientific sources and integration of Open Targets summary information with other data. [13-15]. The works presented include an application of machine learning for predicting target-disease therapeutic status from evidence on the Open Targets Platform and a work demonstrating a novel framework for

integrating additional biological information with this resource to improve prediction performance for indication status.

1.4 Generation and Network Analysis of Biological Data

In Problem 2, biological information networks are used together with network embedding to assign similarity scores between genes. Networks are representations of information where relationships among nodes/entities are shown with edge connections. A set of nodes connected by edges is called a graph, and the study of graphs in mathematics has a long history, with the first recognized theorem of graph theory dating back to Leonard Euler's solution to the Konigsburg bridge problem in 1736 [16]. In biomedical science, network methods are increasingly applied for the study of gene expression datasets where large gene regulatory network (GRN) models can be readily derived from high-throughput gene expression datasets which produce gene expression measures for multiple samples and/or cells [17]. With the rapid expansion of available datasets, generation of numerous networks becomes possible but further steps remain to effectively explore network topology, derive novel insights, and identify suitable networks for algorithmic use and mining as in the framework from Problem 2. Work for Problem 3 is a network analysis performed for human brain single nucleus datasets which demonstrates a breadth of network analysis directions.

1.5 Dissertation Task and Contributions to Data Science

An overview of dissertation tasks and contributions to Data Science is the following:

<u>Problem 1: CCA for Correlation Analysis and Derived Feature Embeddings</u>

- Novel dataset derived from integration of US EPA and State-level mortality data resources **engineered specifically for ML algorithms**

- Innovative **CCA-derived epidemiological analysis provides a novel quantification of exposure-outcome association** which more strongly and significantly quantifies air quality and health outcomes relationships through covariation models then linear regression.
- **Alignment of multiple datasets by CCA-based features leveraged to extract new insights** into regional cell state biologies and their relationship to disease states from transcriptomic data, overcoming the challenges of comparing results from **unsupervised clustering.**

Problem 2: Link Prediction in Evidence Networks Using Local Information and Features Generated by New Collaborative Filtering Methods

- Created new search procedure based on ML using **local graph topology** for predicting drug-target relationships from Open Targets platform aggregated association score data using platform API **which outperforms previously published approaches for this task**.
- Lead development of **novel feature engineering** project for use with Open Targets association evidence which **expands the utility of the platform information network** and **achieves a substantial improvement in performance for drug-target relationship prediction over all previously published methods for this task**.
- Application of collaborative filtering concepts in paper 2 for omics datasets integration for query applications is especially significant because the **volume, size, and variability of public Omics data archives creates challenges in access and processing times for query and data mining applications which are met by work in paper 2**.

Problem 3: Network Construction for Data Mining and Biological Insight

- Gene co-expression network topologies vary with multiple factors including disease state, tissue of interest, and cell type. Such networks can be used as inputs for custom feature generation as shown in problem 2.
- For this problem, **multiple novel gene networks are created from single-cell RNAseq transcriptomic data to achieve association rule learning, classification, and clustering tasks.**
- Data mining performed from these networks derives new insights into cell-specific and regional brain biology.

2 | Canonical Correlation Analysis (CCA) for Correlation Analysis and Feature Embedding

2.1 Air Quality and Cause-Specific Mortality in the United States: Association Analysis by Regression and CCA for 1980-2014

2.1.1 Background

Passed in 1970 and subsequently amended in 1977 and 1990, the United States Clean Air Act §7401 et seq. (1970) requires that the United States Environmental Protection Agency (EPA) set national air quality standards for six air pollutants: ground-level ozone, particulate matter, carbon monoxide, lead, sulfur dioxide, and nitrogen dioxide. To date, numerous studies have associated air pollution exposure with increased risk for adverse health events, specifically including incidences of cardiovascular events and strokes [18-23]. Notably, negative health impacts of air pollution exposure have been linked at levels of exposure even below current federal regulatory limits in the United States, including a study of all-cause mortality in Medicare beneficiaries where higher levels of exposure to small-diameter particulates and ozone in air pollution, although within federal regulatory limits, were linked with all-cause mortality [18]. Revision of federal standards may be undertaken based on evidence of exposure-outcome associations, but this requires quantified risk estimates for each level of exposure to inform this process. With respect to air pollution, quantifying such associations can present a considerable analytical challenge due to correlations among both exposure and outcome variables. This is because air pollution exposures are typically not single-exposure events and each pollutant type impacts morbidity and mortality risk via multiple interdependent organ systems and interactions with

individual underlying health conditions. Thus, it is a considerable challenge to assess air pollution effects on populations health: each health outcome is not independent of eachother and neither is each pollutant exposure.

Despite these challenges, dose-response and predictive models are necessary. One of the most common approaches for such analyses is multiple linear regression with interaction effects, where pollutant measures are the independent variables and the dependent/response variables are health effects [24]. Yet such models do not fully capture relationships between interrelated responses. This challenge motivates this work which compares multiple linear regression with an alternative approach to association analysis, canonical correlation analysis.

2.1.2 State of the Art

As noted in background above, regression model are commonly applied to quantify linear associations between exposure (independent/predictor) variables (features) and outcomes (dependent/response) measures, and these methods have been widely applied in the study of air pollution effects on human health [18-23]. Multiple linear regression is a parametric method whose parameter estimates provide coefficient estimates which are interpretable as the strength of association between exposure and outcome. Commonly reported outputs of such models are estimates of additional mortality predicted from unit increases in air pollutants, or quantifications of the relative contributions of specific measured air pollutants to a variety of health outcomes. Canonical correlation analysis (CCA), which is a method for examining associations among multi-dimensional vectors, offers a complementary perspective.

Canonical correlation analysis was first described in 1936 by Hotelling as a method by which to examine latent (canonical) relationships between multi-dimensional vectors $X = (x_1, x_2, \ldots x_n)$

and $Y = (y_1, y_2, \ldots y_n)$ which have non-zero Pearson correlations (ρ) among variables such that $\rho(x_i, x_j)$, $\rho(y_q, y_r)$, $\rho(x_k, y_p)$ are non-zero for some variables [11, 12]. Existence of non-zero intercorrelations implies that linear combinations of variables in the two sets may be predictable by or predictive of the others. The procedure underlying CCA seeks to find linear combinations of X and Y with maximal correlations with each other. In effect, these linear combinations may be used to examine and characterize possible latent relationships between multidimensional X and Y domains, with correlations between X and Y sets in the canonical dimensions taken to represent latent factors accounting for correlated set covariations [11, 12]. CCA is specifically suited for the situation where we have multiple intercorrelated exposure measures and multiple interrelated health effects with the relationship between variable sets being driven by complex, multidimensional latent phenomena, broadly, the effects and interactions of air pollutants on interdependent body systems in individuals with a wide range of susceptibilities and underlying health conditions. A notable strength in applying CCA for questions in environmental epidemiology as we present here is that CCA does not require assumption of independence pf predictors or application of domain knowledge to generate or interpret interaction terms. Unmeasured latent phenomena can be further explored through canonical weight comparisons as will be shown, by focusing on statistically significant high magnitude cross-set collaborations as will be further detailed in the following sections.

CCA has been further extended in recent years to kernel [25] and deep [26] adaptations of this method. In kernel CCA and deep CCA, data sets are projects into high-dimensional kernel or embedding spaces before CCA is performed, with these approaches permitting non-linear representation of latent relationships between datasets. Kernel and deep adaptations of CCA permit extension of this method to model complex variable interrelationships, but step away

from the use of canonical weights as interpretable coefficients (as work in this dissertation will move in the next discussed work). Thus challenges of both kernel and deep CCA are appropriate kernel selection, interpretability, and overfitting. The further development of kernel and deep CCA should then necessarily be driven by requirements of its applications. In epidemiology use as demonstrated here, interpretability of canonical coefficient relationships is an essential aspect of this project. In the following work, we will show an application of CCA where interpretability is approached from a different perspective.

2.1.3 Problem Definition

The problem defined for this analysis is that we aim to quantify associations between two sets of variables which are not independent of each other (air pollution measures and cause-specific mortality rates) which are interrelated by complex latent phenomena. To characterize this problem and provide a benchmark against which to compare our analysis and its interpretations, the work first outlines correlation relationships among variables within and between sets and results from multiple linear regression for single mortality rates are presented. We then seek to approach our problem by applying CCA to find combinations of air quality and cause-specific mortality rates which are maximally correlated and consider their canonical weight in significantly correlated dimensions.

2.1.4 Challenges

One significant challenge in this analysis comes from the lack of a readily available data set suited for our analysis. CCA requires that the two variable sets to be intercorrelated share a common identifier for pairwise analysis. While the EPA is mandated by the federal government to collect air quality measures, it is not within the scope of this agency to monitor health

outcomes. In order to perform this analysis, then, a significant part of the work was to identify and integrate appropriate data resources and develop a strategy and workflow for dataset integration.

2.1.5 Proposed Method

Proposed methods for this work include first identification, preprocessing, and integration of public datasets suitable for our task followed by implementation and analysis. Data sources for this study were selected to provide a set of variables quantifying air pollution exposure and a second set of variables quantifying cause-specific mortality rates. The common identifier used to join these records was the combination of United States federal county code and year. The **Air Quality** feature set used for this work was obtained from AirData, a website maintained by the United States EPA which offers public access to air quality measurements from outdoor monitors at 4000 sites across the United States, Puerto Rico, and the United States Virgin Islands. Available tables from AirData are annual and daily summary tables with measurements of ambient air quality rating, regulated pollutant quantities, particulate concentrations, meteorological conditions, ozone precursors, and lead. For this analysis, 'Annual Summary' tables are used for the time period 1980-2014 [27]. The **Cause-Specific Mortality** data set used for this analysis came from United States county-level age-standardized rates for the years 1980-2014 aggregated by the Institute for Health Metrics and Evaluation (IHME). IHME reports estimates for US county-level mortality rates for 21 causes of death including chronic respiratory diseases, cardiovascular disorders, and other causes. This dataset is made publicly available through the Global Health Data Exchange and is the product of a significant body of work to collect and report these data: age-standardized mortality rates reported for males, females, and combined genders as number of deaths per 100,000 people in this population are estimates

generated from review of death records from the National Center for Health Statistics, population counts from the US Census Bureau, and the Human Mortality Database referenced using the cause list from the Global Burden of Disease Study [28]. Age-standardization of mortality rates is a particularly important step in ensuring comparability of data across county locations, as population age demographics vary across the US – this dataset was selected for integration with EPA records because of its systematic methods as well as its broad coverage of the US regions for which EPA data is available.

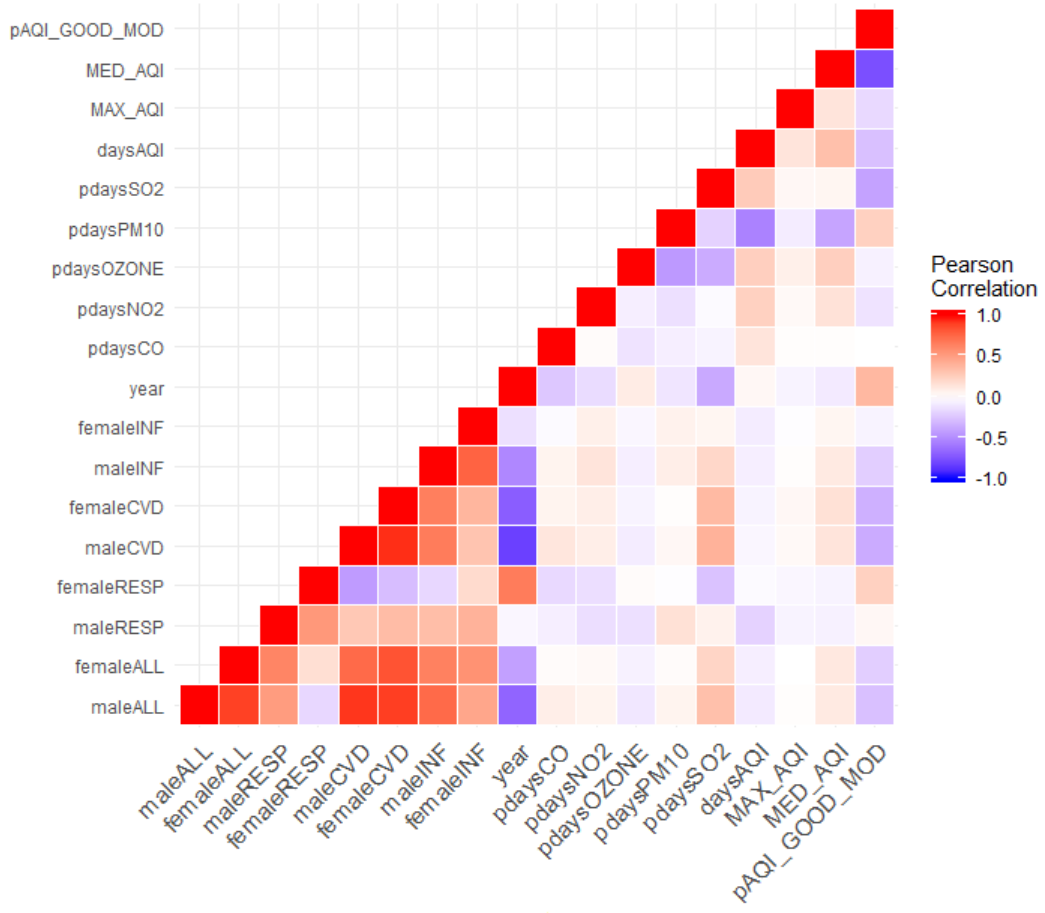
2.1.6 Analysis and Methodology

Data pre-processing and table joins were implemented in Python, version 3.6 using year and county cod as the unique row identifier. The final .csv for the integrated dataset contained 31,019 data rows uniquely identified by county location and year with sets of mortality rates and air quality measurements. Mortality rates used in this analysis were the following: male and female age-adjusted mortality rates for the following causes: all (ALL), respiratory disorders (RESP), cardiovascular diseases (CVD), and lower respiratory and other common infectious diseases (INF). These causes of mortality were selected for inclusion based on previous studies linking air pollution exposure with systemic inflammation and adverse effects on the cardiovascular and respiratory systems [18-23]. The set of air quality measurement data included the following fields: median annual Air Quality Index (AQI); maximum annual AQI; the proportion of recorded days on which AQI fell into each of the following categories: Good (0-50), Moderate (51-100), Unhealthy for Sensitive Groups (101-150), Unhealthy (151-200), Very Unhealthy (201-300), and Hazardous (301-500); and the proportion of days on which the AQI was attributed to one of the following pollutants: Carbon Monoxide (CO), Nitrogen Dioxide (NO2), Ozone, Particulate Matter (PM10), and Sulfur Dioxide (SO2). AQI is a summary measure of air quality, with the

scores ranging from 0 to 500, with lower values corresponding to better air quality. Only year-county rows with complete data for both mortality and air quality data sets were included in the final analysis. Project code is archived on github: (https://github.com/erinteeple/CCA_air).

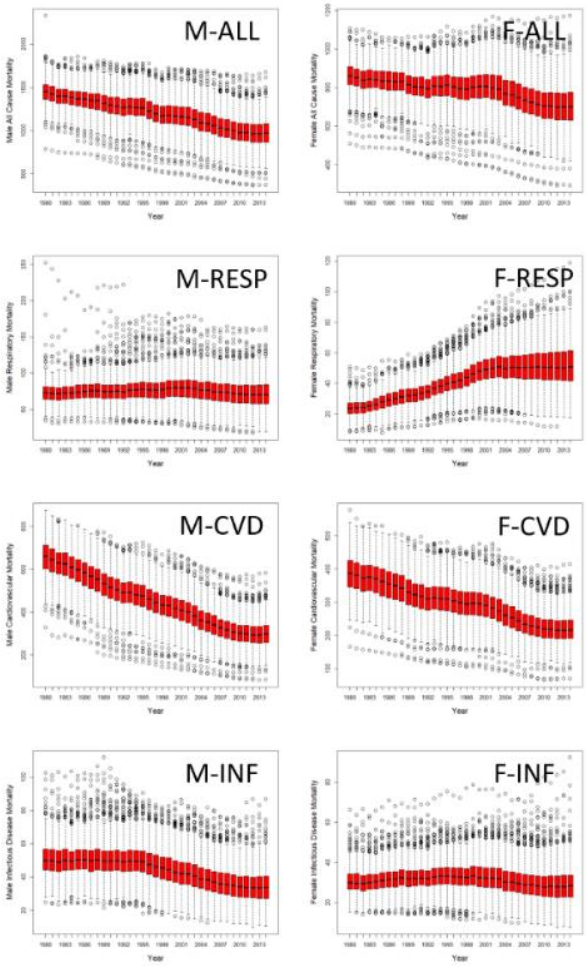
The first step in analysis of the integrated datasets was to perform an initial exploration of the data. Figure 1 from the publication of this analysis is show which presents Pearson correlations.

Linear correlations



REF [4] Fig. 1: Pearson correlations mortality rates and air quality exposure measures.

among and between air quality measures and mortality rates are observed. These multiple intercorrelations confirm that this dataset is suitable for analysis by CCA. Also included in this initial exploration was to examine whether there are relationships between year and any of the studied cause-specific mortality rates. Figure 2 shows relationships between year and each of the cause-specific mortality rates included in our analysis. By these plots, mean mortality rates by cause were observed to exhibit different general trends for males and females particularly with respect to changes in respiratory causes of mortality, thus it was decided to perform separate analyses for males and females.



REF [4] Fig. 2: Box plots showing variations in age-standardized mortality rates per 100,000 persons for time period 1980-2014 by gender and cause: M: male; F: female; ALL: all causes; RESP: respiratory; CVD: cardiovascular; INF: infectious disease.

In regards to air pollution exposure features, formal relationships between those variables warrants specific consideration and discussion. Proportions of days in each AQI rating category were combined into a single measure, proportion of days on which the AQI was in the good or moderate air quality categories. Similarly, median AQI, maximum AQI, and proportion of days on which AQI was good or moderate have a natural relationship and capture interrelated exposure patterns. These are non-independent and separate linear regression models using each of these AQI summary measures. A further consideration in using annual summary data is that reporting of proportions of specific pollutants reflects only the proportion of days on which the maximal AQI is attributed to that maximal type of pollution, meaning that air pollutants present at other levels of exposure are not captured by this reporting convention and the specific magnitude of exposure is not captured. To compensate for this lack of information, interaction terms were added in the linear regression models between AQI summary measure sand pollutant proportion terms.

Multiple Linear Regression Analysis: Multiple linear regression analysis is a multivariate parametric statistical model in which we examine linear relationships between a dependent variable and one or more independent variables [24]. Analyses of this type produce coefficient estimates for each predictor which quantify estimate magnitude and direction of a given predictor contribution to the dependent variable value, with confidence intervals indicating a probability-based range for this estimate (e.g. if the estimated of a coefficient includes 0 then no relationship between that predictor and the outcome is possible). In a multiple linear regression model, the adjusted R-squared value quantifies the ability of the model to explain variation in the dependent variable (mortality rate in this analysis) from the independent variables (air quality features).

Canonical Correlation Analysis: CCA is formulated as follows: for a data matrix M composed of feature sets X and Y which have *p* and *q* measurements, respectively, for each of N observations:

$$M = [\,\mathrm{X} \mid \mathrm{Y}\,] \begin{cases} \mathrm{X} : \mathrm{N} * p \\ \mathrm{Y} : \mathrm{N} * q \end{cases} \qquad (1)$$

CCA seeks independent, linear combinations of the X and Y set variables U_a and $V_b$which maximize $corr(U,V)$:

$$U_a = a^{\mathrm{T}} X = \sum_{i=1}^{p} a_i X_i \qquad (2)$$

$$V_b = b^{\mathrm{T}} Y = \sum_{i=1}^{q} b_i Y_i \qquad (3)$$

$$corr(U,V) = \frac{cov(U,V)}{\sqrt{var(U)\, var(U)}} \qquad (4)$$

For the analysis of intercorrelated mortality rates and intercorrelated air quality measures, where correlations between observation sets are mediated through biological mechanisms, we can see that this method generates a correlation between sets with an intuitive interpretation which is correlation between covarying set variables [29].

The results of multiple linear regression are compared with CCA for quantification of relationships between air pollution exposure and multiple cause of mortality. Proportion of variance in specific mortality rates explained using multiple linear regression is compared with cross-set correlation magnitude and significance in the canonical dimensions as well as the weight coefficients assigned in each model to particular air quality variables (β_i in linear models and canonical weights in CCA). Interpretation of the variable weights in CCA is quite different from parameter estimates in multiple linear regression. This is because in a multiple linear regression model, the variable coefficient for an independent variable is an estimate of a true linear contribution parameter of that variable to the dependent variable value derived from the regression model. In contrast, variable canonical weights in CCA are those which are assigned in a specific canonical dimension in the correlated projection. The canonical weights assigned to variables in the two sets indicate relative contributions in the correlated project in a given dimension. As such, these weightings are most informative taken relatively and together for both sets and in relation to the magnitude and significance of their specific correlated projection – this will be further explained by example in this and the subsequent work in this dissertation.

2.1.7 Experiments

For these experiments, data integration is first performed in Python and multiple linear regression and CCA analyses use R and Python [12, 30]. Adjusted R-squared values for regression models predicting annual mortality from year only as a benchmark are compared with models predicting annual mortality from year plus air quality measures. Table 1 from the referenced publication summarizes results of this first set of experiments including adjusted R-squared values for regression models predicting annual mortality i) from year only and ii) from year plus air quality measures, including interaction terms between proportions of days on which the leading pollutant

was of a specific type. For assessment of model performance, we apply the general linear test [31] and observe that compared with year-only models, a significantly greater proportion of variation in mortality is explained by models which include year together with air quality measures. These results show significant associations between air quality exposures and different cause-specific mortality rates (Table 1).

ADJUSTED R-SQUARED FOR MULTIPLE LINEAR REGRESSION MODELS

OUTCOMES		Predictor Sets			
Cause	Gender	Year Only	Year + Max AQI	Year + Med AQI	Year + Good Days
ALL	M	0.45	0.46**	0.48**	0.46**
	F	0.17	0.19**	0.21**	0.19**
CVD	M	0.002	0.06**	0.06**	0.06**
	F	0.43	0.44**	0.44**	0.44**
RESP	M	0.68	0.70**	0.71**	0.70**
	F	0.49	0.52**	0.53**	0.52**
INF	M	0.27	0.28**	0.31**	0.29**
	F	0.02	0 .03**	0.06**	0.03**

**indicates p-value < 0.05 for general linear test of significantly better model fit for year plus air quality measures model versus reduced model predicting outcome from year only.

REF [4] Table 1: Mortality rate prediction.

Air quality measure coefficient confidence intervals are then shown in in Table 2 from the referenced publication. From this table, it can be seen that for some air quality measure variables, the coefficient intervals include 0 (no effect), and some coefficient estimates are negative, which runs counter to our expectation that higher air pollution exposures are significantly positively associated with worse health outcomes. For examples, surprisingly, ozone has a negative coefficient interval for female respiratory mortality, but several considerations should be taken into account in the interpretation of this result. Particularly, other variables in the regression model are correlated both with ozone level and with the mortality outcome, such as year, which

can be seen in our exploratory plots where there are trends which differ for cause specific mortality which are changing over the study period, and looking at predictive model for respiratory mortality alone may bias results if related mortality rates such as cardiovascular events are also affected by the exposure and being recorded as the proximate cause of death.

REF [4]Table 2: Linear model coefficients.

