Developing and Applying Machine Learning and Artificial Intelligence Approaches to Leverage Multi-Omics to Advance Understanding of Alzheimer's Disease Pathology

By

Jerome J. Choi

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Epidemiology)

at the

University of Wisconsin-Madison

2025

Date of final oral examination: 05/12/2025

This dissertation is approved by the following members of the Final Oral Committee:

Corinne D. Engleman, Professor, Population Health Sciences
Daifeng Wang, Associate Professor, Biostatistics and Medical Informatics
John Svaren, Professor, Comparative Biosciences
Shaneda Warren Anderson, Assistant Professor, Population Health Sciences
Megan Zuelsdorff, Assistant Professor, School of Nursing
Tianyuan Lu, Assistant Professor, Population Health Sciences

Table of Contents

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder driven by complex, multifactorial processes involving genetic, epigenetic, transcriptomic, and metabolic dysregulation. Understanding the molecular mechanisms underlying AD requires approaches that can capture the dynamic interactions across these biological layers. The disease is not the result of a single molecular event but emerges from disruptions in coordinated biological networks, making it difficult to identify clear causal pathways or biomarkers. Multi-omics technologies provide a means to interrogate these layers simultaneously, offering a comprehensive view of disease biology by linking genetic variation to downstream molecular and cellular phenotypes. However, the integration and interpretation of multi-omics data pose significant analytical challenges due to high dimensionality, heterogeneity, sparsity, and biological noise. Machine learning (ML) offers powerful tools to address these challenges by modeling non-linear relationships, uncovering hidden patterns, learning from high-dimensional data, and enabling accurate and interpretable phenotype prediction. This dissertation develops and applies novel ML-based frameworks for integrative multi-omics analysis in the context of AD. First, the roles of cooperative transcription factor (TF) pairs in regulating target genes within the gene regulatory networks (GRNs) of oligodendrocytes were characterized, utilizing multi-omics data and deep learning approaches. Second, COSIME: Cooperative Multi-view Integration and Scalable and Interpretable Model Explainer was developed with applications to Alzheimer's Disease (AD). Third, multi-omics integration of transcriptomics and metabolomics with machine learning was used to uncover novel risk factors for AD. By combining computational innovation with biological insight, this work advances systems-level understanding of AD and contributes new tools for precision medicine and biomarker discovery.

Dedication

As I reflect on my Ph.D. journey, I am deeply aware that none of this would have been possible without the love, support, and encouragement of those closest to me. The challenges and triumphs along the way have been shared with many who have been my unwavering source of strength. It is with immense gratitude that I dedicate this achievement to them.

I would like to express my deepest gratitude to my family. To my parents, thank you for your endless love, encouragement, and belief in me. Your support has been the foundation of everything I've achieved. To my wife, thank you for your unwavering love, patience, and strength. You have been my constant source of support, balance, and inspiration throughout this journey. To my brother and his family, thank you for your constant support and for being there throughout this journey. I would also like to express my heartfelt gratitude to my in-laws. Your warmth, kindness, and support have been a great source of comfort and strength, and I am so thankful for the family you've embraced me into. And to our beloved dogs, thank you for your comforting presence and loyal companionship during the many long days and late nights of this work.

Finally, I want to thank all those who have quietly supported me along the way. Your belief in me, your encouragement, and your constant presence have made this journey possible. I am deeply grateful for the role each of you has played in helping me reach this point.

Acknowledgements

I am deeply grateful to my co-advisors, Drs. Corinne Engelman and Daifeng Wang, for your invaluable guidance, mentorship, and support throughout my graduate studies. Your insight, high standards, and trust in my work have profoundly shaped my development as a researcher. I am also grateful to my additional supervisors, Drs. John Svaren and Tianyuan Lu, for their support, advice