TABLE II
PARAMETERS IN MULTIPLE LINEAR REGRESSION MODELS PREDICTING MORTALITY RATES FROM MEDIAN AQI AND POLLUTANTS

PREDICTORS		Parameter 95% CI for Mortality Targets			
Variable		ALL	RESP	CVD	INF
Year	M	-15, -15	-.1, -.1	-11, -10	-1, -1
	F	-5, -5	9, 9	-5, -5	-.1, -.1
MedAQI	M	5, 5	0.1, 0.1	2, 3	0.3, 0.3
	F	3, 3	0, 0	2, 2	.2, .2
Days CO	M	-29, 38	-11, -3	-6, 24	2, 6
	F	-11, 29	-2, 3	-5, 15	1, 4
CO x AQI	M	-5, -3	-.2, 0	-3, 2	-.3, -.2
	F	-3, -2	0, 0	-2, -2	-.2, -.1
Days NO2	M	68, 149	-1, 8	9, 45	16, 21
	F	31, 80	-3, 2	5, 31	12, 15
NO2 x AQI	M	-7, -5	-1, -1	-3, -2	-.4, -.2
	F	-4, -2	-.2, -.1	-1, -1	-.3, -.2
Days Ozone	M	110, 149	-9, -5	76, 94	8, 10
	F	47, 71	-6, -3	41, 53	4, 6
Oz. x AQI	M	-5, 4	0, 0.1	-3, -2	-.3, -.2
	F	-2, -2	0, 0	-2, -1	-.2, -.1
Days PM10	M	164, 201	3, 7	82, 98	14, 17
	F	84, 107	.6, 3	55, 67	9, 11
PM10 x AQI	M	-7, -5	0, 0.2	-4, -3	-.5, -.4
	F	-4, -3	0, 0	-3, -2	-.3, -.2
Days SO2	M	206, 236	6, 9	114, 128	14, 16
	F	101, 119	0, 3	70, 80	8, 10
SO2 x AQI	M	-6, -5	-0.2, -0.1	-3, -2	-.4, -.3
	F	-3, -3	-0.1, 0	-2, -1	-.3, -.2

**95% confidence intervals for predicting mortality rates from median AQI, year, and proportion of days on which AQI score is attributed to a specified pollutant with interaction terms between median AQI and proportion of pollutant days included in the regression to represent exposures x median severity of exposure.

Given these considerations, CCA provides a complementary perspective to multiple linear regression from the same data set. Table 3 from the reference publication summarizes the results of CCA. By CCA, we observe a statistically significant and high magnitude correlation of 0.91 between the sets of air quality exposure measurements and all-cause mortality variables. As shown in reference Table 3, in the first canonical dimension, positive canonical weights are assigned to multiple adverse air quality measures together with positive weights assigned to multiple cause of mortality rates. In addition, in this first canonical dimension, a negative (protective) weight is assigned to the proportion of days on which AQI is rated as good or moderate. Where multiple linear regression find low-moderate proportions of variation explained in individual cause-specific mortality rates for air quality measures, CCA quantifies strong significant correlation between variations in air quality across the United States and variations in cause-specific mortality. Relative to the formulation of CCA, effects of air pollution on human physiology are captured as a latent phenomena relating covariations between the mortality and air quality data matrices.

STANDARDIZED CANONICAL COEFFICIENTS

VARIABLE NAMES	CANONICAL DIMENSIONS		
	1*	2*	3*
ALL male	0.68	0.033	-0.03
ALL female	0.43	0.002	-0.02
RESP male	0.05	0.21	-0.04
RESP female	-0.65	0.05	0.003
CVD male	0.84	-0.01	-0.01
CVD female	0.71	-0.04	-0.03
INF male	0.52	-0.02	-0.13
INF female	0.13	-0.01	-0.12
Year	-0.99	-0.004	-0.02
Days CO	0.19	-0.02	0.45
Days NO2	0.12	-0.611	-0.54
Days Ozone	-0.10	-0.51	0.12
Days PM10	0.06	0.71	-0.58
Days SO2	0.46	0.06	-0.14
Max AQI	0.04	-0.25	0.004
Med AQI	0.12	-0.64	-0.0002
Good Days	-0.40	0.49	-0.06

*Indicates $p < 0.05$ for test of null hypothesis that canonical correlation in that dimension is zero. Magnitudes of correlations in the canonical dimensions were found to be 0.91, 0.31, and 0.20 for dimensions 1, 2, 3.

REF [4] Table 3: CCA coefficient estimates.

2.1.8 Conclusions

The results and scientific contributions of this work can be summarized as follows:

1) A novel 34-year national county-level air pollution exposure - mortality outcomes dataset is created and made publicly available. The approach uses federal county identifiers, permitting easy integration with other geographically coded data sets for investigators who may wish to build on this work.
2) We identify significant associations between variations in cause-specific mortality and air quality measures, with impacts on human health shown to occur even within regulated pollution exposure levels.
3) We explore the use of CCA alongside multiple linear regression for environmental epidemiology applications and show its use in relating complex exposure patterns with multiple correlated outcomes.
4) These findings have important public health and policy implications: associations between worse air quality and cause-specific mortality can inform regulatory limit revision efforts. We show that covariation in interrelated outcomes can be used to quantify harm alongside quantitation of specific cause effects.
5) The publication reporting the data science framework for this analysis highlights how this method can be adapted more broadly for application in environmental epidemiology and public health research.

In summary, the first work in this dissertation demonstrates strong and significant first- and second-dimension canonical correlations relating variations in air pollution exposure and multiple causes of mortality using data from the United States for the period 1980-2014. These results complement and extend our understanding of results from multiple linear regression analysis. Of particular note,

relationships between air pollution exposure and mortality are shown to occur across locations subject to United States federal regulatory limits which are under continual monitoring. Harm to human health from air pollution at a range of exposure limits has been shown by other investigators [18-23]. This work provides further confirmation of the relationship between air quality and health outcomes.

2.2 CCA Application: Integrated Label Transfer Oligodendrocyte Population Profiling in Parkinson's Disease and Multiple System Atrophy

2.2.1 Background

The second work in this dissertation applies CCA-derived methods in a different context, but shared between these two applications is the task of relating two sets of data with common observations (data rows) but for two matrices of intercorrelated features. For this second work, CCA is applied to generate embedded projects of data rows so that label applied to one matrix can be transferred to another. In this case, the data used is single-nucleus RNAseq data (snRNAseq). For this data type, gene expression (RNA transcripts) are quantified at the single cell level from processed tissue. Integration of expression data across samples and data sets is frequently performed with the goal of transfering cell type labels. Clustering of cells by gene expression is also performed, often followed by comparisons within clusters to identify differences in gene expression among similarly labelled cells. Label transfer saves type by not requiring that samples of the same general type be re-labelled. Another potential use for label transfer methods is to identify cell states which might be shared by samples from different anatomic regions or be altered under disease conditions. The analysis presented here presents a work of this type, with alignment between accomplished through the application of CCA not for

examination of correlated features, but rather for derivation of an embedding which is then used for the purpose of clustering similar cell states. this work describes integrated analysis of oligodendrocyte lineage nuclei sequenced from human brain putamen region tissue samples for healthy Control (n = 3), Parkinson's Disease (PD; n = 3) and Multiple System Atrophy (MSA; n = 3) subjects with label transfer to substantia nigra region tissue samples for healthy Control (n = 5) subjects.

PD and MSA are both progressive neurodegenerative diseases. PD and MSA are both synucleinopathies, which are disorders in which nervous system aggregates of α-synuclein, the protein encoded by the SNCA gene, are found in different cell types where they are linked with cell death and dysfunction. Tissue histology and genetic analyses have suggested that oligodendrocyte cell biology may be linked with synucleinopathy pathogenesis. The motivation for this work is to use snRNAseq data to examine oligodendrocyte population heterogeneity in disease and healthy control brain tissue, identify disease-associated differences in cell states and gene expression, and to relate our finding to another disease-relevant tissue dataset. Our task in this work is to group related populations of cells across datasets with varying number of cells but the same genes quantified in each dataset. CCA is applied for the purpose of embedding cells so that cell states shared between datasets can be identified for label transfer and as will be shown, the results of this work provide novel insights relating regional variations in oligodendrocyte biology to disease features of PD and MSA. This application of CCA further provides insight guiding how CCA may be further developed in its kernel and deep forms for embedding tasks, where CCA coefficient interpretability is not a primary concern of the domain application.

2.2.2 State of the Art

As background for this analysis, a brief explanation is needed on the development and ongoing refinement of single nucleus RNAseq technologies in molecular biology. snRNAseq permits individual cell-level resolution transcription profiling and has dramatically expanded our ability to understand activities and variations of cells all over the body, and particularly in the brain and nervous system with their heterogeneous, densely packed, interacting cell populations. Cells are the foundational unit of multicellular organisms, with all cells in an individual organisms sharing common DNA but varying in their transcription of this code, as reflected by differences in RNA transcripts, which are quantified at the cell level by snRNAseq methods [32]. snRNAseq data and methods have particular potential for improving our understanding of central nervous system cell biology, as the cells in the brain form dense, interconnected, and diverse networks which support dynamic and complex processes such as memory encoding, vision processing, and motor control and coordination, with these activities distributed over macro and microscopic anatomic tissue regions [32]. Sequencing of nuclei in a tissue sample generates a matrix of cell-level gene transcript counts which is called a unique molecular identifier (UMI) count matrix. In the analysis of snRNAseq data, variations in gene expression between cells are used both to cluster cell types (by similar patterns of gene expression) and also for differential expression analysis to compare between cells in a particular cluster or other grouping. Standard data pre-processing steps for UMI count matrices consist of filtering to remove low quality rows (e.g. rows with very few counts or very many more than other data rows which are taken to represent empty droplets or those with more than one cell) and data rows with high numbers of mitochondrial genes [33]. Sequencing depth variations can result in different total counts of genes being detects in different cells. Normalization of UMI count matrices is therefore undertaken as a preprocessing step to support comparability and clustering workflows within and between datasets. Normalization

methods that are commonly used are to log-transform the UMI counts matrix followed by scale factor multiplication [33] as well as an alternative, sctransform, which takes sequencing depth as a covariate in a generalized linear model and yields the residuals of a regularized negative binomial regression for use as effectively normalized data [34]. The sctransform modelling framework has been proposed as a method by which to remove technical characteristics from data while preserving cell-to-cell biological heterogeneity as sequencing depth can vary. Joint analysis of multiple samples has additional challenges, in particular, the need to integrate different datasets so that cell subpopulations are matched. A workflow developed and implemented by Stuart et al. 2019 presents a comprehensive strategy for such integrations [35]. This method applies concepts from statistical learning and combines single cell datasets through the application of canonical correlation analysis (CCA) and mutual nearest neighbors profiling for the task of identifying ‘anchors’, pairwise correspondences of cell states between datasets. Based on this integration procedures, processed reference datasets may also be used to transfer predicted cell type labels to a query sample through anchor-based transformation, as well. While presented as an efficient method for labelling cell types, in the application presented for this part of the dissertation, we first create an integrated reference constructed from Control, MSA, and PD putamen data and enlist label trans methods using Control substantia nigra data as our query.

2.2.3 Problem Definition

This work in the dissertation describes a complementary application of CCA where instead of interpretation of canonical weights being using to explore quantitative relationships between two intercorrelated sets of variables, CCA is applied for feature generation before clustering. The problem in this work and generally in snRNAseq data analysis is to take two or more UMI count matrices and assign a cell type label to each sequenced nucleus using gene expression data. For a

given data matrix, many methods exist for clustering – the cluster solution which is desired for snRNAseq is the one that resolves known cell types and/or subpopulations of interest. In order for cell populations to be compared between multiple data matrices, cells of similar types must be clustered together while cells of the same type are not to be separated due to variation between datasets, variation due to treatment or disease effects, and variation due to differences in tissue type.

Integration of multiple sample data matrices is necessary to move beyond descriptive summary of samples. Alignment of multiple datasets allows for identification of corresponding cell states based on gene expression. Once cells are aligned, population proportion and differential expression comparisons can be performed. The dimensions of UMI count matrices are a further consideration for data integration and clustering work. Data matrices typically contain transcript information for up to several thousand cells and 10-16000 genes and gene co-expression patterns differ by cell type, with some genes expressed predominantly or exclusively in specific cell types and other genes expressed variably among many cell types. Clustering of snRNAseq data can be performed using common and straightforward methods – a very straightforward method for this is that after quality filtering and normalization of data sets, those genes which have the greatest variation among cells are identified (typically about the top 2000 most variable genes); PCA is performed on this subset of highly variable genes for dimension reduction; an elbow plot is used to determine how many principle components to include in clustering, and unsupervised clustering is applied to the PC-transformed data for community detection. In Seurat, the package used for this work, the basic clustering procedure is to use a KNN graph constructed from Euclidean distance in PC space where edge weights are computed between pairs of cells from the overlap in their local neighborhoods (e.g. Jaccard distance) followed by application of the

Louvain clustering algorithm which partitions the graph into clusters based on iterative optimization of cluster modularity, with a resolution parameter which can be varied impacting cluster size [33, 35]. Once cluster identities are assigned from dimension-reduced features, analysis returns to the normalized data matrix for cluster profiling and intra-cluster comparison of gene-level expression.

The scope of this work is not to propose an alternative method for integration and clustering of single cell data but rather to present an application in a specific use case which is enabled by assignment of cell state correspondences using CCA-based embedding. CCA-based embeddings for single cell data were introduced to address the interest in finding pairwise correspondence in cell states. Due to the requirement that input data matrices related by CCA must share observation N and the fact that snRNAseq data matrices contain different numbers of cells, this issue is resolved in the CCA-based analysis of single cell data by transposing matrices so that N reflects the genes for which measurements are obtained from both sets [35, 36] while canonical weights assigned to cells are used to related them [36].

2.2.4 Challenges

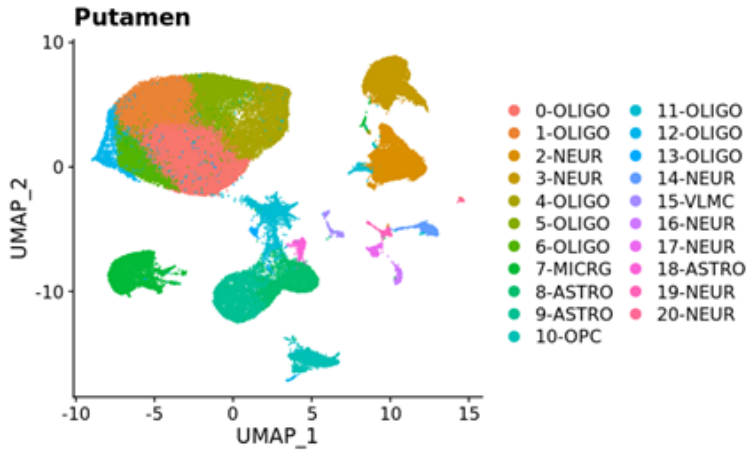
Two issues generally in unsupervised clustering which come up when working with snRNAseq data are that (1) while it is possible to vary resolution and try a number of different algorithms, it is not possible to define in advance what populations will be clustered out and (2) variability due to sample characteristics and treatment effects can separate similar cell types into different clusters. Cell composition differences in different samples can also have effect. With respect to tissues from the brain as in this analysis, neuron proportions in a sample vary by location and for a sample with many neurons and few other cell types, application of the above workflow for unsupervised clustering will lead to selection of highly variable genes that are variable in the

data and clustering is likely to separate neurons and neuron subtypes and if we wanted to compare clusters in such a dataset with clusters derived from a dataset with fewer neurons and greater proportions of other cell types (such as astrocytes or oligodendrocytes), we would not have direct correspondence in clustered cell states. The same problem arises if a treatment effect or disease state is causing greater variability that cell type differences. This issue of aligning cell populations of interest and transferring cell type labels has motivated the development of label transfer workflows. In this work, we apply such a workflow as a demonstration of a novel use case for CCA with follow up validation analyses used to obtain new insights into disease biology.

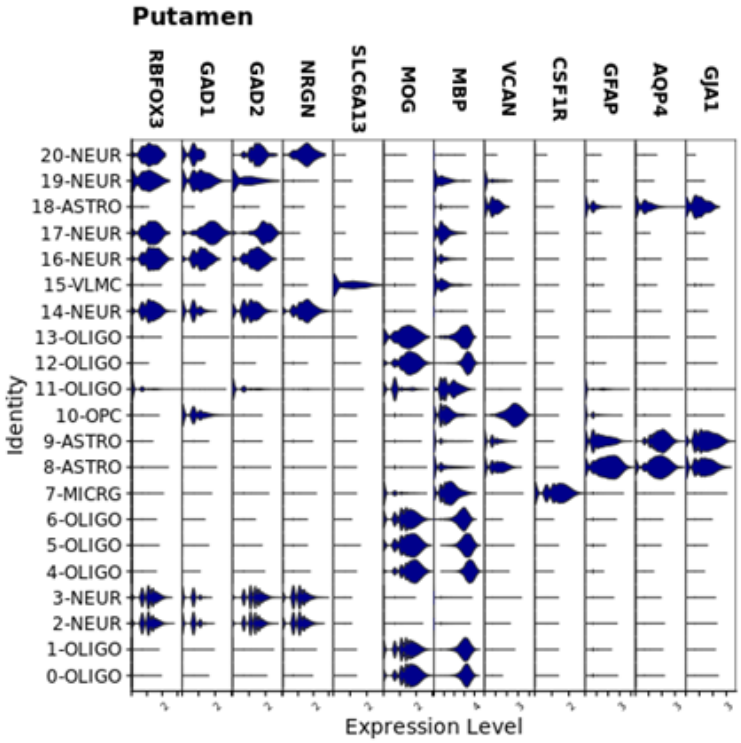
2.2.5 Proposed Method

CCA-based integration and clustering were performed for snRNAseq data obtained from human post-mortem brain putamen region tissue samples. Samples were obtained through partnerships with licensed organizations with completed pre-mortem consent for donation and ethical committee approval for sample acquisition and use. Samples came from nine human donors (n =3 per group, PD, MSA, and Control). Methods for sample processing and sequencing are detailed in related publications. Putamen samples gene-count matrices were analyzed using R version 4.0.0/RStudio for CCA-based sample integration and unsupervised clustering using Seurat Package version 4.0.1. Cell clusters corresponding between samples identified by this workflow were assigned broad type annotations based on cluster-level gene feature expression patterns (oligodendrocyte precursor cell (OPC; VCAN), oligodendrocyte (OLIGO; MOG, MBP), neuron (NEUR; RBFOX3, SNAP25, GAD1, GAD2, NRGN), astrocyte (ASTRO; GFAP, AQP4, GJA1), microglia (MICRG; CSF1R), and vascular leptomeningeal cells (VLMC; SLC6A13)) (Reference publication Fig. 2). Differential expression analysis was applied at cluster level for

pairwise comparisons of disease and control sample gene expression using the Seurat FindMarkers() function and MAST [37]. Pathway enrichment analysis is an approach that compares overlaps of sets of differentially expressed genes with genes linked with biological functions which can be applied to identify pathways characteristic of particular cell populations and pathways differentially affected in disease states. For this analysis, pathway analyses were performed for gene sets of interest using Qiagen Ingenuity Pathway Analysis (IPA) software [38] with significance cutoff adjusted p-value<0.05 and abs(log2 fold change) cutoff 0.35 and functional enrichments for cell cluster profiles were queried using the Enrichr platform [39].



REF [1] Figure 2: Putamen sample nuclei integrated and clustered. Plots are UMAP of principle components coloured by cluster identity. Expression levels for type-specific markers are shown in violin plots by cluster.

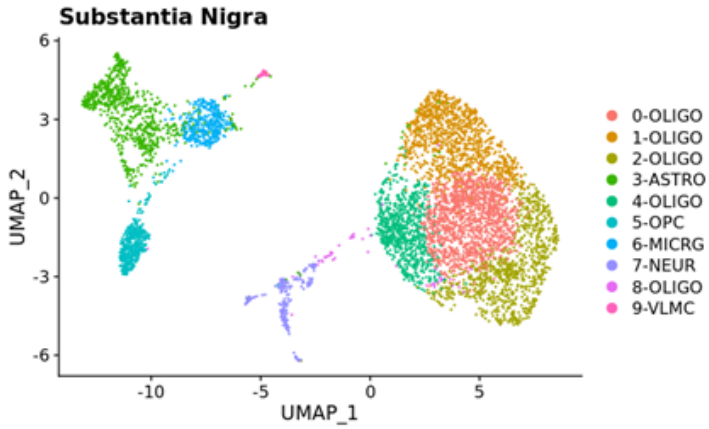


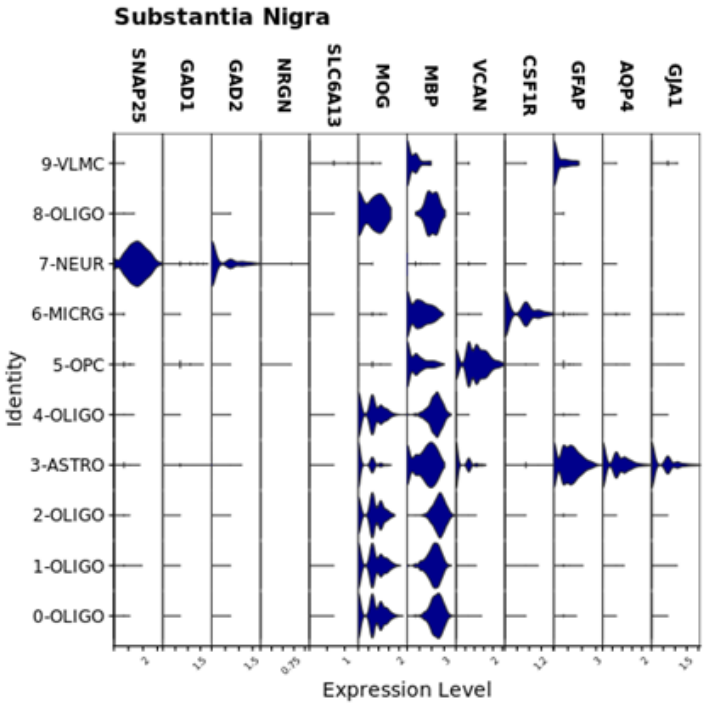
Following CCA-based integration and profiling of putamen brain samples, we then applied the transformation of gene-count matrices derived from this analysis to an independent dataset with the aim of identifying cell populations in the dataset from another brain region that might be linked with disease-specific transcriptional alterations identified from the reference data. Data used for this exploratory analysis comes from work by Agarwal et al. 2020 for substantia nigra samples from 5 human donors (NCBI interface:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140231. For this data, sample integration and cell types annotation was performed as for putamen so that broad cell types assignments could be compared after label transfer. (Reference publication Fig. 2). Label transfer was performed for annotated data in Seurat using the functions FindTransferAnchors and TransferData. Two sets of labels were assigned in substantia nigra data – broad types and numbered cluster identities, which are clusters identified by the application of the Louvain clustering method which capture heterogeneities within broad cell populations.

2.2.6 Experiments

The aims of this analysis were to profile cell states in human putamen, identify disease-related transcriptional changes in these populations, and then to apply label transfer methods to relate these findings to data from another brain region, substantia nigra, which is particularly affected in PD. A first step in comparing the two datasets was to assess broad populations of cells (Publication Figure 3). Table 1 from the reference publication shows these proportions. We see from these that oligodendrocytes are the most common cell type in all samples. Subsequent analyses focus on findings in this cell type.



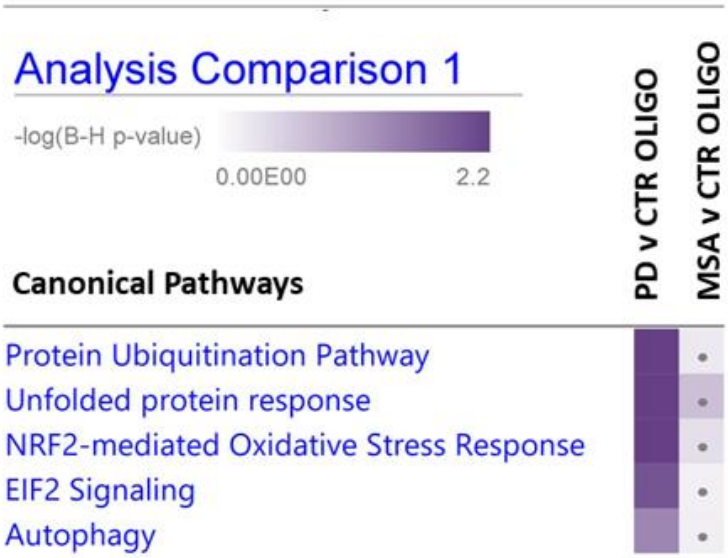


REF [1] Figure 3: Substantia nigra sample nuclei integrated and clustered. Plots are UMAP of principle components colored by cluster identity. Expression levels for type-specific markers are shown in violin plots by cluster.

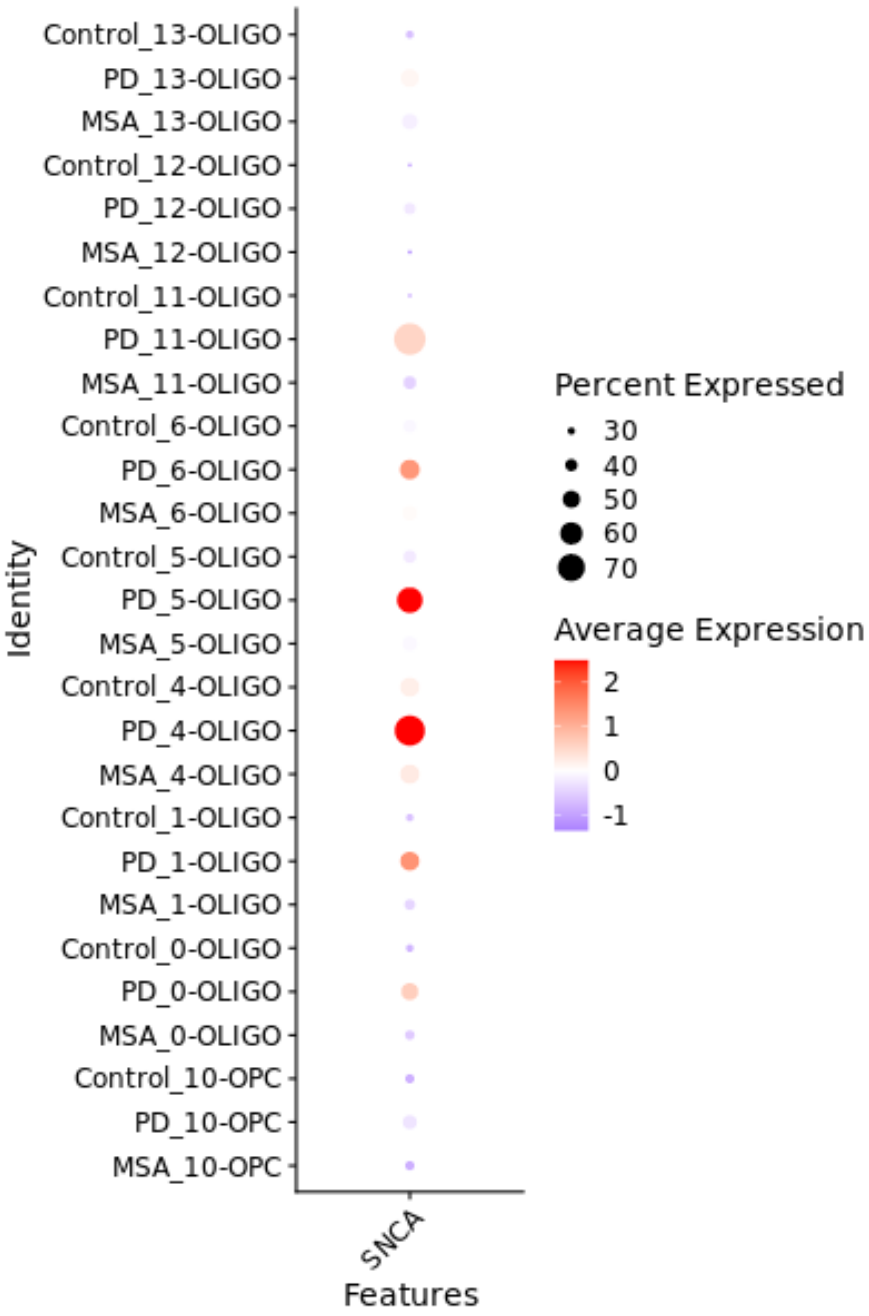
REF [1] Table 1: Broad Cell Types Proportions

Cell Type	Tissue Source - Condition	Mean Proportion ± Standard Deviation
Oligodendrocyte	Putamen - Control	66.5±14.3
	Putamen – PD	64.2±24.5
	Putamen - MSA	64.6±13.2
	Subst. Nigra - Control	63.8±16.9
Neuron	Putamen - Control	14.5±6.6
	Putamen – PD	13.9±14.8
	Putamen - MSA	18.7±12.1
	Subst. Nigra - Control	5.5±5.5
Astrocyte	Putamen - Control	9.4±5.9
	Putamen – PD	12.6±7.4
	Putamen - MSA	7.4±1.4
	Subst. Nigra - Control	16.0±8.6
Microglia	Putamen - Control	4.1±0.8
	Putamen – PD	6.3±2.1
	Putamen - MSA	4.8±1.2
	Subst. Nigra - Control	5.4±3.7
OPC	Putamen - Control	5.1±1.6
	Putamen – PD	2.4±0.3
	Putamen - MSA	3.8±1.5
	Subst. Nigra - Control	8.4±4.5
VLMC	Putamen - Control	0.4±0.4
	Putamen – PD	0.7±0.5
	Putamen - MSA	0.7±0.5
	Subst. Nigra - Control	0.9±0.4

Putamen data includes both healthy and disease state samples. Louvain clustering before annotation identified eight clusters subsequently annotated as oligodendrocyte type. Differential gene expression comparisons were performed for all oligodendrocytes and for oligodendrocyte subclusters. Notable results of these analyses include identification in PD of prominent differences in unfolded protein response and stress signaling in comparison to Control which are not observed in MSA (Reference publication Fig. 4) as well as pronounced increases in SNCA gene expression oligodendrocyte subclusters number 4-OLIGO and 5-OLIGO particular to PD (Reference publication Fig. 5).



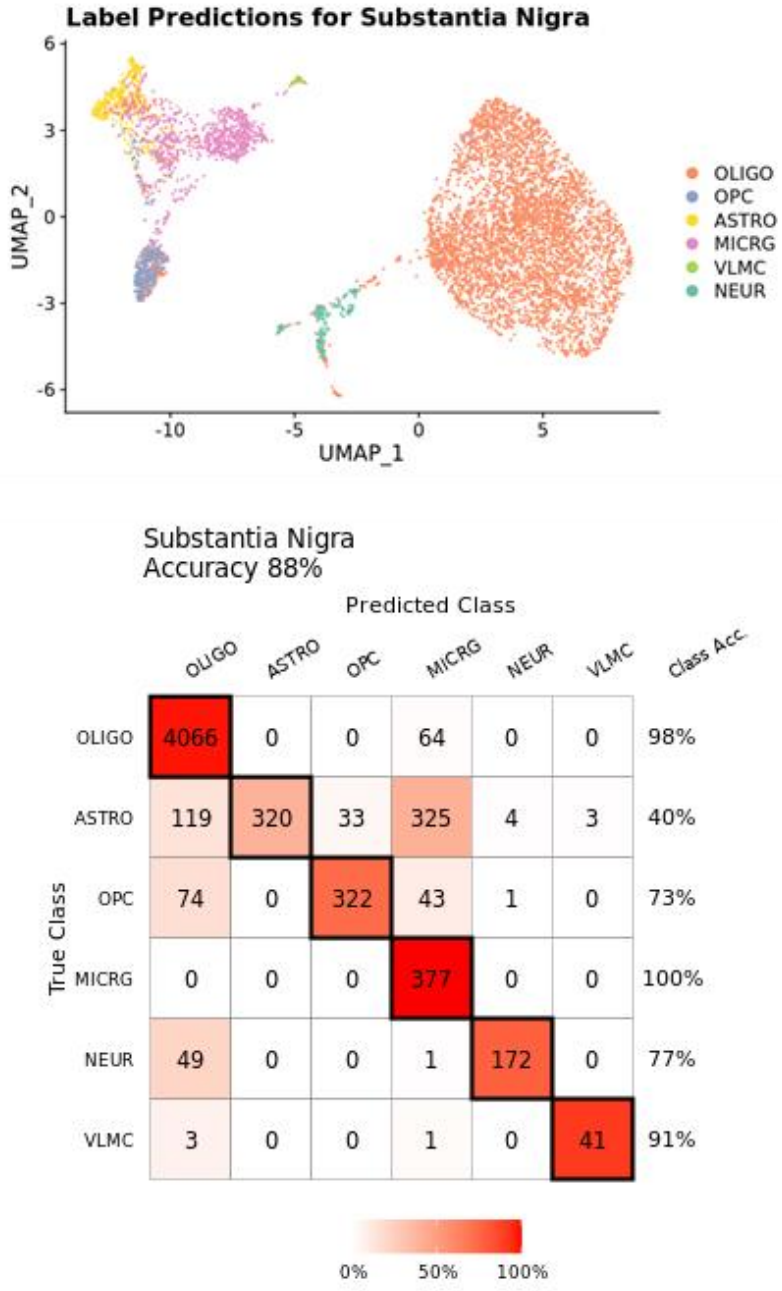
REF [1] Figure 4: Pathway enrichments for oligodendrocyte nuclei differentially expressed genes. (grey dot: p-adj>0.05)



REF [1] Figure 5: Comparative proportions and average expression of SNCA in oligodendrocyte lineage clusters.

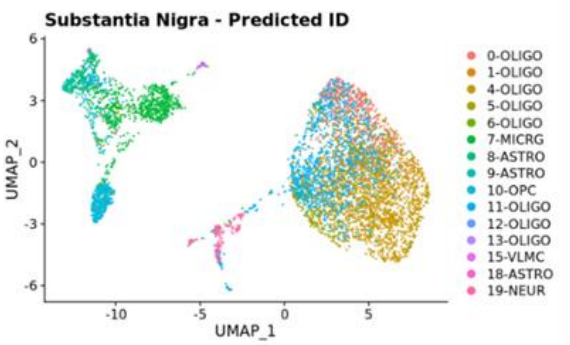
Transfer of cell type annotations to substantia nigra was then performed to determine how oligodendrocyte population heterogeneity might relate to results from putamen.Label transfer performance was evaluated at two levels – first, accuracy of broad types annotations was assessed by comparing transferred broad types labels with broad types labels already assigned to substantia nigra data. As can be seen from the confusion matrix in reference publication Fig. 6,

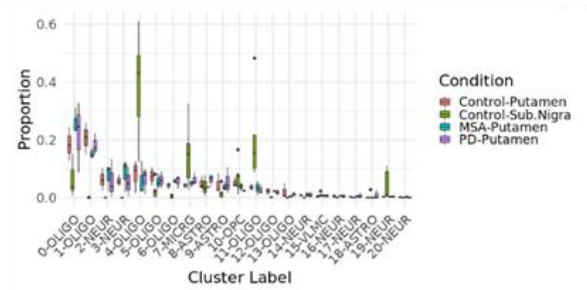
classification accuracy for oligodendrocytes was highly concordant, with 98% accuracy. Interestingly, some other cell types had lower classification accuracies – contributing factors in such inaccuracies include both biological and analytical method aspects. Here, though we focus on oligodendrocytes.

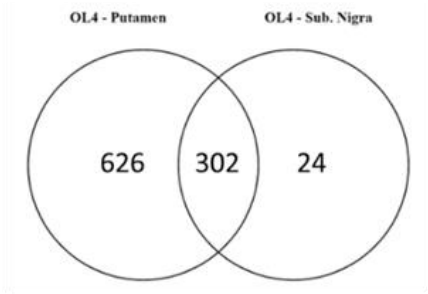


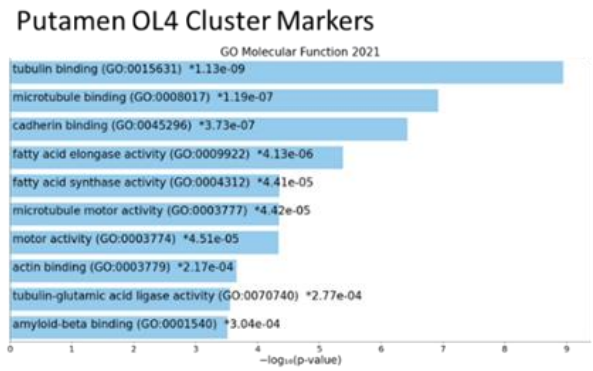
REF [1] Figure 6: Prediction of broad cell types from putamen reference and confusion matrix with class accuracies.

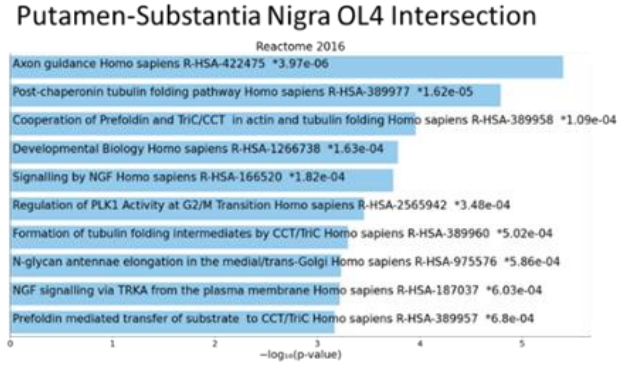
We then applied label transfer from putamen reference data to substantia nigra to predict oligodendrocyte subcluster population membership in substantia nigra data. Interestingly, in substantia nigra oligodendrocytes, 4-OLIGO types comprises a greater than expected proportion of all oligodendrocytes (Reference publication Fig. 7). Classification accuracies were imperfect among different cell types in broad classification, but assessment of classification accuracy in this situation is more challenging because population heterogeneity lacks a ground truth for this prediction. To compare the populations then, we compared sets of genes which are more highly expressed in this cluster versus others within the respective datasets – this comparison confirms a high degree of similarity of the clusters as seen in the significant overlap between marker gene sets for OL4 in putamen and OL4-predicted in substantia nigra (Reference publication Figure 7).The interpretation of these gene sets is further enhanced by matching them with cellular functions as is done in the functional enrichment analysis shown in reference publication Fig. 7.











REF [1] Figure 7: Predictions shown in UMAP project and predicted nuclei population proportions. Overlap of markers for 4-OLIGO cluster and functional enrichments.

2.2.7 Conclusions

This work is included to show the successful alternative application of CCA for identification of corresponding cells states as a preprocessing step in single-cell RNAseq analysis. Understanding the derivation and relatedness of these cell states between datasets guides this successful application. Results and scientific contributions of the work are summarized here:

1) In this study, through application and evaluation of label transfer methods, we successfully generalize a comparison analysis of PD and MSA disease versus control putamen region single cell data to another brain region particularly affected in PD.
2) This analysis newly identifies an expanded subpopulation of oligodendrocytes observed to overexpress SNCA in putamen in PD. From this result, it remains to be further understood how functional activities in oligodendrocyte subpopulations relate to α-synuclein biology and synucleinopathy disease processes.

While oligodendroglial inclusions of α-synuclein protein, the product of expression of gene SNCA, are reported as prominent neuropathology findings in MSA and in PD, neuronal α-synuclein protein aggregations are prominent; varying degrees of neuronal and oligodendroglial involvement are reported in both disorders [40-42]. Multiple studies have linked SNCA mutations and SNCA gene duplications and triplications with familial PD [43-45]. Studies have also found associations between genetic variants within the SNCA locus with MSA [46, 47]. Yet the connections between SNCA, its functions in CNS cell populations, and the pathobiology of PD, MSA, and other synucleinopathies remains incompletely understood. This work provides new insight into SNCA expression patterns by linking disease-specific changes in oligodendrocytes in PD with regional cellular population heterogeneity. In the third section, further work with single-cell RNAseq data will be presented. In the third section, cell-gene tables will be used for network

analysis of gene expression patterns. Preceding this, the second section discuss tasks in evidence aggregation for gene-disease relatedness which demonstrate the usefulness of transforming biological datasets such as single cell RNAseq into networks.

3 | Link Prediction in Aggregated Evidence Networks Using Local Information and Integration of Related Information Networks by Collaborative Filtering

3.1 A Target-Specific Evidence Function for Indication Expansion Queries in the Open Targets Platform

3.1.1 Background

Open Targets is a public-private research initiative that began with the formation of the Centre for Therapeutic Tagret Validation (CTTV), a collaboration between GSK, the Wellcome Sanger Trust, and the European Bioinformatics Institute (EMBL-EBI) (www.opentargets.org). CTTV was renamed to the Open Targets Initiative in 2016 and the aim of work by this organization is to advance the development of methods for exploring and integrating large volumes of scientific data for the support of target validation analyses. Open Targets makes available both a web-based search platform and an application-programming interface (API) where systematically aggregated association evidence may be searched and/or downloaded. Since the creation of CTTV and its renaming as the Open Targets Initiative, a number of pharmaceutical companies have joined this collaboration to provide input on platform design and use, including Sanofi, Biogen, and Takeda [13-15]. The foundation of searches in Open Targets is its defined ‘target’ entity, which is a protein, protein complex, or RNA molecule. Targets in Open Targets are named by their official Human Gene Nomenclature (HGNC) gene name and annotation by ENSEMBL stable ID is also recorded. The use of these standardized identifiers facilitates aggregation of information from diverse scientific sources and integration of Open Targets summary information with other data. Disease

terms in Open Targets are standardized through use of the Experimental Factor Ontology (EFO) classification system where hierarchical relationships are mapped between disease terms, for instance stroke and myocardial ischemia are each subtypes of vascular disease as well as subtypes of neurological and cardiac disease respectively [13-15]. Table 1 from the reference publication shows the types of evidence aggregation scores compiled in Open Targets for Target-Disease pairs. Evidence aggregation in Open Targets at the time of this first work included data for 6,752,528 target-disease associations. A challenge which has remained is how to effectively use this systematically aggregated information.

REF [3] Table 1. Association evidence scores in Open Targets

Association Type	***Evidence Sources***
Genetic Association	Genomics England PanelApp, ClinVar (EVA), PheWAS, Gene2Phenotype, Genomics England PanelApp, Open Targets Genetics Portal, Uniprot, ClinGen
Somatic Mutation	Cancer Gene Census, ClinVar somatic (EVA), IntOGen
Pathways & Systems Biology	Reactome, Sysbio, SLAPenrich, PROGENy, Project Score
RNA Expression	Expression Atlas
Text Mining	EuropePMC
Animal Models	PhenoDigm: mouse-human similarity score
Known Drug	ChEMBL: Bin score by trials phase: [(0); (I); (II); (III); (IV)]

3.1.2 State of the Art

Drug repositioning describes work aimed at identifying greater numbers of therapeutic uses for compounds outside of an initially identified set of indications [48]. Ongoing challenges for repositioning analyses include a need for more complete and nuanced integrated data resources and further conceptualized and validated scientific approaches for high volume screening pipelines [49,

50]. Association scores published on the Open Targets platform provide evidence summary information which is generated by Open Targets analysts through the systematic processing of published scientific information. Open Targets can be queried by EFO disease term to obtain targets with non-zero association scores of the reported types and target entities can be queried to return disease terms with non-zero associations in one of the seven categories as shown in Table 1. Association scores range between zero to 1 and can be sorted to screen for types and relative strength of evidence connecting target and disease pairs. An important consideration in the comparison and interpretation of these association scores is that their distribution reflects state of knowledge on these relationships, not absolute truth. Furthermore, there exists considerable overlap between scores for target-disease pairs with known drugs and those which do not have known drugs. And these association scores are further complicated by the practice in Open Targets of propagating association scores up through the EFO hierarchy, so that direct and indirect associations are included.

These considerations underly the ongoing challenge of using the Open Targets platform to identify true opportunities for indication expansion from aggregated evidence where score magnitude and type do not suffice for this purpose when taken for consideration on their own. Previous work examining the use of Open Targets for drug repositioning searches includes application of association score cutoffs to estimate potential indications by therapeutic area [51] and a ligand - receptor composite association score analysis for g- protein-coupled receptor targets [52]. Machine learning strategies have also been trialed for prediction of target therapeutic status [53]: in this work by Ferrero et al., input data for this prediction task was table with one row per target and input features were computed as a pan-disease term score which was defined as the mean of association scores across disease terms, excluding literature and known drug associations which achieved an

AUC of 0.76 for their best-performing model. However, this model predicts only target indication status, not indications which is the prediction which is useful for drug repurposing. This analysis approaches that task, examining how association evidence may be used to discriminate between target-disease pairing with and without a known drug.

3.1.3 Problem Definition

In drug repurposing, a compound and its target(s) are known, but not all diseases which might have clinical benefit are known. The features available from Open Targets for making disease predictions are different types of aggregated association evidence scores. This is a positive and unlabeled problem – we have labels for known target-indication pairs for a given compound and it is not possible to evaluate all other indications to know the true negatives but we would like to predict potential positives from our available information. The magnitude of association evidence cannot be assumed to linearly relate to the likelihood of a target-disease pair having a known drug because association evidence scores reflect the state of knowledge and caudal relationships and information such as participation of targets in disease pathobiology is not encoded in this resource.

3.1.4 Challenges

The nonlinearity of association scores, missing information from unstudied target-disease relationships, and absence of biological information thus presents a main challenge when using Open Targets association evidence for searches to identify new indications. In this project, we engage with these challenges by hypothesizing that patterns in association evidence types might more effectively identify target-specific drug indications.

3.1.5 Proposed Method

The approach taken in this work is to leverage patterns in association evidence to identify disease terms with similar types of association evidence to predict new druggable target-disease pairs. We compare performance of several core machine learning methods, benchmarking model performance against a harmonic sum overall summary score of all available evidence (excluding known drug status). It is also informative to understand if any specific type of association evidence is most predictive of druggable target-disease relationships. Such an understanding could guide heuristic use of association score evidence, so we include in our work examination of feature importance scores for trained target-specific models.

To generate input data for this analysis, starting from the list of targets included in the Open Targets platform, the Open Targets API was used to obtain tables of disease terms for each target and labels for whether a given target-disease pair had a known drug. Known drug status was used as the target label and the other association scores were used as input features for model training. Prediction performance was compared for four different models: logistic regression, decision tree, random forest, and xgboost methods. The benchmark overall score used for prediction was computed as the harmonic sum of all association scores, which is a method used for summarization in Open Targets. Targets eligible for inclusion in this analysis were those listed in the "20.09_target_list.csv" reference file from Open Targets, with targets used for predictive models required to have a minimum of 1 target-disease pair with a known drug and at least 15 non-zero target-disease associations without a known drug (n=1220 targets). For training and validation, since target-disease pairs with known drugs were less common than those without, target-disease-drug data rows were resampled in 1:1 ratio to generate target data tables.

3.1.6 Experiments

Python version 3.6 and R/Rstudio was used for Open Targets API queries [54], data preprocessing, and model training and evaluation. Classification models and feature importance scores analyses use Python sklearn (logistic regression, decision tree, and random forest)) [30] and xgboost libraries [55]. Data tables were split into training (2/3) and test (1/3) sets. Model training used 3-fold cross-validation. Best-performing models identified from validation were applied to the held-out test data to compute test AUC. Feature importance scores for targets, methods, and association evidence types were then recorded for the best-performing models. Harmonic sum association scores used for benchmarking performance were calculated according to the following formula (with association scores sorted in descending order):

$$s_1 + \frac{s_2}{2^2} + \frac{s_3}{3^2} + \cdots + \frac{s_i}{i^2} \qquad (1)$$

This is the formular used in Open Targets for Overall score calculation. In this work, an overall score is recalculated excluding known drug association as this is the variable to be predicted so the overall score used for benchmark analyses is comprised of association scores for literature, RNA expression, genetic association, somatic mutation, animal model, and affected pathway evidence.

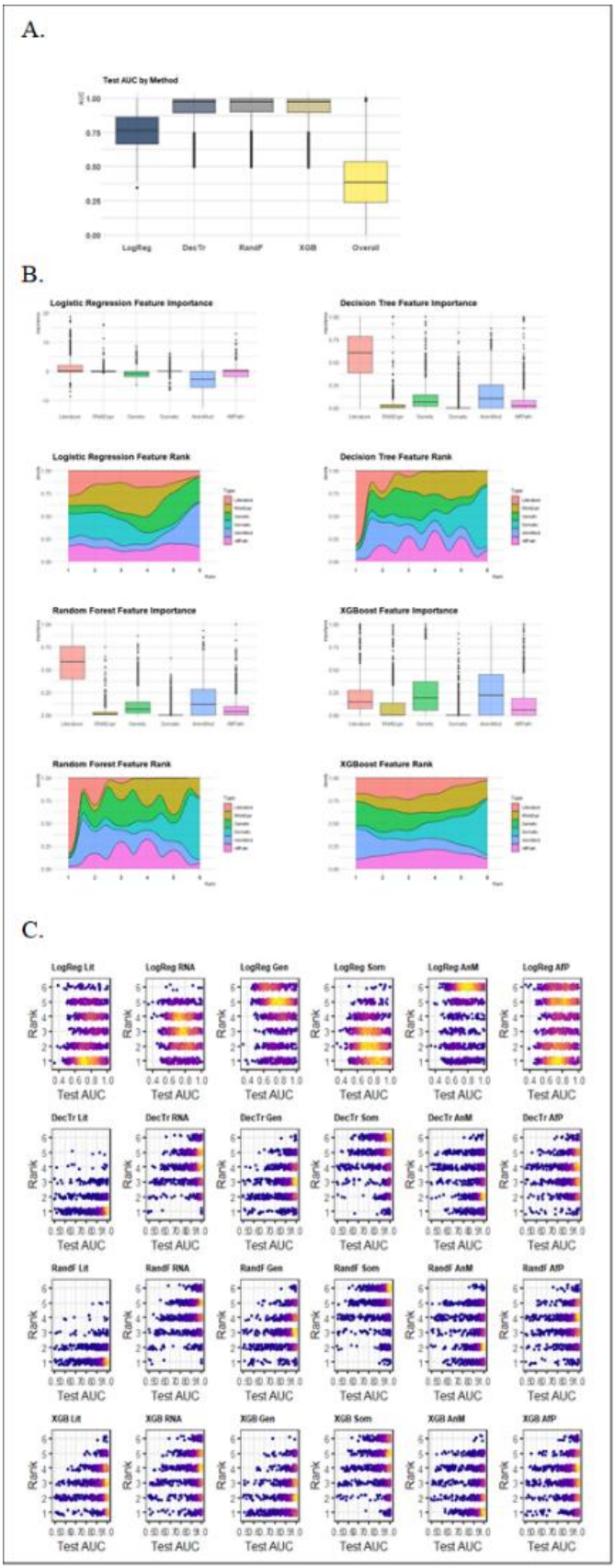
A. Target Prediction Performance and Important Features

Mean test AUC (± Standard Deviation) for all targets was compared for overall harmonic sum (0.401 ± 0.221), logistic regression (0.762 ± 0.132), decision tree (0.920 ± 0.109), random forest (0.923 ± 0.110), and xgboost (0.922 ± 0.110) (Fig. 1A from referenced publication). Not unexpectedly, nonlinear models with more flexible decision boundaries are found to have better performance versus linear regression and harmonic sum. Feature importance scores for these models also offer an interesting result. As can be seen in Fig. 2B from the reference publication, feature weight and rank comparisons reveal that considerable heterogeneity among targets exists

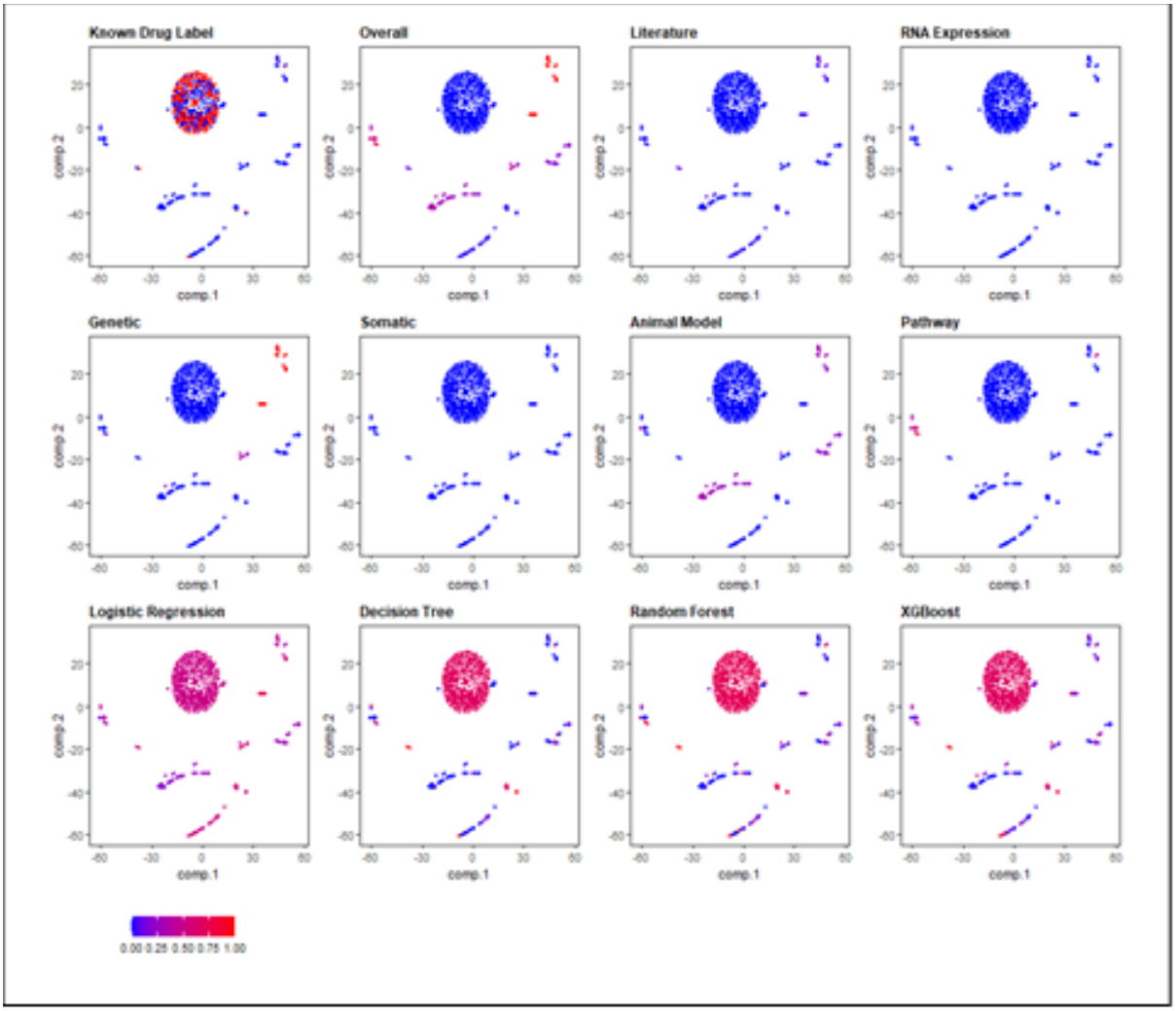
for the feature types which are scored as most important for target therapeutic status prediction. A related question then is whether models which highly weight certain types of evidence are better-performing – from the density plots in Fig. 1C, we observe that high-performing models of all core types weight different types of evidence among targets.

B. Indication Expansion Filtering by Evidence Function

The proposed application of this work is to use the trained models to make predictions on potential novel target-indication pairs using association evidence. Best-performing models of each type were therefore applied to the set of all target-disease pairs with and without known drugs. This procedure yields an exploratory result, which can be visualized by t-SNE as shown in Figure 2 from the reference publication which shows this plot for target SCN9A in which each point represents a target-disease association colored either by known drug status (top left) association evidence type (middle and upper row right), or target evidence function output score (bottom row). Evidence function score were found to highlight novel target-indication pairs not readily identifiable form single association score evidence.



REF [3] Fig. 1. Test set AUC by model (a). Feature importance and rank distributions by evidence type (b). Association evidence rank plotted by model AUC performance. Color indicates relative density by model and association evidence rank: low (blue) to high (yellow) (c).



REF [3] Fig. 2. t-SNE plots for association evidence from target-disease pairs linked with target 'SCN9A'. 'Label' plot shows known_drug status (top row, left; red for existing known drug for that target-disease pair). Color scaled from blue (scores closer to 0) to red (scores closer to 1). Note that trained evidence functions (bottom row) assign greater scores to known drug pairs and a number of candidate indications.

3.1.7 Conclusions

The results and scientific contributions of this work can be summarized as follows:

1) This work presents a novel method for predicting target-disease therapeutic status from association evidence patterns which outperforms (by AUC) previously published prediction methods.

2) Our approach can supplement existing Open Targets platform association evidence score comparisons.

3) We show that heterogeneity exists among targets with respect to which types of association evidence are most important for distinguishing among target-disease pairs with and without drug indications. The import implication of this result is that heuristic reliance of one type of association evidence over others does not identify all target-disease pairs with a known drug.

Further work is needed to make integrated use of Open Targets aggregated evidence combine with other data resources. Given the complex and multidimensional interactions which occur among targets across different cell types, cell states, tissues, developmental stages, and numerous other conditions, network-based methods for modelling these relationships become an attractive solution.

3.2 Empowering the Discovery of Novel Target-Disease Associations via Machine Learning Approaches in the Open Targets Platform

3.2.1 Background

As detailed in the preceding section, Open Targets (https://www.targetvalidation.org/) is a public-private research partnership which provides platform and API access to systematically aggregated evidence resources linking target and disease pairs [15]. Evidence is summarized using association scores, which range between 0 – 1 and are reported for genetic, somatic mutation, pathway biology, transcriptomics, text mining, animal model, and known drug status. An overall score is also computed as the harmonic sum of ordered association evidence [14]. Evidence aggregation in Open Targets is ongoing, with consolidated evidence provided in the 21.04 release for more than 11 million target-disease pairs (https://blog.opentargets.org/next-gen-platform-released/). The purpose of this organization is to provide processed aggregated data for data mining and algorithmic exploration by academic and industry scientists [13]. As shown in the preceding dissertation work, association evidence features can be used to discriminate between target-disease pairs with and without known drug at the target level and such trained models have the potential to identify EFO terms with similar association evidence patterns which may be opportunities for drug repurposing. As noted in the previous work, a limitation of the current Open Targets platform is that target-disease association scores reflect only the state of current knowledge and discovery work would benefit from connecting Open Targets information with orthogonal data resources as detailed in this section.

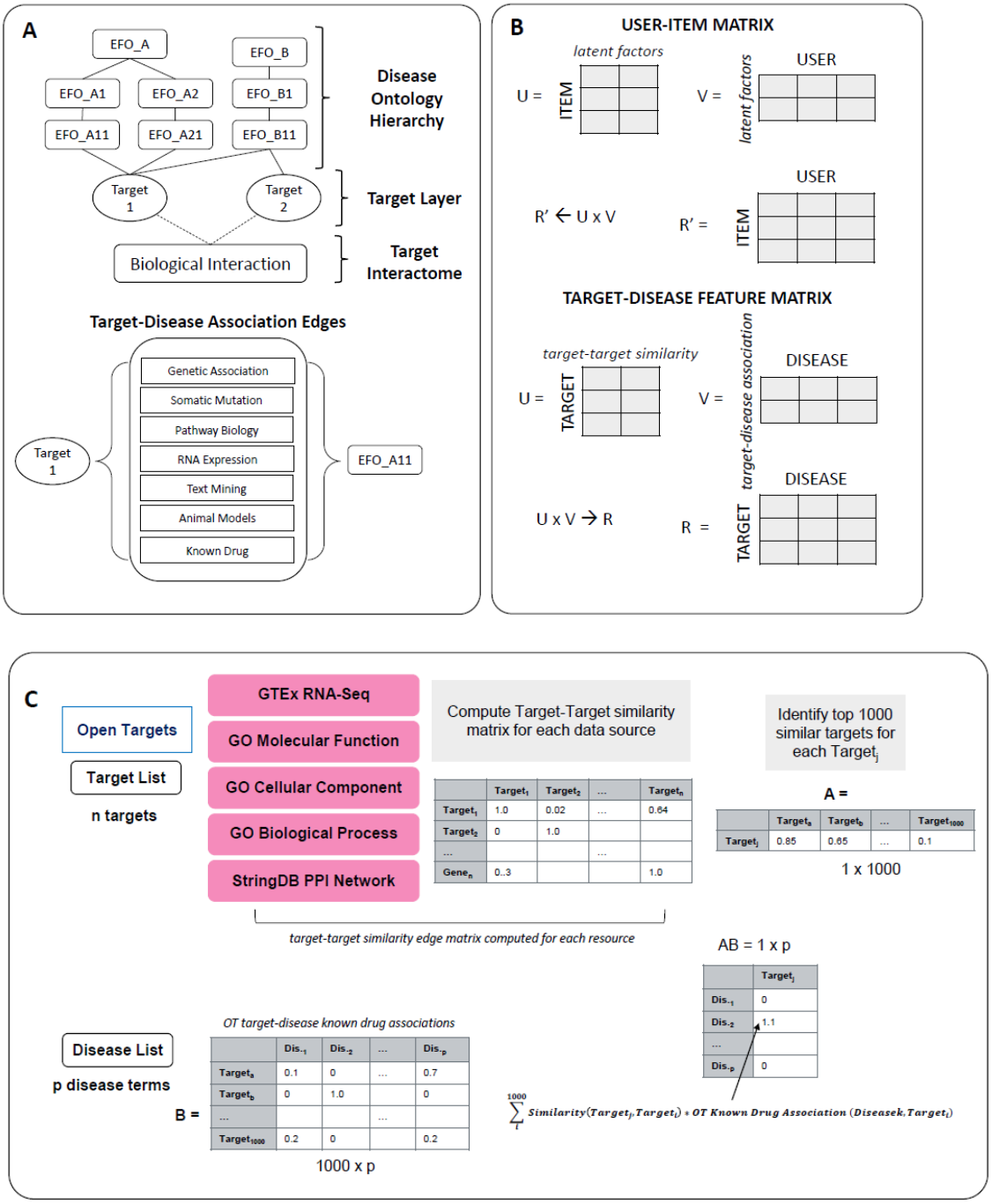
3.2.2 State of the Art

Previous work using Open Targets association scores to predict known drug status is mentioned in the previous section. These previously described works focus on the use of Open Targets association evidence to predict whether a particular target-disease pairing has or might have a suitable drug. Yet the persistent need that remains is for integration of Open Targets platform associations which summarize current states of knowledge and biological information resources

which can be searched to enhance discovery workflows where the task is not focused on successful identification of known druggable target-disease relationships but rather identification of potential druggable target-disease relationships.

3.2.3 Problem Definition

In previous work, association scores used as predictors comprise a small feature set [53] and success at the task of identifying whether a target-disease pair has a known drug is of limited further use beyond the prediction task. The more useful output is whether a target-disease pair might be suited for drug targeting. Furthermore, a wide selection of biological data resources can be considered to expand the features used for prediction tasks using Open Targets data, but this requires development of an integration strategy, since we would want to use target-target relationships for a target-disease prediction task. For the problem definition in this work, we consider Open Targets as an information network with EFO term and target nodes linked by association score edges. Thinking about Open Targets data in this way, connectivity among EFO terms is extensive through disease term hierarchies. But targets lack target-target biological interaction edges. Fig. 1A from the reference publication depicts this network interpretation of Open Targets. The first aspect of this problem is to generate target-target relatedness. The next is use such graph information to generate target-disease relatedness features from biological target-target information sources. We can understand the value of such newly generated features as encoding undiscovered relationships generated out of biological data [56, 57]. The motivating hypothesis of this work is that newly generated features encoding unaggregated biological relationships could boost prediction performance for target disease drug status prediction by machine learning models and provide enhanced insights for discovery applications.



REF [2] Figure 1. Overview of Open Targets data and generation of newly computed features. Open Targets association evidence network edge weights are annotated for evidence from multiple sources (**A**). Novel target-disease association features generated from target-target similarity and target-disease matrices compared with factors used in calculation of a user-item matrix (**B**). Target-disease arrays are generated for each information source and association evidence for known drug status (**C**).

3.2.4 Challenges

The two challenges to be addressed in this work are first, identification of relevant high quality and comprehensive target-target relationship score data resources and second, the conversion of target-target information into target-disease relationship features. Target-target relationships can take several potentially relevant forms: expression of targets in the same tissue location (where their biological activities may interact and influence one another), participation of targets in the same pathway or process (where their biological effects may be interacting), and/or targets may have known protein-protein physical interactions (e.g ligand receptor or other pairings). Public data resources used to quantify target-target relationships for this work were Genotype-Tissue Expression (GTEx) [58]; Gene Ontologies on Molecular Function (MF), Cellular Component (CC) and Biological Process (BP) annotations [59, 60]. Semantic similarity between Gene Ontology (GO) terms is a method in bioinformatics research to study gene functional similarities [61]. Protein-protein interactions information was less directly available as functional interactions among targets were available in protein-protein interaction (PPI) network form from the STRING database (version11) [62]. Protein-protein networks were therefore embedded for this work using the Node2Vec algorithm [63], and from these embeddings, target-target relatedness may be computed as angular distance. One further challenge not specifically addressed in this work is consideration of how to derive and validate novel target-target information resources – this question will be explored in the third and concluding of the dissertation. For this project, we use established biological target-target data resources.

3.2.5 Proposed Method

Methods for this work can be broken down into several sequential parts: target-target edge scoring, generation of new target-disease scores from target-target scores and Open Targets

associations, machine learning model training and evaluation for known drug status prediction using open Targets association scores and newly generated features, and validation of novel prediction quality.

Target-Target edge scoring: As noted above, multiple data resources were selected for use to quantify target-target relationships: Genotype-Tissue Expression (GTEx) which captures expression of pairs of genes in the same tissues [58]; Gene Ontologies on Molecular Function (MF) , Cellular Component (CC) and Biological Process (BP) annotations [59, 60]. Target-Target associations scored by these resources indicate co-occurrence of targets in the same tissue (Gtex) or semantic similarity (GO MF, CC, and BP). Calculation of semantic similarity of targets from GO resources was developed as a method to study gene similarities [61]. GO ontologies are non-overlapping and organize biological domain knowledge with respect to three areas: MF ontologies describe molecular-level activities such as transports or catalysis performed by single gene products or functional complexes; CC ontologies describe cellular structure locations, and BP ontologies refer to larger processes involving coordinated activities of multiple gene products (http://geneontology.org/docs/ontology-documentation/). Protein-protein interactions information for targets is available in protein-protein interaction (PPI) network form from the STRING database (version11) [62] and embedded for this work using the well-established Node2Vec algorithm [63].

Generation of target-disease scores from target-target data resource scores and Open Targets Association evidence: To combine target-target information with target-disease information from Open Targets, a collaborative filtering approach is used [64]. To understand how this is calculated to make logical sense, it is helpful to think of how in a recommender system, the product of an item-factor matrix and a user-factor matrix is a user-item matrix is a user item

matrix (Fig. 1B from reference publication). Factors are quantities which have variability among and meaningful relationships with items and users. Latent factors are unmeasured and can be learned from data, but in this case, we define factors as relationships with genes from the specified data resources between genes-genes and between diseases and genes for Open Targets known drug association scores. This process is a feature generation step and is performed for gene-gene associations from each of the referenced sources (GTex, GO MF, GO CC, GO BP , and STRING). However, since these features are both generated from Open Targets target-disease known drug association scores and known drug status is being used as the target for prediction allocation of target-disease pairs for training and testing sets is done before this step so that test set target-disease associations are set to 0 before feature generation so this information is not encoded in a manner that constitutes so-called “data peaking”. Figure 1C from the reference publication shows the matrix multiplications procedure for feature encoding from the data resources used.

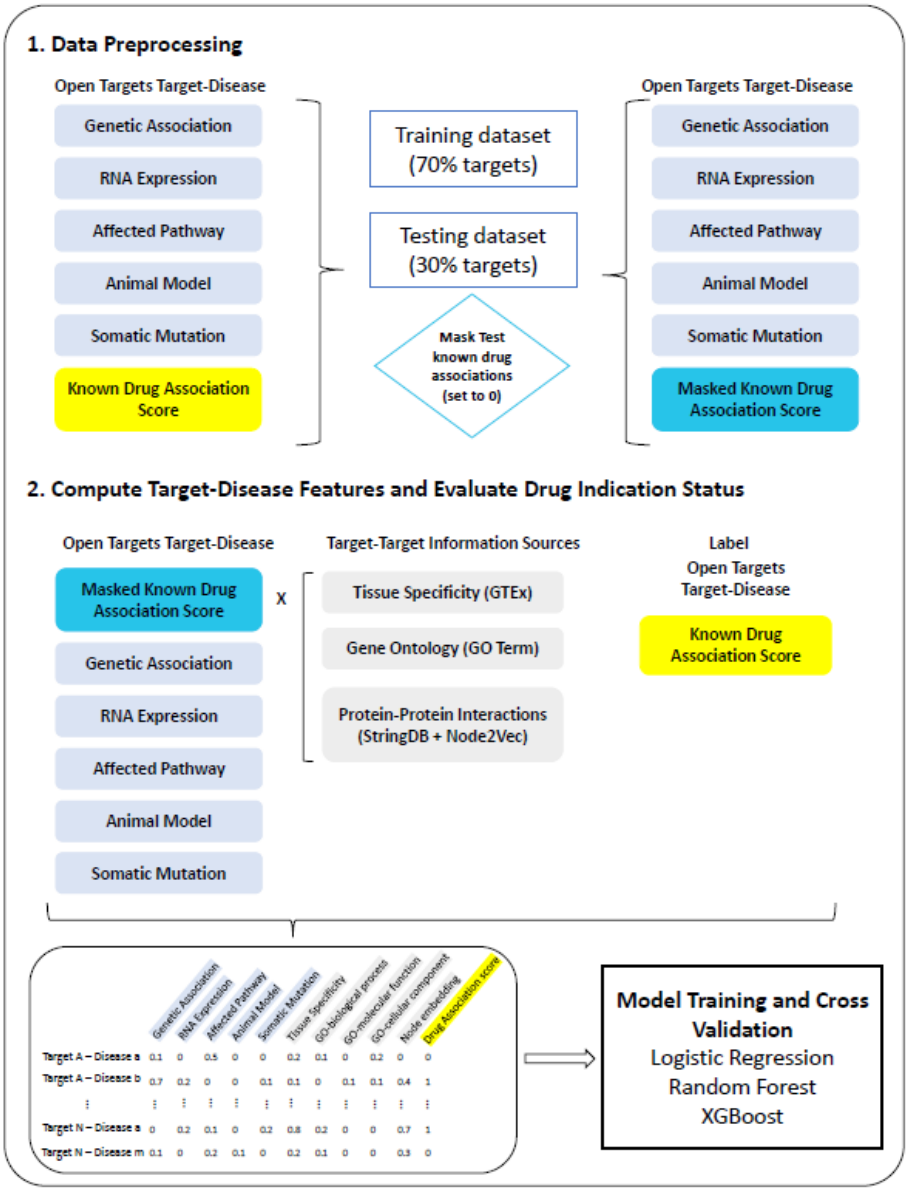
Model training and evaluation for prediction of known drug status from Open Targets association evidence and novel features and prediction quality assessment: Three core machine learning models were trialed for use of Open Targets association evidence only and Open Targets association evidence plus the newly generated features. Trained models were compared by prediction performance and feature importance scores were compared for Open Targets association scores and newly generated features. This was followed by validation of prediction scores for target-disease pairs without known drugs including case studies using external literature information.

3.2.6 Experiments

Data sources access and processing: Association evidence data for target-disease direct associations was downloaded from the Open Targets platform via their API. Disease terms included in downstream workflows were filtered to remove nonspecific terms by removing terms with therapeutic area “measurement”, “phenotype”, “biological process”, and “cell proliferation disorders”. Filtered data then included 1,378,786 target-disease associations for 24,064 unique targets with 990 of these unique targets having at least one indication in clinical trials. For model training and evaluation, 229,228 target-disease pairs were allocated and split into training (159,249 target-disease pairs for 693 unique targets) and testing (69,979 target-disease pairs for 297) sets in a 70%:30% ratio. 23,074 target-disease pairs were held out for validation.

Feature generation: Open Targets association features genetic, somatic mutation, affected pathway, RNA expression, and animal model were used as a benchmark set of predictors. Known indication association was used as the target variable, binarized to label 1 if a target-disease pair has a known drug in clinical trial or approved or label 0 otherwise (Reference Fig. 2). New features were generated using a collaborative filtering-derived approach. For this procedure, first each data source (gene ontologies (MF, CC, BP), GTex, and embedded PPI networks was used to generate a target-target similarity matrix. The similarity matrix was computed from semantic similarity (ontologies), co-expression in tissues (GTex), and embedded networks representing physical interactions of gene products (PPI). From each of the target-target similarity matrices for each data source, for each target, a set of 1000 most similar targets was identified. New features for target-disease similarity were then computed as the product of this array of target-target similarities and the matrix of Open Targets known drug association scores for these 1000 most similar targets. This procedure yielded a new set of target-disease

associations for the target whose 1000 most similar targets were used in the calculation. Note that since known-drug status is the target for prediction and known drug status is used for feature generation, masking of test cases was required before feature generation; target-disease test cases with known drugs were masked by assignment of association score value 0 before feature generation. Fig.2 from the reference publication provides a schematic overview of the process.



REF [2] Figure 2. Workflow schematic for feature generation and therapeutic status prediction evaluation.

Model training and prediction performance evaluation

Machine learning models logistic regression, random forest (RF) [65] and XGBoost [55] were trained using 5-fold cross-validation. Models were trained to predict whether a target-disease pair had a known drug or not. We compared performance of models trained using OT features only with models trained using OT plus newly generated features. In 5-fold cross-validation,

The best-performing model based on validation set AUPR (area under precision-recall) was found to be XGBoost using OT plus computed features (validation set AUPR=0.73 and test set AUPR 0.69). AUPR is used to compare model performance for this task where we evaluate performance in an unbalanced dataset (more negative than positive instances) and are particularly interested in correctly identifying positive instances. OT features-only models perform especially poorly by AUPR – this suggests that newly computed features introduce important information useful for model fitting in this prediction task.

REF [2] Table 1. Number of data instances used for training and validation after removal of all-zero value rows. Held-out testing data comprised of 46290 instances (7382 positive: 38907 negative).

Set	**Fold1**	**Fold2**	**Fold3**	**Fold4**	**Fold5**
Train					
Positive	15137	14382	15120	14435	14918
Negative	70945	67020	70210	73575	71941
Total	**86082**	**81402**	**85330**	**88010**	**86859**
Validation					
Positive	3369	4085	3404	4098	3561
Negative	18132	20313	18424	15344	16194
Total	**21501**	**24398**	**21828**	**19442**	**19755**

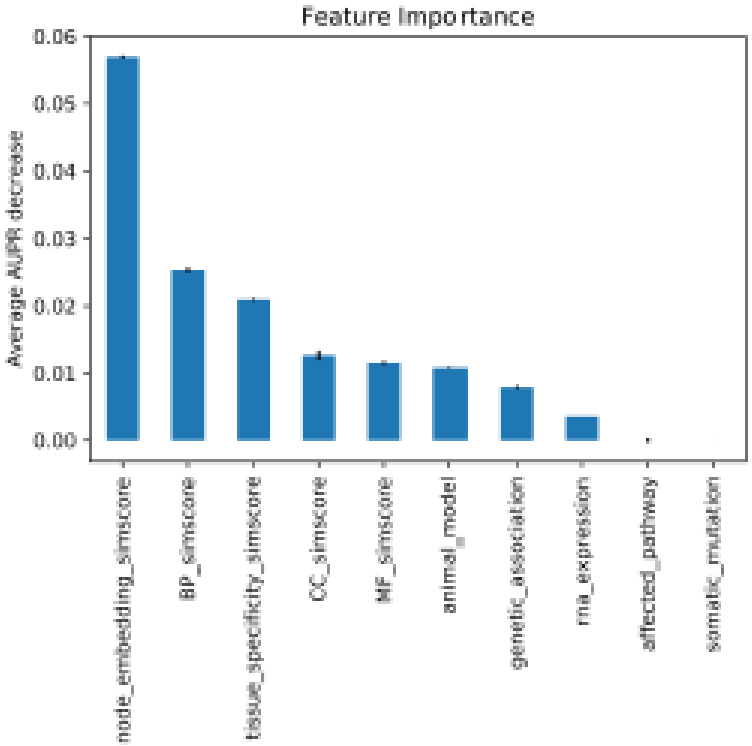
REF [2] Table 2. Known drug status prediction (± standard deviation across 5 folds)

Train Set			
Method	**LogReg**	**RF**	**XGB**
OT Association Evidence			
AUROC	0.7603 (±0.0088)	0.8685 (±0.0075)	0.8784 (±0.0051)
AUPR	0.0685 (±0.0025)	0.2074 (±0.0093)	0.2072 (±0.0103)
Computed Features + OT Association Evidence			
AUROC	0.8867 (±0.0027)	0.9262 (±0.0019)	0.9406 (±0.0018)
AUPR	0.6442 (±0.0069)	0.7500 (±0.0070)	0.7969 (±0.0065)

Validation Set			
Method	**LogReg**	**RF**	**XGB**
OT Association Evidence			
AUROC	0.7625 (±0.0357)	0.8143 (±0.0314)	0.8076 (±0.0335)
AUPR	0.0707 (±0.0102)	0.0872 (±0.0118)	0.0888 (±0.0177)
Computed Features + OT Association Evidence			
AUROC	0.8864 (±0.0076)	0.9103 (±0.0140)	0.9137 (±0.0142)
AUPR	0.6452 (±0.0226)	0.7092 (±0.0459)	0.7264 (±0.0457)

Validation of prediction performance and feature importance scoring: Feature importance comparisons are useful to compare relative contributions of different features to model performance. In this work, for the best-performing model, feature importance was assessed by calculating average decrease in AUPR by randomly shuffling variables in model training. By this method, an important feature is one which decreases AUPR. Feature importance scores obtained by this method identify features generated from embedded protein-protein interaction networks

as most important and all computed features are scored more highly than OT association features for this task.



REF [2] Figure 4B. Feature importance scores indicate the feature types we generated strongly predict known drug therapeutic status.

3.2.7 Conclusions

This work builds on and extends the scientific contributions of the previous work in this section:

1) This work presents a second novel method for predicting target-disease therapeutic status from association evidence patterns which outperforms the performance of models trained with OT platform association evidence features.

2) Our approach can supplement existing Open Targets platform association evidence score comparisons and can be readily adapted for integration of other network information sources.

3) We show that integration of computed features based on functional network interactions achieves improved prediction performance, highlighting the utility of supplementary biological knowledge representation when using OT association evidence.

These results motivate the next section of the dissertation which examines methods for the generation of biological network information from single-cell RNA seq data which can be another source of information on gene-gene relationships within cell states.

4 | Derivation of Biological Information Networks in Validation and Discovery in Single-Cell RNAseq

As a broad introduction, networks are representations of information where relationships among nodes/entities are shown with edge connections. A set of nodes connected by edges is called a graph, and the study of graphs in mathematics has a long history, with the first recognized theorem of graph theory dating back to Leonard Euler's solution to the Konigsburg bridge problem in 1736 [16]. In biomedical science, network methods are increasingly applied for the study of gene expression datasets where large gene regulatory network (GRN) models can be readily derived from high-throughput gene expression datasets which produce gene expression measures for multiple samples and/or cells [17]. As background on gene expression datasets for readers with diverse scientific backgrounds, genes are sequences of nucleotides in DNA and RNA. Polymeric DNA stores living organisms' genome, which is the complete DNA sequence carrying instructions for development, growth, and function of cells – this is stored in cells as chromosomes. RNA is transcribed from DNA, and RNA transcripts have a number of functions essential to cell survival and proliferation, including not only acting as code templates for protein synthesis which is the most well-recognized function of RNA but also functioning in gene expression regulation and other activities. Substantial variation exists in what DNA is actively being transcribed into RNA at a given time in different cells and tissues, and factors such as cell type, tissue, development stage of the organism, growth, disease states, and external and internal stimulation all influence gene expression patterns. High throughput sequencing of RNA generates data tables where transcripts aligned to genes yield counts of gene transcripts over

sample or cell. These matrices can be analyzed to understand which genes are underactive transcription and what differences in transcription may be present between two or more comparison conditions. Humans are estimated to have about 20,000 genes [66], so typically next generation sequencing data tables have more features than instances/samples.

This chapter focuses on the application of network analysis methods for the study of single-nucleus RNA sequencing (snRNAseq) datasets. Sequencing methods for snRNAseq generate for analysis a data matrix which has counts of gene expression transcripts at the level of individual cells (e.g. each gene is a feature and each cell can be considered a unit of observation). From a matrix of cells with varying patterns of gene expression, it then becomes possible to generate a gene-gene co-expression network in which genes share an edge if they are expressed in the same cells and no edge if they are not. One commonly used procedure for construction of a gene-gene network from an expression datasets is to calculate pairwise correlations between each two cells from their cell vectors. Correspondingly, cell-cell networks based on gene expression patterns can also be generated from such matrices. Once such network representations have been produced, these have many potential uses and an array of available methods to study them. For example, a gene-gene network can be studied to understand which sets of genes are transcribed together, and returning to the cell-gene table, one can identify cell subsets where the co-expression of multiple genes can be observed. This approach is suited to identify processes perturbed in disease states (e.g. is a gene co-expression pattern observed more frequently in cells from one state or another). Another use is to cluster related cells in a cell-cell network generated from gene expression data.

4.1 Derivation and Validation of SNCA Region-Specific Gene Networks .

4.1.1 Background

In this work, single nucleus RNAseq data from the Allen Cell Types Database for post-mortem tissue from the human middle temporal gyrus ((MTG, 15,928 nuclei) [67] and Anterior Cingulate Cortex (ACC, 7,258 nuclei) [68] are used to derive a conserved gene-gene co-expression network for the SNCA gene. Expression patterns for this SNCA gene network are then studied for expression patterns among different cell types, and a second validation dataset published by Agarwal et al., 2020 for matched brain cortex (Middle Frontal Gyrus (MFG); 10,706 nuclei) and Substantia Nigra (SN) samples (5943 nuclei) [69] is explored to validate the biological conservation this network across different brain regions.

SNCA is the gene for alpha-synuclein (αSyn) protein, a protein which is characteristically found in nervous system aggregates in Parkinson's Disease (PD), Lewy Body Dementia (LBD), Multiple System Atrophy (MSA), and Pure Autonomics Failure (PAF), disorders collectively termed synucleinopathies [70-74]. Intriguingly, the synucleinopathies differ in which nervous system regions and cell locations are the primary sites of αSyn pathology, and other neurodegenerative disorders may have features of αSyn pathology, as well, for example up to 50% of Alzheimer's Disease patients have evidence of αSyn aggregates in post-mortem studies [70]. Epidemiological studies of SNCA-implicated diseases support polygenetic inheritance, environmental factors, and epigenetics as contributing to lifetime risk, with genetic variants linked with synucleinopathy risk including not only SNCA overexpression and structural modification variants [75], but also gene mutations linked to mitochondrial dysfunction, lysosomal storage disorders, oxidative stress responses, and alterations in potassium channel function [76-78]. Susceptibility genes shown to increase PD susceptibility in particular include GBA, PARK2, PARK7, PINK1, and LRRK2, which connect PD risk with the multiple pathways

[79-82]. The specific processes by which αSyn aggregates form among different cell types, in diverse brain regions, and among different neurodegenerative diseases remain areas of ongoing research and motive the presented analysis.

4.1.2 State of the Art

Weighted gene co-expression network analysis (WGCNA) is a framework developed for biological network generation and data mining which begins with the construction of a Pearson correlation-based gene network followed by downstream exploratory analyses. WGCNA was first described by Horvath and Zhang [83]. By this method, gene coexpression similarity (*sim*) is defined for a given pair of genes *i* and *j* as sim_{ij} using correlation: $sim_{ij} = cor(x_i, x_j)$ where *x* is an expression vector across multiple instances. In the case of single-nucleus or single-cell data, *x* is an expression vector of a gene across cells. Computing sim_{ij} for all gene pairs yields similarity matrix $S = [sim_{ij}]$ and thresholding applied to matrix *S* yields network adjacency matrix $A = [a_{ij}]$. Choices can be made in how to threshold *S* to obtain *A*. For example, one can use a cutoff value for sim_{ij} to dichotomize adjacency as 1 or 0. Alternatively, since the determination of an optimal threshold for a given analysis is non-obvious and can impact results, soft thresholding is the approach used in the WGCNA framework. By soft thresholding, a power function is applied to for thresholding rather than a cutoff value: $a_{ij} = (sim_{ij})^{\beta}$ where β is a parameter selected as the smallest value achieving approximate scale free topology based on the scale free topology criterion plot (insert from CRAN doc ref) for a given dataset. Selecting higher values of β creates greater separation in the transformation of similarity to adjacency values.

Once calculated, adjacency matrix *A* is the relatedness network structure for genes within the input dataset. From this, clustering can be applied to identify groups of genes expressed together

within samples (cells in the case of single nucleus or single cell data). In WGCNA, gene clustering is performed using network proximity, where $gene_p$ and $gene_q$ are defined to have greater proximity if they are more interconnected. The topological overlap measure (TOM) [56] is applied in WGCNA to quantify interconnectedness and is calculated as the overlap of adjacency neighborhoods for pairs of genes including all *m*-step neighbors with a value normalized to fall between 0-1. Larger values of *m* have the impact of including larger neighborhoods in the quantification of overlap. The selection of *m* is made empirically based on resulting cluster sizes. The TOM matrix contains the calculated similarity of each gene pair. In WGCNA, 1-TOM is then generated and is termed the dissimilarity matrix (dissTOM). In the dissimilarity matrix, higher numbers indicate greater dissimilarity. For clustering gene sets into modules (groups of co-expressed genes), WGCNA uses average linked hierarchical clustering applied to the dissimilarity matrix which in the formulation of this framework was observed to lead to more distinct gene modules [83]. In R, clustering of the dissimilarity matrix ix performed using function flashClust(dissTOM, method= “average”) which takes as input a dissimilarity structure [84] and outputs a clustering tree where modules of co-expressed genes are identified by selecting a cut height for the clustering tree branches. Average linkage clustering calculates the distance between two clusters as the average distance between objects from a first cluster and objects from a second cluster. In biological systems, smaller distinctive function gene sets are often related through larger and less cohesive functional systems which provides further intuition for this formulation [85, 86]. In the following work, we apply the WGCNA framework to profile SNCA biology in several single cell datasets, with attention to the reproducibility and biological relevance of these data-derived networks.

4.1.3 Problem Definition

WGCNA requires parameter tuning during analysis, and the quality and scientific meaning of the outputs of these analyses generally requires further steps for interpretation and understanding. This work is an application of the WGCNA framework to human brain single cell datasets to identify genes co-expressed with SNCA in different brain regions. This analysis is followed by profiling of gene expression patterns by cell types in the datasets used, identification of conserved co-expression patterns across brain regions, and validation of the biological significance of the identified SNCA coexpression gene set (module) with respect to functional annotations of protein-protein interaction networks generated using module genes and identification of genes linked with genetic risk for Parkinson's Disease among genes in the identified networks.

4.1.4 Challenges

Genes identified by WGCNA as being in the same co-expression cluster (module) with SNCA could be correlated with one another by random chance rather than having a true functional interaction. Therefore, a substantial remaining challenge when applying this method is to validate analysis outputs relative to the domain-specific question for which the work is undertaken.

In this work, we aim to better understand expression of SNCA and its participation and connection to different cell functions by the following additional analyses applied to our identified SNCA module genes:

1) Gene expression data provides only a partial view of cellular activities, since RNA transcription precedes, rather than being concurrent with the presence of protein in cells or enaction of its other functional activities. Thus one challenge in understanding the quality and

significance of clustered gene sets is how they relate to cell functions. To follow up this issue, a conserved network of genes co-expressed with SNCA is combined in this analysis with a reference network of protein-protein interactions followed by identification of functional pathways linked to this larger network.

2) Genes identified as being coexpressed with SNCA are compared with genes linked with known PD risk variants to assess genetic evidence for the relatedness of these expression patterns to synucleinopathy disease risk.

3) Functional experiments were undertaken for selected network genes to assess the effects of modulating their expression on SNCA expression *in vitro*. This experimental work will be included in the manuscript resulting from this analysis but is out of scope for the dissertation. (Figure 8 reporting in vitro validation results is therefore not included for review as part of the dissertation).

4) A major consideration in the generalizability of WGNCA-derived gene networks is how the data used to derive a particular network relates to other datasets and questions. The initial derivation of the set of SNCA co-expressed genes is performed in cortex, where the dominant cell type is neurons. This challenge is addressed in this work through the inclusion of a validation dataset which comes from human midbrain (substantia nigra).

4.1.5 Proposed Method

Data sources: Publicly available single-nucleus RNAseq data from the Allen Cell Types Database [68] was downloaded and processed for human middle temporal gyrus (MTG; 15,928 nuclei) and anterior cingulate cortex (ACC; 7,283 nuclei). ACC and MTG samples come from frozen human brain samples for 8 healthy donors ranging in age from 24-66 years. For the

presented analyses, included genes were filtered to include only protein-coding genes, excluding those on the X and Y chromosomes [87]. Validation of results from ACC and MTG were performed using independent single nucleus RNAseq dataset from matched samples of cortex and substantia nigra (SN) for 5 human donors [69]: (middle frontal gyrus (MFG); 10,706 nuclei) and substantia nigra (SN) samples (5943 nuclei).

SNCA Module Detection Using WGCNA: WGCNA was performed for ACC and MTG data in R using the WGCNA package [83]. For each, MTG ad ACC, the gene-gene co-expression matrix was generated from normalized cell-gene arrays extracted from the Seurat data objects for data from these locations after loading the files. Following the WGCNA framework as described above, a scale free-topology fit index was plotted as a function of potential values for the soft thresholding power, and a soft threshold power of 8 was selected empirically for optimal transformation of each co-expression similarity matrix into a topological overlap matrix (TOM). As discussed, selection of this relatively high value supports greater separation between genes with relatively high and low correlations. The TOM matrix reflects relationships of topological similarity between genes. The dissimilarity matrix (1 – TOM) is used to represent dissimilarity, and as discussed, this dissimilarity matrix is used to cluster groups of genes into co-expression modules based on its better performance for distinct modules when in biological systems there are multiple overlying higher processes in which these coexpression sets participate [83]. Hierarchical clustering and dynamic cutting were then applied to identify modules of co-expressed genes. The dynamic tree cutting algorithm (deep split = 2) was used to detect gene modules (e.g. clusters of densely interconnected genes in the computed co-expression network).

The modules containing the SNCA gene (the cluster of genes co-expressed with SNCA) was then identified for ACC and MTG, respectively as these are of particular interest for this

analysis. ACC and MTG SNCA modules were then overlapped to find common genes between them and to identify a robust module with conserved expression across sample locations for further downstream use. Statistical comparison of module overlaps was performed in R using the Hypergeometric test function phyper(), where q = number of genes overlapping between MTG and ACC; m = gene MTG module size; n = total number of genes (estimated to be 20,000) – MTG genes to find number of non-MTG genes; and k = gene ACC module size.

SNCA Module Protein-Protein Interaction (PPI) Network: As noted above, genes in a WGCNA-derived SNCA co-expression module conserved across MTG and ACC locations could be correlated with SNCA by random chance. To explore functional relationships that might underly coexpression, a protein-protein interaction (PPI) network was then generated from genes in the overlap set of the co-expression modules for ACC and MTG. PPI network generation is performed for a given gene set using reference information on known protein-protein interactions to link genes by their physical interactions [88]. NetworkAnalyst and the STRING database were used to generate and visualize this PPI network and to identify additional interacting proteins connected within the resulting network [88].

Single Nuclei RNA-Seq Analysis of Human MTG and ACC: Single-cell RNA-Seq analyses for MTG and ACC were performed in R using the Seurat package, version 3.0 [35]. A standard data pre-processing workflow was applied using cutoffs 200 < nFeatureRNA < 9500 and percent.mt < 0.01 for MTG and 200 < nFeatureRNA < 8500 and percent.mt < 0.01 for ACC, with cutoffs selected based on initial QC plots. Data were normalized using global-scale log normalization, scaling by a factor of 10,000 for each data set. Identification of highly variable features, linear dimension reduction by PCA transformation, and cell clustering were performed using standard Seurat package workflows with PCA dimensions optimized to achieve separations by cell type

markers (n = 17). Cell type annotations for each cluster were determined by cluster marker visualizations and marker distributions quantified using feature and violin plots. SNCA expression within each cluster was examined by labelling nuclei by expression level of the SNCA gene. Differential SNCA expression was then compared by examining cluster staining on TSNE plots and by using dot plots to compare differences in mean expression by cluster and proportion of cells within each cluster expressing SNCA.

SNCA Module Gene Ontology and Comparison with PD Genome-Wide Association Studies: Genes identified from the SNCA module and PPI interaction network were also compared with a list of genes identified as being nearest genes to single nucleotide polymorphisms (SNPs) linked with significantly increased risk for Parkinson's Disease in genome-wide association studies [78]. We then compared the set of nearest genes with the 197-gene union set of the ACC and MTG modules and the ACC and MTG co-expression module PPI network. Statistical tests for overlap enrichment were performed in R using the hypergeometric test. Overlap visualizations for each of these comparisons were generated using Venny [89]. The Ingenuity Pathway Analysis (IPA) tool (Qiagen) was used to identify networks and processes involving genes for the robust SNCA module obtained from the intersection of the ACC and MTG locations and the intersection of this module with genes represented in the PPI network analysis and by comparison with nearest genes adjacent to PD GWAS loci [38]. The Enrichr online tool was used to perform DisGeNET queries for the overlap sets for the SNCA PPI network and the PD-GWAS nearest genes. Disease enrichment p-values are calculated in Enrichr using the Fisher exact test, which is a test of proportion which models the probability of any gene belonging to any set as a binomial distribution, and with adjusted p-values reflecting deviation from an expected rank based on prior results obtained from multiple trials of random gene sets [90, 91].

Validation of SNCA and Co-expression Module Expression Patterns Using Independent Cortex and Substantia Nigra Single Nucleus RNA-Seq Data: Single-nucleus RNA-Seq integration and cell type annotations were performed for validation MFG and Substantia Nigra data separately in R using the Seurat package, version 3.0. A similar workflow as for ACC and MTG samples was followed. Pre-processing cut-offs were selected based on initial QC plots: 200< nFeatureRNA < 6000 and percent.mt < 5. Data were normalized at the individual sample level and then integrated using the Seurat functions FindIntegrationAnchors and IntegrateData as described in the Seurat data integration workflow with the number of PCs used for clustering (n = 20) chosen to optimize separation between clusters. Broad cell types were assigned for each cluster based on marker expression levels as for ACC and MTG, and SNCA gene expression was similarly compared by FeaturePlot and DotPlot for each region and cell type cluster.

For comparison with the ACC-MTG SNCA module, a SNCA co-expression module was also derived by WGCNA. As for ACC and MTG regional data, a normalized cell-gene array was extracted from the Seurat data object, including only protein coding genes. The plot of the scale-free topology fit index versus potential soft thresholding values was examined to choose a soft threshold power of 2 for transformation of the co-expression matrix and deep split =2 for dynamic tree cutting. Statistical comparison of module overlaps between the ACC-MTG SNCA co-expression module intersection set and the substantia nigra SNCA co-expression module was also performed in R as for the overlap enrichment testing for the ACC and MTG modules.

Genes identified by WGCNA as being part of the SNCA co-expression module conserved across MTG and ACC locations and for the Substantia Nigra could be correlated with SNCA by random chance rather than having a true functional interaction with SNCA. A protein-protein interaction (PPI) network was thus created for genes identified as belonging to the conserved

SNCA co-expression module for ACC-MTG in order to integrate the identified robust module with information on known protein-protein interactions and to determine hub genes/proteins within this network [88]. A second PPI network was also identified for genes identified as belonging to the SNCA co-expression modules for ACC-MTG-SN. NetworkAnalyst and the STRING database were used to generate and visualize this PPI network and to identify additional interacting proteins connected within the network generated for the robust SNCA co-expression module genes [88]. Functional ontology analysis for this shared PPI network was then performed using the Cytoscape ClueGo application [92, 93].

Statistics and Reproducibility: Data used in this study is publicly available; sources are detailed below in section titled 'Data Availability'. Analysis code is available as supplementary files. Statistical methods are presented in each of the above sections in the context of their use and interpretation. Supplemental files provide gene lists for ACC, MTG, SN SNCA co-expression modules as identified by WGCNA as well as PPI networks identified from 197-gene overlap of ACC-MTG modules and 29-gene overlap of ACC-MTG-SN modules.

4.1.6 Experiments

Identification of a conserved SNCA co-expression module for human MTG and ACC regions: Prior to clustering and nuclei type annotations, WGCNA and hierarchical clustering was applied to normalized gene count transcription matrices for all nuclei to identify clusters of genes co-expressed with SNCA for MTG (n = 427 genes) and ACC (n = 333 genes) samples. From these SNCA co-expression clusters for MTG and ACC, we identified a statistically significant intersection set of 197 genes (hypergeometric p-value = 5.3e-247), which consisted of genes in the SNCA-containing co-expression clusters for both MTG and ACC. This intersection set of SNCA-co-expressed genes comprises a robust co-expression module for cortical SNCA (Fig. 1).

The robust and statistically significant overlap between ACC and MTG suggests the conservation of the SNCA co-expression module across different regions of the cortex.

Cell types annotations in MTG and ACC: Nuclei-gene matrices for MTG and ACC samples from the Allen Cell Types Database were filtered and processed using the Seurat package workflow to annotate cell type for nuclei in these gene count transcription matrices (Fig. 2) [35, 94]. For each location (ACC and MTG), nuclei were clustered based on highly variable genes using an unbiased graph-based clustering approach. Cell types for these unbiased clusters were annotated using broad

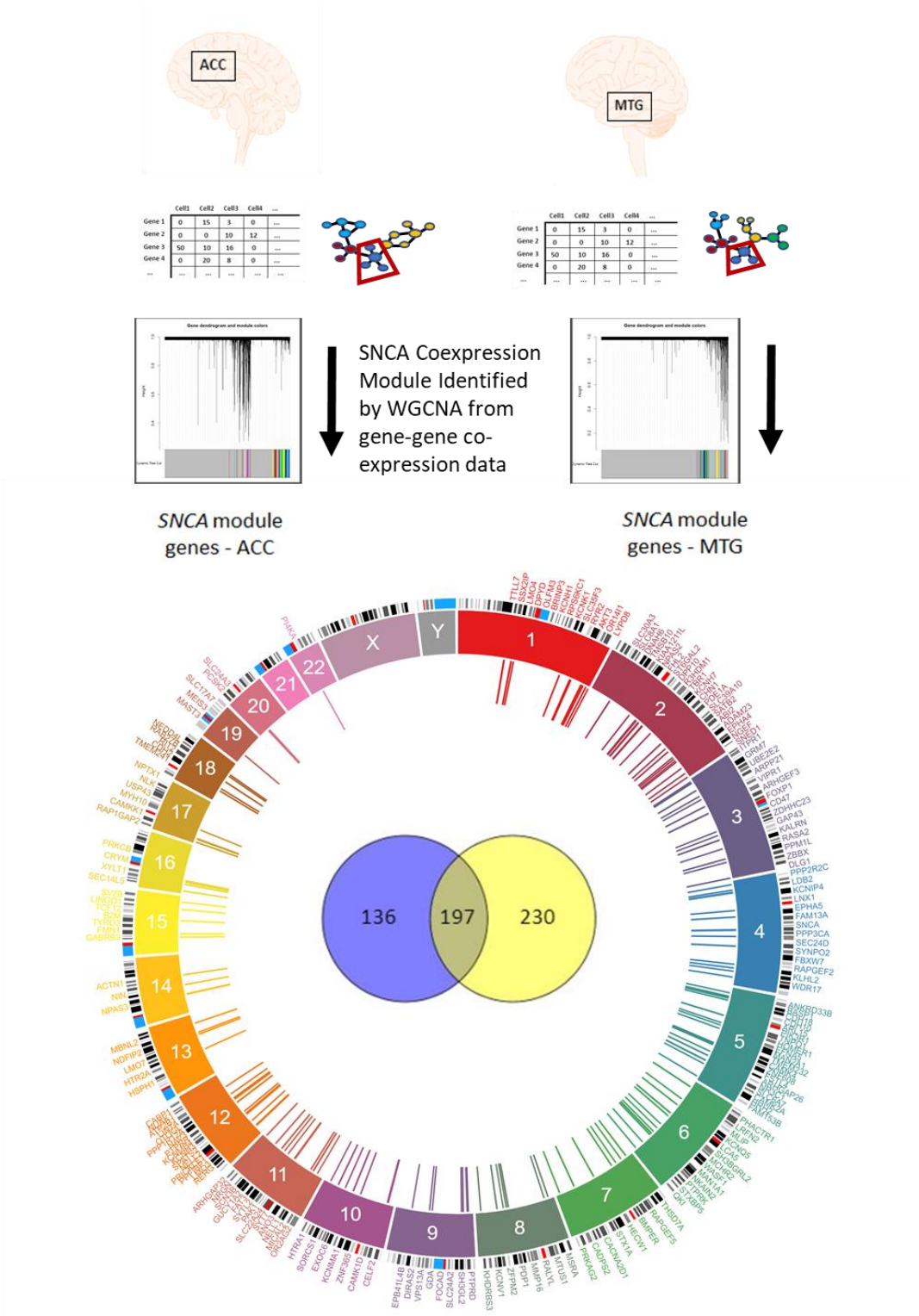


Figure 1. SNCA Co-expression module conserved across sample locations identified by WGCNA hierarchical clustering. Genes belonging to SNCA-containing cluster are those co-expressed with SNCA in nuclei from ACC or MTG samples. SNCA co-expression module 197-gene overlap mapped to human genome.

expression markers for excitatory/glutamergic neurons (GLUT; SLC17A); inhibitory/GABAergic neurons (GABA; GAD2), astrocytes (ASTRO; GFAP), oligodendrocytes (OD; MOG), oligodendrocyte precursor cells (OPC; PDGFRA), and microglia (MG; CSF1R) (Figure 3). These annotations confirmed the presence of all major cell types in both ACC and MTG with a higher percentage of excitatory neurons observed in MTG (67.3% clustered nuclei) versus ACC (55.6% clustered nuclei). Neuronal nuclei comprised the majority of sample nuclei for both ACC and MTG samples (Fig. 3; Table 1).

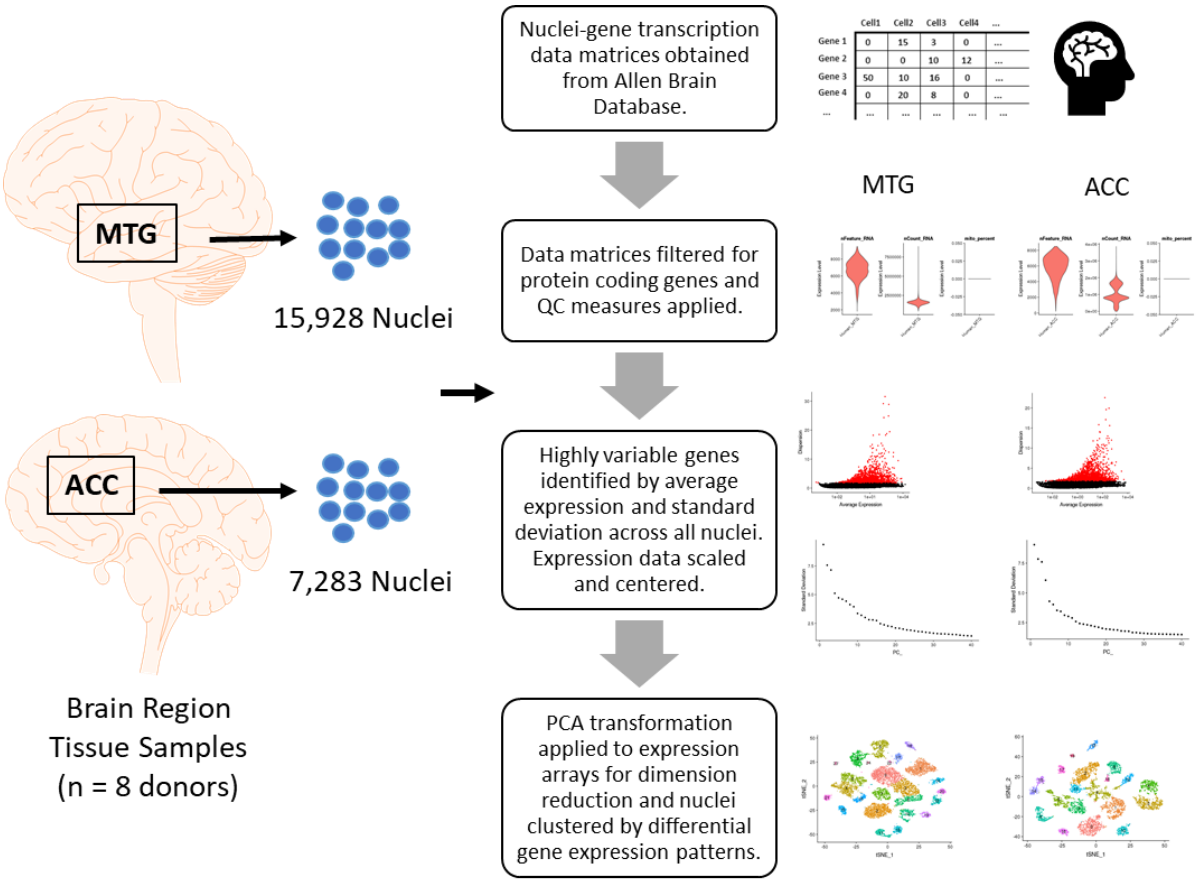


Figure 2. Single-cell RNA-Seq analysis of human MTG and ACC samples. Schematic overview of data preprocessing, quality control (QC), and variable gene and PCA dimension selection workflow.

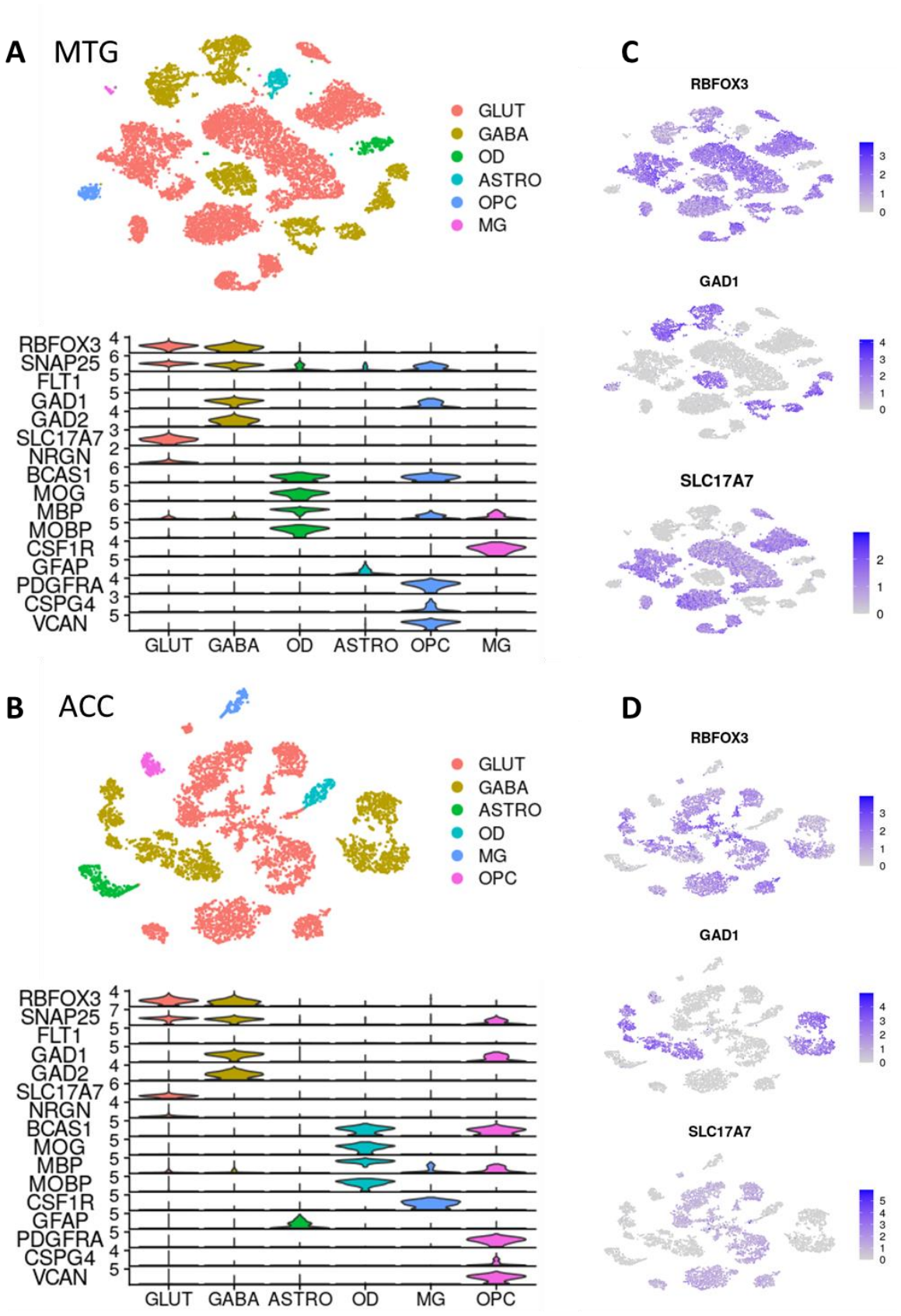


Figure 3. **Cluster annotations using cell type markers** for (**A**) MTG and (**B**) ACC. Feature plots show expression of neuronal cell type markers for (**C**) MTG and (**D**) ACC.

Identification of increased SNCA and SNCA co-expression module gene expression in excitatory (glutamatergic) neurons in ACC and MTG: We then sought whether SNCA and the conserved 197 genes in the co-expression module were differentially expressed by cell type in ACC and MTG. To explore this question, we plotted heatmaps and feature plots from the annotated, normalized single cell data (Fig. 4). Genes in the SNCA co-expression module were observed to be most highly expressed in excitatory neurons in comparison to inhibitory neurons and other cortical cell types for both ACC and MTG samples. Expression levels for genes in the 197-gene co-expression module for the annotated cell type clusters are presented for MTG (Fig. 4A) and ACC (Fig. 4D). Alongside the observed increased expression of SNCA co-expression module genes, greater expression of SNCA itself was also observed in excitatory neurons in comparison to inhibitor neurons and all other non-neuronal cell types (MTG: Fig. 4B, C; ACC: Fig. 4E, F).

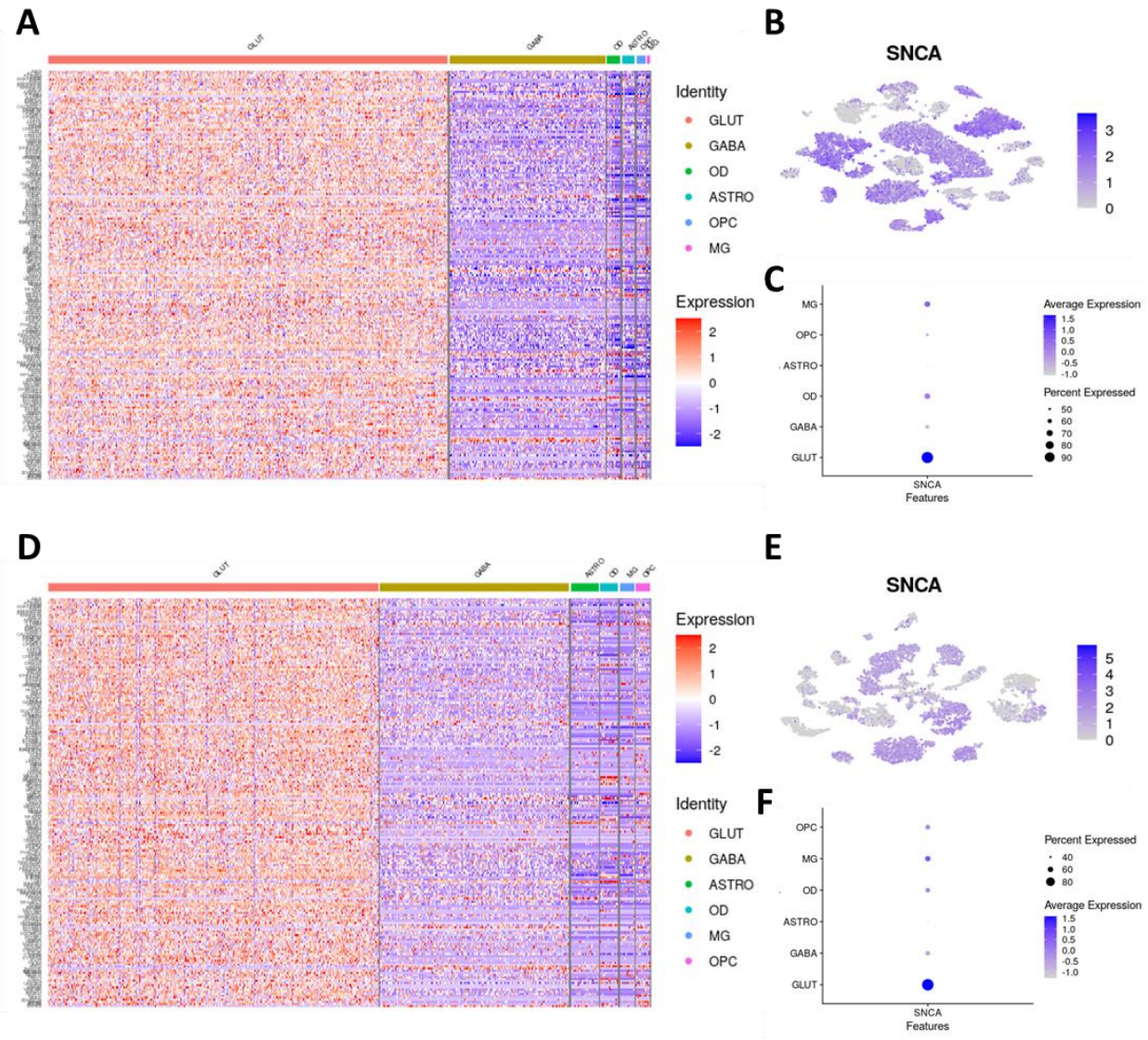
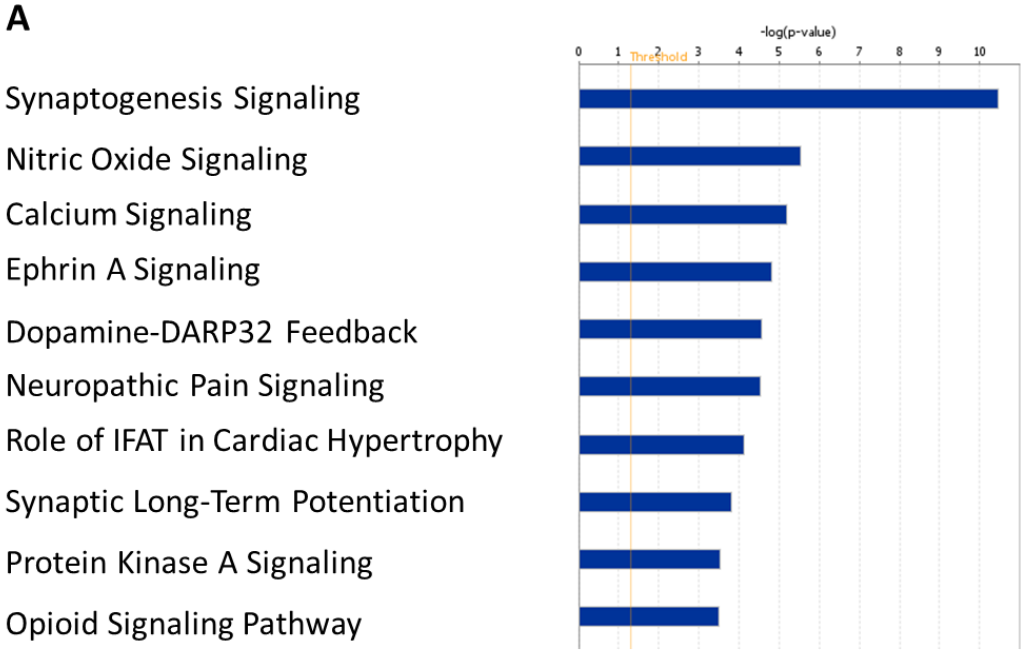


Figure 4. SNCA module gene expression and SNCA expression by cell type for ACC and MTG. Excitatory neurons are observed to have greater expression of SNCA co-expression module genes(MTG-**A**; ACC-**D**) and SNCA (MTG-**B**,**C**; ACC-**E**,**F**) compared with inhibitory neurons and non-neuronal cell types.

Pathway analysis identifying that SNCA co-expression module genes are implicated in synaptic biology and dopamine processing: Synaptogenesis, nitric oxide, calcium and ephrin A signalling, as well as dopamine feedback pathways were found to be the top pathways significantly enriched for among genes co-expressed with SNCA in ACC and MTG when this gene set was analysed using IPA (Fig. 5A). Pathway analysis for the 197-gene conserved SNCA co-expression module also identified statistically significant enrichment for molecular targets of the upstream regulators Levodopa, Histone Deacetylase (HDAC1), cAMP Responsive Element Binding Protein 1 (CREB1), and SNCA, among others (Fig. 5B). Gene ontology (GO) analysis further identifies cellular localization, synaptic transmission, and ion transport as among functions enriched for in the set of 197 conserved SNCA-co-expressed genes (Fig. 5C) [95, 96]. In the network shown in Fig. 5C, nodes correspond to gene sets and edges represent overlap between gene sets; edge width proportionate to the number of overlapping genes. The identified pathways and regulators found to be enriched for among SNCA-co-expressed genes are central to synaptic biology and dopamine processing, supporting a role for αSyn as a multifunctional protein acting at the intersection of multiple cellular pathways.

SNCA is found to be the major hub protein in the substantia nigra-specific PPI network derived from co-expression module genes: To obtain further insight into the functional pathways connecting SNCA and its co-expressed genes, the 197 genes in the conserved co-expression module were then used to generate a protein-protein interaction network using NetworkAnalyst [88], restricting our network model to substantia nigra-specific interactions. Network analysis yielded a network of 1495 proteins connecting the genes in our identified module via known protein-protein interactions (Fig. 6A). Although SNCA was not found to be a top hub in the WGCNA network for either ACC or MTG, SNCA was empirically found to be the top hub of this

A



B

Regulators	Target Molecules in Data Set
Levodopa**	ABI2,ARPP21,CRYM,EFNA5,EPHA5,FSTL4,GABRB3,HECW1,HOMER1,KCNK1,LDB2,MBNL2,NEDD4L,NGEF,NUAK1,PCSK2,PIK3R1,PPM1L,PPP2R2C,PRICKLE1,RERG,RIT2,SH3GL2,SLC24A2,SLC6A7,VPS13A,WDR17
HDAC1**	CACNA2D1,CAMK2A,CHN1,HOMER1,KLHL2,LDB2,PPP3CA,PRKCB,SH3GL2,SLC17A7
CREB1**	AKT3,ARL15,CACNA2D1,CADPS2,CAMK1D,CAMK4,CDH10,CRYM,GDA,HOMER1,HTR2A,NKAIN2,PTPRR,RALYL,RERG,SCN3B,ZDHHC23
SNCA**	CRYM,DPP10,GDA,KCNIP4,LMO7,NPTX1,RIT2,SLC17A7,SNCA,SV2B,TBR1

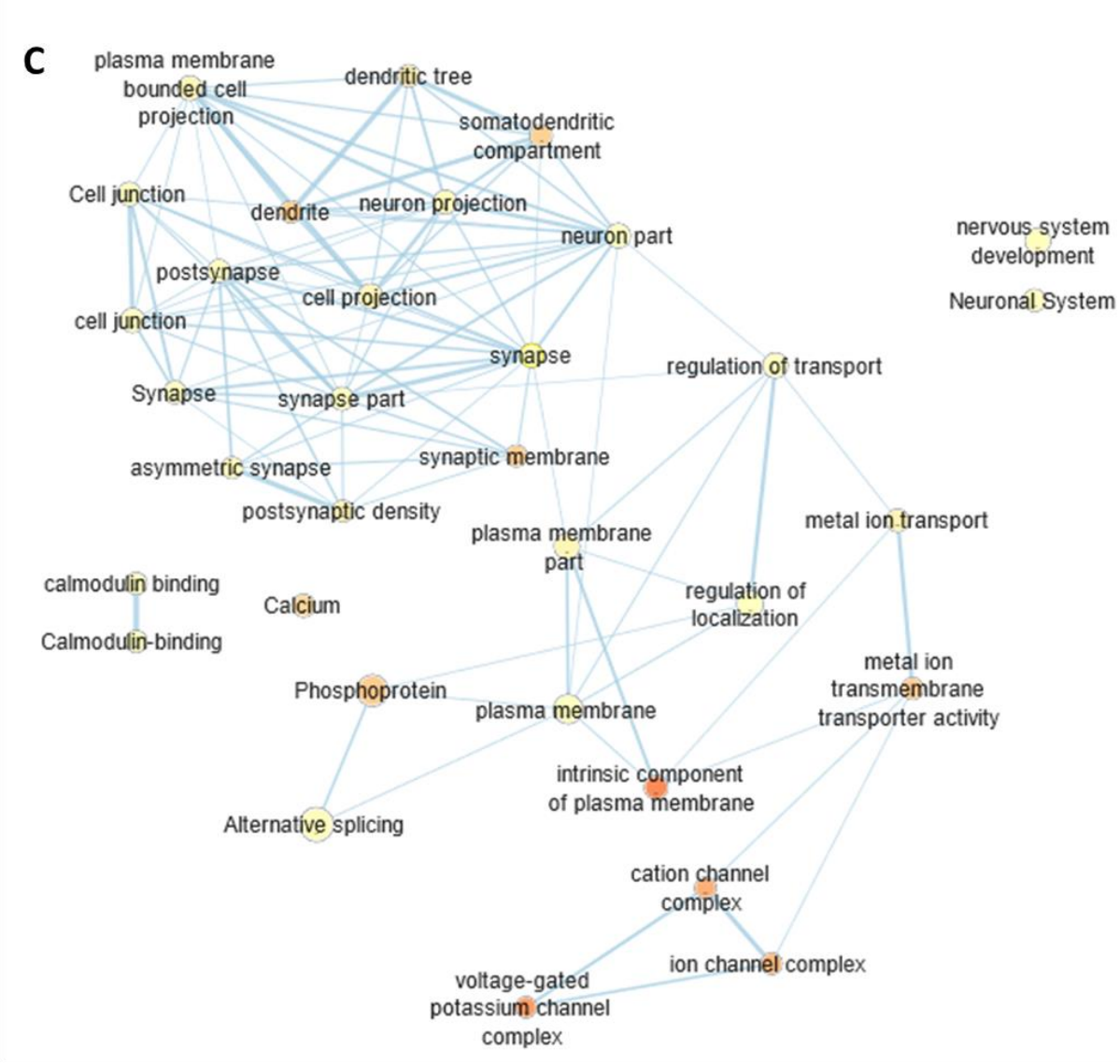
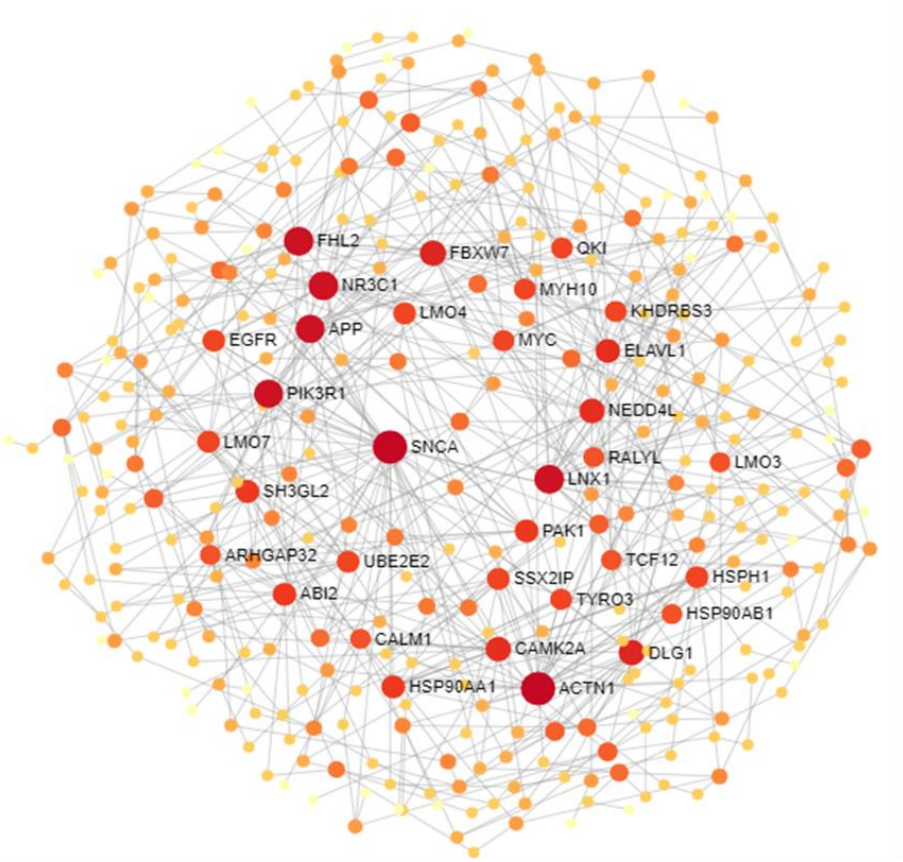


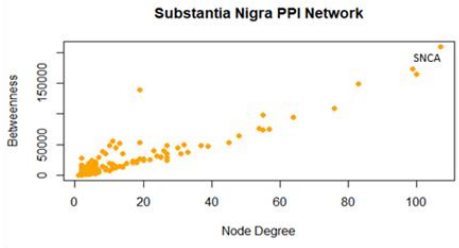
Figure 5. Ingenuity Pathway Analysis for 197-gene conserved SNCA co-expression module for nervous system tissue excluding cancer pathways (**A**) Top canonical pathways and (**B**) Top regulator molecules enriched for among module genes (**overlap $p < 10^{-4}$). (**C**) **EnrichmentMap Analysis** reveals functional enrichment for cellular localization, synaptic transmission, and ion transport ($q < 10^{-5}$; darker red coloration indicates more significant pathway enrichment; nodes = gene sets; edges = overlap between gene sets; edge width = # overlapping genes).

A



B

Top Full PPI Network Hubs		
Gene Symbol	**Degree**	**Betweenness**
SNCA	107	209937.1
FHL2	100	165397.1
LNX1	99	173174.3
ACTN1	83	149510.8
PIK3R1	76	109032.4
NR3C1	64	94574.99
SH3GL2	57	75580.77
NEDD4L	55	98650.93
QKI	55	73363.88
FBXW7	54	75971.5



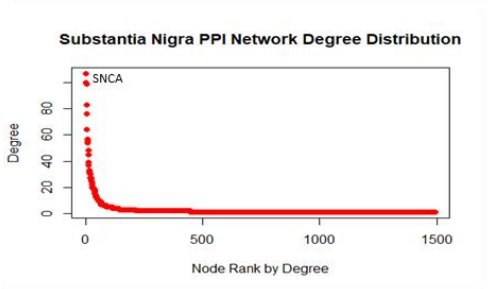
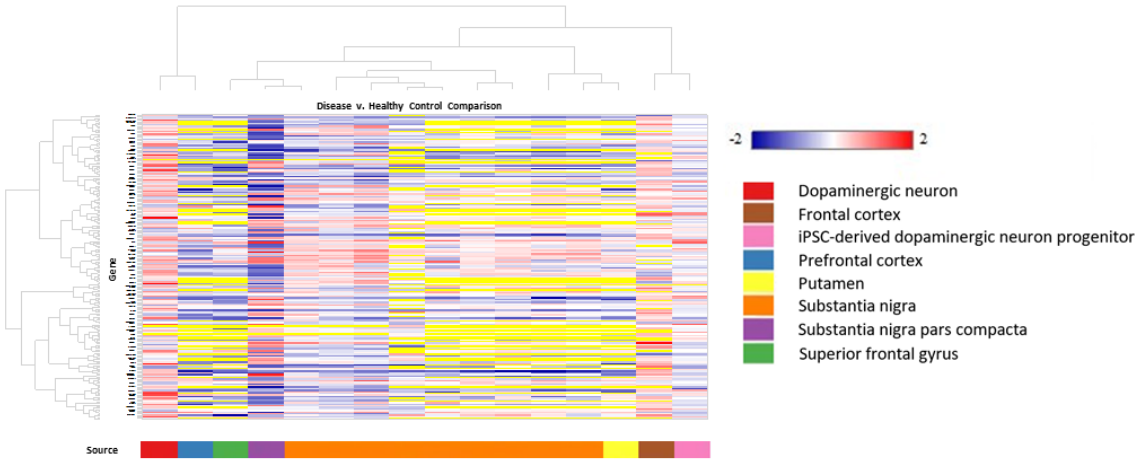


Figure 6. Protein-protein interaction network generated for the genes in the conserved *SNCA* co-expression module. (**A**) Minimally connected protein-protein interaction network for genes co-expressed with *SNCA* is shown with degree represented by node size and color. (**B**) *SNCA* is the top hub in this network by both degree and betweenness measures.

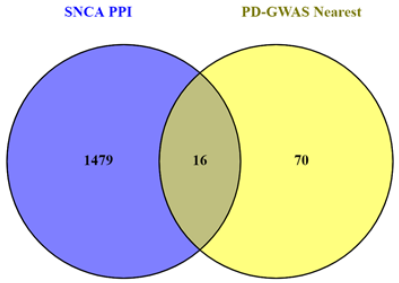
protein-protein interaction network by both degree and betweenness criteria (Fig. 6B), further establishing SNCA as a central participant in interacting neural pathways. Interaction pathways within this PPI network potentially offer opportunities for targeted intervention to modulate SNCA expression or biomarker identification to monitor SNCA expression pathways in the CNS. We also used this network to identify other top hubs (Fig.6B); top PPI network hubs ordered by degree (highest to lowest) are the following: SNCA, FHL2, LNX1, ACTN1, PIK3R1, NR3C1, SH3GL2, NEDD4L, QKI, and FBXW7.

SNCA co-expression module genes are also observed to be differentially expressed in CNS tissues for PD Case versus Control Samples: To further explore whether SNCA and its 197-gene conserved co-expression module might be altered in the synucleinopathy Parkinson's Disease (PD), differential expression of these 197 SNCA co-expression module genes was queried for in archival studies comparing PD case and control samples obtained for human cortex, putamen, substantia nigra, and dopaminergic neuron tissues. Records for this search were extracted from OmicSoft DiseaseLand Database (release HumanDisease_B37 20171220_v7) using an adjusted p-value cutoff of 0.05. Heatmap comparison of these search results reveals numerous significant differences in PD case versus control expression of SNCA co-expression module genes as well as tissue-specific variations in SNCA module gene expression (Fig. 7A). Consistent with our findings in MTG and ACC, frontal cortex samples have a number of genes with significantly increased expression for PD samples versus controls, and in addition, dopaminergic neurons are also found to have increased expression of these genes for PD versus Control samples (Fig. 7A; tissue sample type annotation indicated by x-axis color bar).

A



B



SNCA
SH3GL2
RIT2
LRRK2
FYN
BAG3
DYRK1A
MED13
GCH1
MBNL2
CAMK2D
STK39
DLG2
FCGR2A
GBA
SYT17

PD-GWAS Nearest ∩ SNCA PPI

DisGeNET Disease Name	P-value	Adjusted P-value
Parkinson Disease	1.04e-9	1.02e-5
Parkinsonism, Juvenile	3.00e-8	9.80e-5
Motor symptoms	2.45e-8	1.21e-4
Lewy Body Disease	4.97e-8	1.22e-4
Autosomal Dominant Late Onset Parkinson's Disease	6.90e-8	1.36e-4

PD-GWAS Nearest ∩ SNCA PPI excluding LRRK2 SNCA GBA

DisGeNET Disease Name	P-value	Adjusted P-value
Parkinson Disease	1.53e-6	1.50e-2

Figure 7.SNCA Co-expression module and PPI network genes: (**A**) **Co-expression module genes** are differentially expressed in PD Cases versus Controls in tissue samples from human cortex, putamen, and dopaminergic neuron (p-adjusted cutoff = 0.05 for differential expression in Qiagen DiseaseLand Gene Set Analysis*; yellow indicates gene expression missing values*). (**B**) **PPI network gene** overlaps with Nalls et al. 2019 PD-GWAS loci nearest genes. DisGeNET set enrichment tables show Fisher Exact test p-values and adjusted p-values for top five most significant diseases for the set of all overlapping genes and excluding high risk PD genes GBA, SNCA, and LRRK2.

SNCA protein-protein interactome is enriched for genes associated with PD-risk loci: We then examined which pathways within the present SNCA PPI network might be of particular relevance to synucleinopathy biology, focusing specifically on PD, for which population-level genomic analyses have previously identified a number of risk variants with high statistical significance. Among 86 unique genes mapped as the nearest genes to highly significant variants linked with PD risk in Nalls et. al, 2019, meta-analysis of genome-wide association studies (GWAS) [78], 16 were found in our SNCA PPI network (hypergeometric p = 0.0006), demonstrating significant enrichment of the SNCA PPI network for genes most closely linked with significant PD-GWAS risk loci (Fig. 7B). Genes that not only have protein-protein pathway interactions with SNCA but which are also near to single nucleotide polymorphisms with highly statistically significant population-level genomic associations with PD are particularly interesting genes, as these characteristics together support a potential causal genetic contribution to disease risk. DisGeNET queries for the overlap sets for the SNCA PPI network and the PD-GWAS nearest gene set confirms enrichment for Parkinson's Disease-linked genes, as well as other neurodegenerative disorders (Fig. 7B) [97]. We also repeated DisGeNET queries excluding SNCA, GBA, and LRRK2 to confirm enrichment for PD-linked genes. This identification of a particular set of genes with potential causal associations with PD from within the SNCA PPI network not only provides further validation of our network but also highlights the value of this newly identified network, as each node and edge represent potential targets for measuring or modulating SNCA-related activities relevant to other synucleinopathies, as well.

SNCA and SNCA co-expression module genes have enriched expression in excitatory (glutamatergic) neurons in independent samples from human MFG and in neuronal and oligodendrocyte lineage populations in SN: In MFG samples, SNCA and ACC-MTG SNCA co-

expression module genes were similarly found to have enriched expression in GLUT neurons in contrast to GABA neurons and other glial cell types, providing validation of our experimental findings in an independent cortex data set (Fig. 9A,B). In SN samples, interestingly, ACC-MTG SNCA co-expression module gene expression was observed to be selectively increased in the single cluster which expressed neuronal cell type marker RBFOX3 (Fig. 9C cluster 12) as well as in clusters 3 and 7 which expressed oligodendrocyte lineage markers (Fig. 9C). Dot plot profiling to examine SNCA expression by cluster revealed that while SNCA expression was highest in neuronal nuclei in SN, SNCA was also expressed at low levels in clusters expressing oligodendrocyte lineage markers (Fig. 9D).

SNCA co-expression module and PPI network conserved across ACC, MTG, and SN is highly enriched for exocytosis and selective autophagy functions: WGCNA and hierarchical clustering applied to normalized gene count transcription matrices for all nuclei identified clusters of genes co-expressed with SNCA for SN (n = 1206) (Fig. 10A). Overlapping ACC, MTG, and SN modules produced a conserved set of 29 genes co-expressed with SNCA for all three regions (Fig.10B). As for conserved ACC-MTG module genes, these 29 genes were used to generate a PPI network which likewise had SNCA as its top hub gene (Fig.10C). The MTG-ACC-SN SNCA co-expression module PPI network was found to be enriched for exocytic processes, neurotransmitter uptake, selective autophagy, and mitochondrial protein localization functions (Fig. 10D).

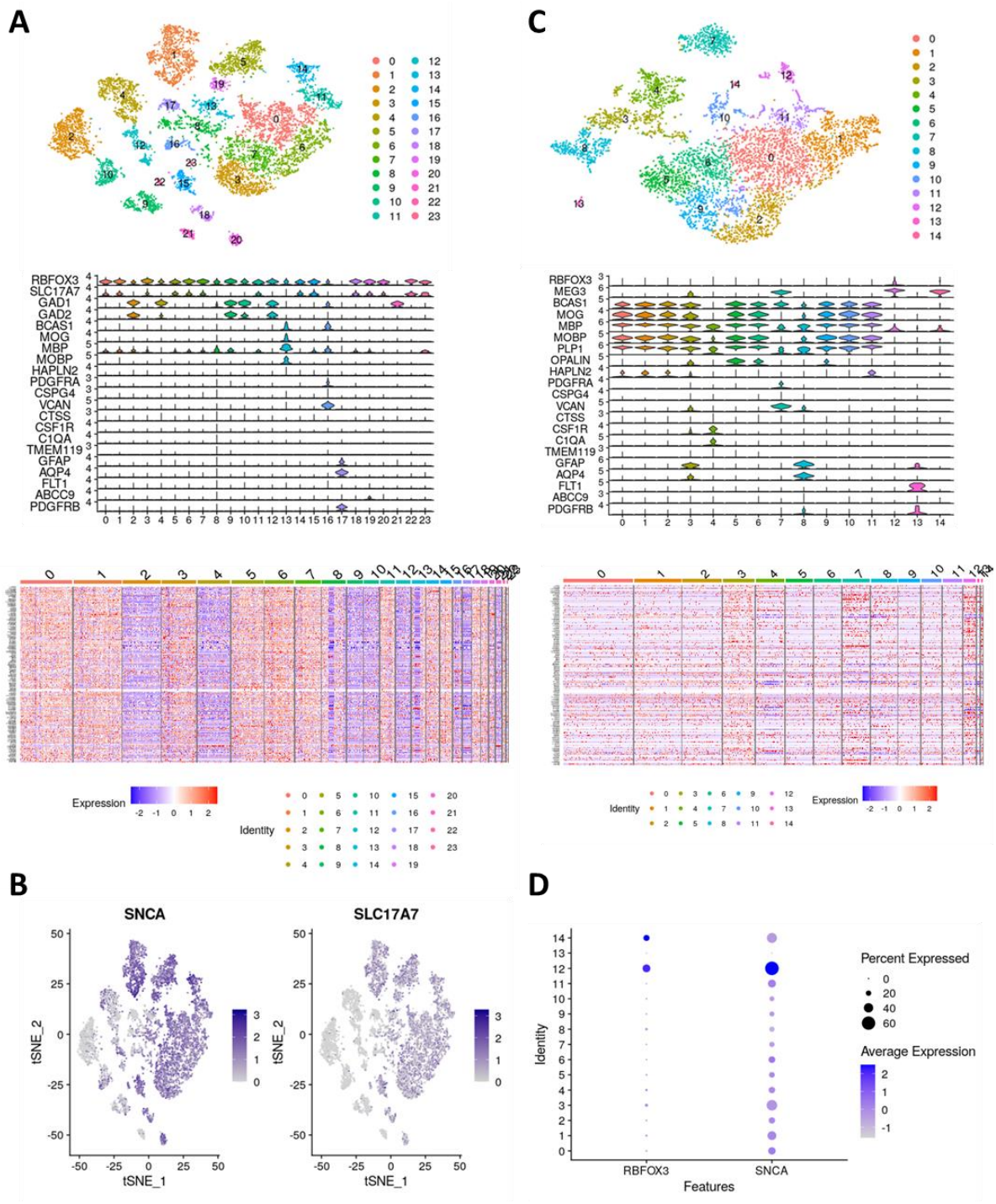


Figure 9. SNCA module gene expression in independent data obtained from human cortex and substantia nigra samples. Excitatory neurons are observed to have greater expression of SNCA co-expression ACC-MTG module genes in cortex (A, B). In substantia nigra, increased expression of module genes is observed in cell clusters expressing oligodendrocyte precursor cell marker genes and neuronal cells (C). Substantia nigra SNCA expression is observed in neuronal and oligodendrocyte clusters (D).

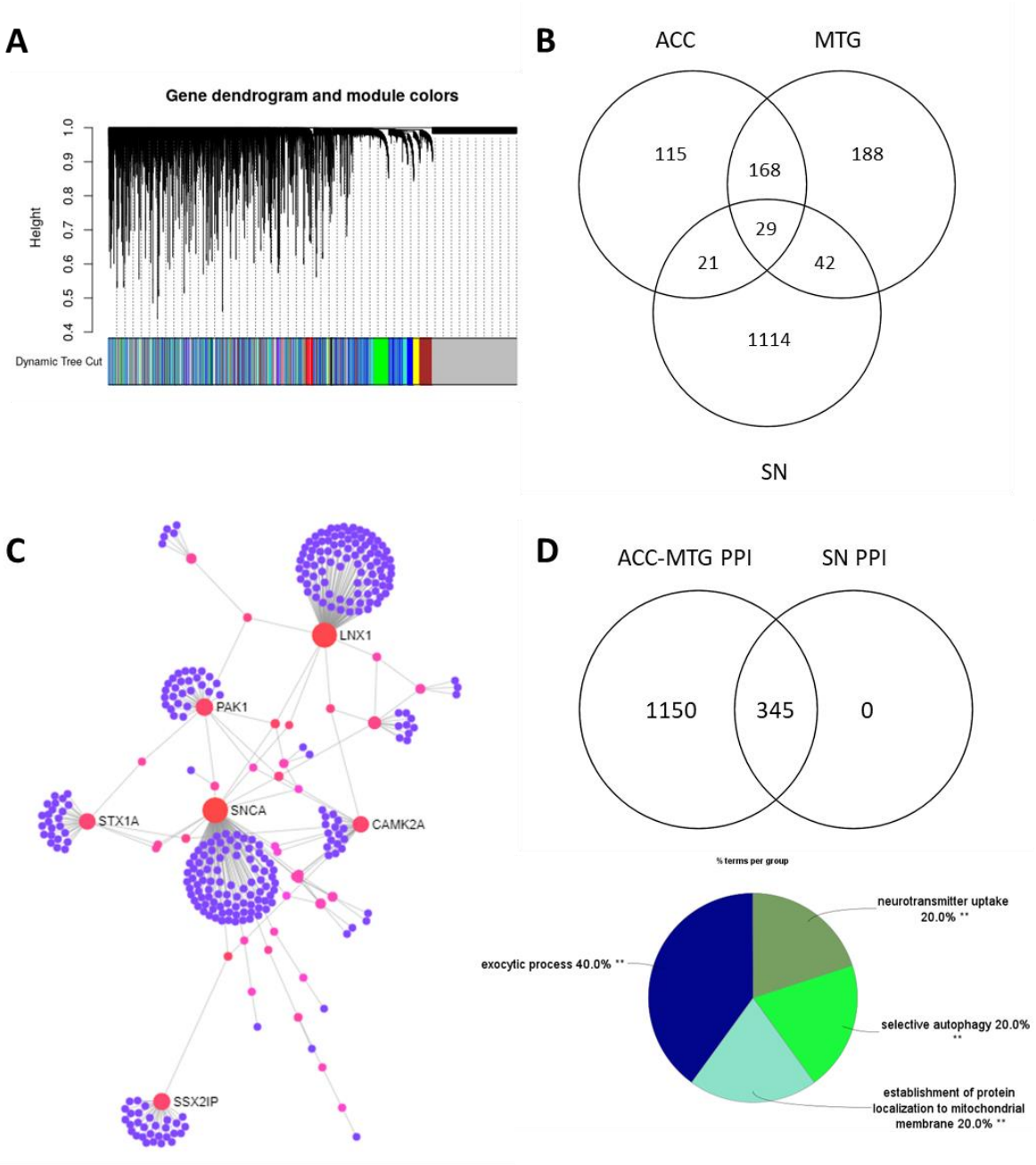


Figure 10. Substantia Nigra SNCA module derived by WGCNA. SNCA co-expression module annotated in gene dendrogram by brown color bar (A). Overlap of SNCA-co-expressed genes among ACC, MTG, and substantia nigra locations (B). Substantia nigra PPI network also has SNCA as its top hub by degree and betweenness measures (C). Substantia nigra SNCA module PPI subnetwork is significantly enriched for functions including signaling, autophagy and protein processing (Gene ontology analysis performed in Cytoscape ClueGo with enrichment p-value < 10e-6 for all terms and with term fusion using GO Tree Interval 3-5 (D).

4.1.7 Conclusions

Among the cell groupings identified in our analysis, we observed conserved patterns of differential neuronal SNCA expression in ACC and MTG regions, with high SNCA expression observed in excitatory (glutamergic; GLUT) neurons and low expression in inhibitory (gabaergic; GABA) neurons. These observations were found to be consistent for GLUT and GABA type neurons identified in MFG samples. A protein-protein interaction network was then constructed for genes in the identified conserved SNCA co-expression module based on known protein-protein interaction pathways.

Significant and novel scientific contributions of this work are the following:

- This analysis newly identifies a protein-protein interaction network conserved across cortical regions which is significantly enriched for PD genetic risk loci [78] and contains a number of genes observed to be differentially expressed in PD versus control brain tissue samples.
- The separate PPI network corresponding to the SN-specific WGCNA SNCA module is shown to be a subnetwork of the ACC-MTG PPI network enriched for exocytosis, neurotransmitter uptake, selective autophagy, and mitochondrial membrane protein localization functions. Separately deriving SNCA networks in tissues of different underlying cell type compositions enables our new identification of a subnetwork comprising SNCA functional biology shared by neurons and oligodendrocytes.
- This work demonstrates derivation and validation of gene co-expression networks suitable for integration with external resources as shown in the preceding section.
-

4.2 Further Discussion

As demonstrated in the preceding section of the dissertation, data-derived biological networks can also be further transformed to serve as input to exploratory workflows using methods such as graph embedding for feature engineering. The research work in this section is an application of network analysis to identify a functional interactome for a gene of interest and validate the quality of the derived network and its biological significance. Once generated and validated it becomes apparent how such a custom network can be integrated to further refine performance in data mining or prediction tasks.

5 | Conclusion

Data science and artificial intelligence offer great potential for advancing biomedical research and drug development. However, adapting data science ideas and methods for domain-specific research and development requires thoughtful management, complementary data resources, and interdisciplinary problem formulations. The work presented in this dissertation reflects problem solving across several important areas in works of this type: data preprocessing, integration, exploratory analysis, network modeling, and formulation of evaluation and validation procedures.

The work included in this dissertation engages with both scientific and data science challenges and makes a number of significant contributions to data science specifically. Repeating the outlined summary from section 1.5:

Problem I: CCA for Correlation Analysis and Derived Feature Embeddings

- Novel dataset derived from integration of US EPA and State-level mortality data resources **engineered specifically for ML algorithms**
- Innovative **CCA-derived epidemiological analysis provides a novel quantification of exposure-outcome association** which more strongly and significantly quantifies air quality and health outcomes relationships through covariation models then linear regression.
- **Alignment of multiple datasets by CCA-based features leveraged to extract new insights** into regional cell state biologies and their relationship to disease states from transcriptomic data, overcoming the challenges of comparing results from **unsupervised clustering.**

Problem II: Link Prediction in Evidence Networks Using Local Information and Features Generated by New Collaborative Filtering Methods

- Created new search procedure based on ML using **local graph topology** for predicting drug-target relationships from Open Targets platform aggregated association score data using platform API **which outperforms previously published approaches for this task**.
- Lead development of **novel feature engineering** project for use with Open Targets association evidence which **expands the utility of the platform information network** and **achieves a substantial improvement in performance for drug-target relationship prediction over all previously published methods for this task**.
- Application of collaborative filtering concepts in paper 2 for omics datasets integration for query applications is especially significant because the **volume, size, and variability**

of public Omics data archives creates challenges in access and processing times for query and data mining applications which are met by work in paper 2.

Problem III: Network Construction for Data Mining and Biological Insight

- Gene co-expression network topologies vary with multiple factors including disease state, tissue of interest, and cell type. Such networks can be used as inputs for custom feature generation as shown in problem 2.
- For this problem, **multiple novel gene networks are created from single-cell RNAseq transcriptomic data to achieve association rule learning, classification, and clustering tasks.**
- Data mining performed from these networks derives new insights into cell-specific and regional brain biology.

While a number of new biological insights were obtained alongside the above-listed data science contributions, these could not have been obtained except by innovative use of data science concepts and methods and their translation into methods for biomedical research. The overarching theme across the projects presented in this dissertation is the necessity in translational data science work to take a systematic approach which begins with the parallel formulation of the biological domain question and data science problem. While complementary in the end results, the foundations and conventions of these parts of the work can also be understood presented separately as found for the data science aspects of these works in this dissertation.

6 | Future Directions

In a recent overview article, Ferrero et al. offered a set of strategic recommendations for adapting organizational structure and culture to make effective use of data science in drug industry work. Among these, their first recommendation was to recognize data science as a separate core discipline focused on access, integration, and knowledge extraction from internal and external data resources, standing alongside longer-established industry domains including biology, medicine, and chemistry [98]. While new computational methods are constantly being introduced across disciplines, data science work focuses particularly on advancing methods that deal with several core information attributes colloquially referred to as the five V's: methods which manage large *volumes* of information, deal with *variety* in data sources, perform with reasonable computing times (*velocity*), yield valid and reliable results *(veracity*), and which fulfill the original task purpose (*value*) [99]. Searching among computational methods to find those which are most suited to a particular task, identifying limitations and potential sources of bias or error in diverse analytical systems, and managing large volumes of data are each the particular concerns and contributions of data scientists on project teams.

Conceptually separating data science considerations in biomedical research applications positions these ideas to be involved at project initiation, where they have the best opportunity to fully understand the concepts and scope of a project. Process models such as the CRISP-DM workflow are used across industries to conceptualize how data science work interfaces with domain-specific applications and tasks [100, 101]. As detailed in the CRISP-DM model, first-step involvement of multiple team perspectives in project development serves several important purposes: first, and most importantly, it is an opportunity for communication among different

domain experts about the aims and scope of a work. These discussions then guide identification of potentially relevant existing data resources, and discussions about formats for newly generated data. Discussions should also be had at this point about evaluation criteria. For instance, a laboratory research team may be most interested in generating data for inferential analysis, biological hypothesis testing, or knowledge extraction from one or more data resources rather than a specific prediction task. For data science-driven projects, shared understanding of concepts and methods among all experts provides the foundation for discussions of exploratory and analytical results that will be necessary to avoid missing unique insights. Bringing such ideas together early in a project is particularly important since customization of methods developed for applications outside of biology and medicine may take considerable time. Framing needs as data questions and identifying both analysis and domain application objectives provides an optimal foundation for project progress and success.

Data Resource Selection Informed by Problem Understanding: As is readily demonstrated in the number and variety of data resources used in this dissertation, unified, scientific data management is a major concern for the application of data science methods in biomedical research domains, as well as more broadly for business organizations and medical institutions [102]. At all levels, systematic protocols for data storage facilitate communication and collaboration and ensure that important and impactful findings can be validated, reproduced, and extended. Good data stewardship practices provide a foundation for ongoing and future work, enhances experiment value, reduces duplication, and are necessary for legal documentation. Yet developing and implementing data strategies that extend within and across organizations remains a significant challenge. The FAIR (Findable, Accessible, Interoperable, Reusable) guidelines are a set of principles set out to guide organizational data management [103-105]. A particular

emphasis of the FAIR guidelines are recommendations to select data formats and storage protocols that make results more widely accessible. For example, standardized metadata annotation protocols make stored data more findable through topic and key word searches. As can be seen in the preceding works, integration of country-level air quality and mortality datasets is made possible through the encoding of county codes, omics data is integrated using standard gene nomenclatures, and archiving published datasets supports validation and extension of initial analysis results. It should be noted that retroactive data cleaning and processing is time-consuming, expensive, and is recognized as a significant barrier in implementing AI- and machine learning internal data mining in many drug industry applications [103].

Data Structures that Influence Discovery: Biological systems, disease pathways, and other relevant scientific phenomena can be represented and studied using numbers, symbols, and algorithms, but these representations are just that, representations and models, which may to varying degrees reflect the trues area of scientific and business interest. The FAIR guidelines highlight the need for data to be both discoverable and easy to access and manipulate [103-105] while an overarching theme of work in this dissertation is the opportunities that exist in at each stage of analysis for thoughtful and creative use of concepts and strategies from data science to extract knowledge and achieve novel insights. At the organizational level, vast archives of previous experimental and clinical data exist. Available, searchable, and easy to manipulate stored data opens up greater possibilities for integration and analysis where there is communication on aims of analysis, collaborative understanding of information contained in different datasets, and collaboration to formulate how data science strategies can be applied.

References

1. Teeple, E., et al., *Integrated label transfer for oligodendrocyte population profiling in Parkinson's Disease and Multiple System Atrophy.* HEALTHINF, 2022.
2. Han, Y., et al., *Empowering the discovery of novel target-disease associations via machine learning approaches in the Open Targets platform.* BMC Bioinformatics, 2022. **23**(1): p. 1-19.
3. Teeple, E., Y. Chang, and D. Rajpal, *A target-specific evidence function for indication expansion queries in the Open Targets platform.* Proceedings IEEE BHI-BSN 2021.
4. Teeple, E., et al., *Air quality and cause-specific mortality in the United States: association analysis by regression and CCA for 1980-2014.* HEALTHINF, 2020.
5. Teeple, E., et al., *Network analysis and human single cell brain transcriptomics reveal novel aspects of alpha-synuclein (SNCA) biology.* Biorxiv, 2020.
6. Ryan, S.K., et al., *Microglia ferroptosis is regulated by SEC24B and contributes to neurodegeneration.* Nat Neurosci, 2023. **26**(1): p. 12-26.
7. Boddupalli, C.S., et al., *Neuroinflammation in neuronopathic Gaucher disease: Role of microglia and NK cells, biomarkers, and response to substrate reduction therapy.* Elife, 2022. **11**.
8. Tasdemir-Yilmaz, O., et al., *Dorsal Root Ganglia Single-Nucleus Transcriptomics Reveal Cellular and Molecular Responses to High Dose AAV-Induced Toxicity.* Molecular Therapy, 2022. **30**(4): p. 528-529.
9. Teeple, E., et al., *Clinical performance evaluation of a machine learning system for predicting hospital-acquired clostridium difficile infection.* HEALTHINF, 2020.
10. Hartvigsen, T., et al., *Early prediction of MRSA infections using electronic health records.* HEALTHINF, 2018.
11. Hotelling, H., *Relations between two sets of variates.* Biometrika, 1936. **28**: p. 321-377.
12. Gonzalez, I., et al., *CCA: An R package to extend canonical correlation analysis.* J Stat Software, 2008. **23**(12).
13. Ochoa, D., et al., *Open Targets Platform: supporting systematic drug-target identification and prioritisation.* Nucleic Acids Res, 2021. **49**(D1): p. D1302-D1310.
14. Carvalho-Silva, D., et al., *Open Targets Platform: new developments and updates two years on.* Nucleic Acids Res, 2019. **47**(D1): p. D1056-D1065.
15. Koscielny, G., et al., *Open Targets: a platform for therapeutic target identification and validation.* Nucleic Acids Res, 2017. **45**(D1): p. D985-D994.
16. Euler, L., *Solutio problematis ad geometriam situs pertinentis.* Comment Acad Sci U Petrop, 1736. **8**: p. 128-40.
17. Emmert-Streib, F., M. Dehmer, and B. Haibe-Kains, *Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks.* Front Cell Dev Biol, 2014. **2**: p. 38.
18. Di, Q., F. Dominici, and J.D. Schwartz, *Air Pollution and Mortality in the Medicare Population.* N Engl J Med, 2017. **377**(15): p. 1498-9.
19. Shah, A.S., et al., *Short term exposure to air pollution and stroke: systematic review and meta-analysis.* BMJ, 2015. **350**: p. h1295.
20. Han, C., et al., *Air quality management policy and reduced mortality rates in Seoul Metropolitan Area: A quasi-experimental study.* Environ Int, 2018. **121**(Pt 1): p. 600-609.
21. Peng, L., et al., *Short-term associations between size-fractionated particulate air pollution and COPD mortality in Shanghai, China.* Environ Pollut, 2020. **257**: p. 113483.
22. India State-Level Disease Burden Initiative Air Pollution, C., *The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017.* Lancet Planet Health, 2019. **3**(1): p. e26-e39.
23. Wang, T., et al., *Mortality burdens in California due to air pollution attributable to local and nonlocal emissions.* Environ Int, 2019. **133**(Pt B): p. 105232.
24. James, G., et al., *Introduction to Statistical Learning with Applications in R*. 2014.

25. Rudzicz, F., *Adaptive Kernel Canonical Correlation Analysis for Estimation of Task Dynamics from Acoustics. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2010.
26. Andrew, G., et al., *Deep Canonical Correlation Analysis.* Proceedings of the 30th International Conference on Machine Learning, 2013.
27. Agency, U.S.E.P., *County Monitor Annual Summary Files*. online.
28. Evaluation, I.f.H.M.a., *United States Combined and Gender-Specific Age-Adjusted Mortality Rates by United States County*.
29. Vineis, P. and D. Kriebel, *Causal models in epidemiology: past inheritance and genetic future.* Environ Health, 2006. **5**: p. 21.
30. Pedregosa, F. and e. al, *Scikit-learn: Machine Learning in Python.* JMLR, 2011. **12**: p. 2825-2830.
31. Kutner, M., et al., *Applied Linear Statistical Models 5th Edition*. 2004: McGraw-Hill/Irwin.
32. Cuevas-Diaz Duran, R., H. Wei, and J.Q. Wu, *Single-cell RNA-sequencing of the brain.* Clin Transl Med, 2017. **6**(1): p. 20.
33. Hao, Y., et al., *Integrated analysis of multimodal single-cell data.* Cell, 2021. **184**(13): p. 3573-3587 e29.
34. Hafemeister, C. and R. Satija, *Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression.* Genome Biol, 2019. **20**(1): p. 296.
35. Stuart, T., et al., *Comprehensive Integration of Single-Cell Data.* Cell, 2019. **177**(7): p. 1888-1902 e21.
36. Butler, A., et al., *Integrating single-cell transcriptomic data across different conditions, technologies, and species.* Nat Biotechnol, 2018. **36**(5): p. 411-420.
37. Finak, G., et al., *MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data.* Genome Biol, 2015. **16**: p. 278.
38. Kramer, A., et al., *Causal analysis approaches in Ingenuity Pathway Analysis.* Bioinformatics, 2014. **30**(4): p. 523-30.
39. Xie, Z., et al., *Gene Set Knowledge Discovery with Enrichr.* Curr Protoc, 2021. **1**(3): p. e90.
40. Jellinger, K.A., *Multiple System Atrophy: An Oligodendroglioneural Synucleinopathy1.* J Alzheimers Dis, 2018. **62**(3): p. 1141-1179.
41. Henderson, M.X., J.Q. Trojanowski, and V.M. Lee, *alpha-Synuclein pathology in Parkinson's disease and related alpha-synucleinopathies.* Neurosci Lett, 2019. **709**: p. 134316.
42. Gilman, S., et al., *Second consensus statement on the diagnosis of multiple system atrophy.* Neurology, 2008. **71**(9): p. 670-6.
43. Ibanez, P., et al., *Causal relation between alpha-synuclein gene duplication and familial Parkinson's disease.* Lancet, 2004. **364**(9440): p. 1169-71.
44. Polymeropoulos, M.H., et al., *Mutation in the alpha-synuclein gene identified in families with Parkinson's disease.* Science, 1997. **276**(5321): p. 2045-7.
45. Singleton, A.B., et al., *alpha-Synuclein locus triplication causes Parkinson's disease.* Science, 2003. **302**(5646): p. 841.
46. Scholz, S.W., et al., *SNCA variants are associated with increased risk for multiple system atrophy.* Ann Neurol, 2009. **65**(5): p. 610-4.
47. Kiely, A.P., et al., *alpha-Synucleinopathy associated with G51D SNCA mutation: a link between Parkinson's disease and multiple system atrophy?* Acta Neuropathol, 2013. **125**(5): p. 753-69.
48. Langedijk, J., et al., *Drug repositioning and repurposing: terminology and definitions in literature.* Drug Discov Today, 2015. **20**(8): p. 1027-34.
49. Vamathevan, J., et al., *Applications of machine learning in drug discovery and development.* Nat Rev Drug Discov, 2019. **18**(6): p. 463-477.
50. Ashburn, T.T. and K.B. Thor, *Drug repositioning: identifying and developing new uses for existing drugs.* Nat Rev Drug Discov, 2004. **3**(8): p. 673-83.

51. Khaladkar, M., et al., *Uncovering novel repositioning opportunities using the Open Targets platform.* Drug Discov Today, 2017. **22**(12): p. 1800-1807.
52. Freudenberg, J.M., et al., *Uncovering new disease indications for G-protein coupled receptors and their endogenous ligands.* BMC Bioinformatics, 2018. **19**(1): p. 345.
53. Ferrero, E., I. Dunham, and P. Sanseau, *In silico prediction of novel therapeutic targets using gene-disease association data.* J Transl Med, 2017. **15**(1): p. 182.
54. *Open Targets python API*. 2021.
55. Chen, T. and C. Guestrin, *XGBoost: a scalable tree boosting system.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: p. 785–794.
56. Yip, A.M. and S. Horvath, *Gene network interconnectedness and the generalized topological overlap measure.* BMC Bioinformatics, 2007. **8**: p. 22.
57. Peng, J., et al., *Improving the measurement of semantic similarity by combining gene ontology and co-functional network: a random walk based approach.* BMC Syst Biol, 2018. **12**(Suppl 2): p. 18.
58. Consortium, G.T., *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.
59. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
60. Gene Ontology, C., *The Gene Ontology resource: enriching a GOld mine.* Nucleic Acids Res, 2021. **49**(D1): p. D325-D334.
61. Zhao, C. and Z. Wang, *GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms.* Sci Rep, 2018. **8**(1): p. 15107.
62. Szklarczyk, D., et al., *STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.* Nucleic Acids Res, 2019. **47**(D1): p. D607-D613.
63. Grover, A. and J. Leskovec, *node2vec: Scalable Feature Learning for Networks.* KDD, 2016. **2016**: p. 855-864.
64. Breese, J.S., D. Heckerman, and C.M. Kadie, *Empirical Analysis of Predictive Algorithms for Collaborative Filtering.* UAI, 1998.
65. Liaw, A. and M. Wiener, *Classification and regression by randomForest.* R News, 2002. **2/3**.
66. Ezkurdia, I., et al., *Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes.* Hum Mol Genet, 2014. **23**(22): p. 5866-78.
67. Hodge, R.D., et al., *Conserved cell types with divergent features in human versus mouse cortex.* Nature, 2019. **573**(7772): p. 61-68.
68. © *Allen Institute for Brain Science: Cell Types Database*. 2015.
69. Agarwal, D., et al., *A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders.* Nat Commun, 2020. **11**(1): p. 4183.
70. Clinton, L.K., et al., *Synergistic Interactions between Abeta, tau, and alpha-synuclein: acceleration of neuropathology and cognitive decline.* J Neurosci, 2010. **30**(21): p. 7281-9.
71. Coon, E.A., J.K. Cutsforth-Gregory, and E.E. Benarroch, *Neuropathology of autonomic dysfunction in synucleinopathies.* Mov Disord, 2018. **33**(3): p. 349-358.
72. Hague, K., et al., *The distribution of Lewy bodies in pure autonomic failure: autopsy findings and review of the literature.* Acta Neuropathol, 1997. **94**(2): p. 192-6.
73. Spillantini, M.G., et al., *Alpha-synuclein in Lewy bodies.* Nature, 1997. **388**(6645): p. 839-40.
74. Inoue, M., et al., *The distribution and dynamic density of oligodendroglial cytoplasmic inclusions (GCIs) in multiple system atrophy: a correlation between the density of GCIs and the degree of involvement of striatonigral and olivopontocerebellar systems.* Acta Neuropathol, 1997. **93**(6): p. 585-91.
75. Devine, M.J., et al., *Parkinson's disease and alpha-synuclein expression.* Mov Disord, 2011. **26**(12): p. 2160-8.

76. Chen, X., et al., *Potassium Channels: A Potential Therapeutic Target for Parkinson's Disease.* Neurosci Bull, 2018. **34**(2): p. 341-348.
77. Alegre-Abarrategui, J., et al., *Selective vulnerability in alpha-synucleinopathies.* Acta Neuropathol, 2019. **138**(5): p. 681-704.
78. Nalls, M.A., et al., *Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies.* Lancet Neurol, 2019. **18**(12): p. 1091-1102.
79. Lesage, S. and A. Brice, *Parkinson's disease: from monogenic forms to genetic susceptibility factors.* Hum Mol Genet, 2009. **18**(R1): p. R48-59.
80. Mata, I.F., et al., *Glucocerebrosidase gene mutations: a risk factor for Lewy body disorders.* Arch Neurol, 2008. **65**(3): p. 379-82.
81. Lesage, S., et al., *Large-scale screening of the Gaucher's disease-related glucocerebrosidase gene in Europeans with Parkinson's disease.* Hum Mol Genet, 2011. **20**(1): p. 202-10.
82. Sidransky, E., et al., *Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease.* N Engl J Med, 2009. **361**(17): p. 1651-61.
83. Zhang, B. and S. Horvath, *A general framework for weighted gene co-expression network analysis.* Stat Appl Genet Mol Biol, 2005. **4**: p. Article17.
84. Langfelder, P. and S. Horvath, *Fast R Functions for Robust Correlations and Hierarchical Clustering.* J Stat Softw, 2012. **46**(11).
85. Ravasz, E., et al., *Hierarchical organization of modularity in metabolic networks.* Science, 2002. **297**(5586): p. 1551-5.
86. Ravasz, E., *Detecting hierarchical modularity in biological networks.* Methods Mol Biol, 2009. **541**: p. 145-60.
87. Cunningham, F., et al., *Ensembl 2019.* Nucleic Acids Res, 2019. **47**(D1): p. D745-D751.
88. Zhou, G., et al., *NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis.* Nucleic Acids Res, 2019. **47**(W1): p. W234-W241.
89. Oliveros, J.C. *Venny. An interactive tool for comparing lists with Venn's diagrams*. 2007-2015; Available from: https://bioinfogp.cnb.csic.es/tools/venny/index.html.
90. Chen, E.Y., et al., *Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.* BMC Bioinformatics, 2013. **14**: p. 128.
91. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic Acids Res, 2016. **44**(W1): p. W90-7.
92. Mlecnik, B., J. Galon, and G. Bindea, *Automated exploration of gene ontology term and pathway networks with ClueGO-REST.* Bioinformatics, 2019. **35**(19): p. 3864-3866.
93. Mlecnik, B., J. Galon, and G. Bindea, *Comprehensive functional analysis of large lists of genes and proteins.* J Proteomics, 2018. **171**: p. 2-10.
94. Zhou, B. and W. Jin, *Visualization of Single Cell RNA-Seq Data Using t-SNE in R.* Methods Mol Biol, 2020. **2117**: p. 159-167.
95. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
96. Merico, D., et al., *Enrichment map: a network-based method for gene-set enrichment visualization and interpretation.* PLoS One, 2010. **5**(11): p. e13984.
97. Pinero, J., et al., *The DisGeNET knowledge platform for disease genomics: 2019 update.* Nucleic Acids Res, 2020. **48**(D1): p. D845-D855.
98. Ferrero, E., et al., *Ten simple rules to power drug discovery with data science.* PLoS Comput Biol, 2020. **16**(8): p. e1008126.
99. Jain, A., *The 5 V's of big data*, in *Watson Health Perspectives*. 2016: online.
100. Shearer, C., *The CRISP-DM model: the new blueprint for data mining.* J Data Warehousing 2000. **5**: p. 13-22.
101. Azevedo, A. and M.F. Sanots. *KDD, SEMMA and CRISP-DM: a parallel overview*. in *Proceedings of the IADIS European Conference on Data Mining* 2008.

102. Fortunato, A., D.W. Grainger, and M. Abou-El-Enein, *Enhancing patient-level clinical data access to promote evidence-based practice and incentivize therapeutic innovation.* Adv Drug Deliv Rev, 2018. **136-137**: p. 97-104.
103. Wise, J., et al., *Implementation and relevance of FAIR data principles in biopharmaceutical R&D.* Drug Discov Today, 2019. **24**(4): p. 933-938.
104. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Sci Data, 2016. **3**: p. 160018.
105. Wilkinson, M.D., et al., *A design framework and exemplar metrics for FAIRness.* Sci Data, 2018. **5**: p. 180118.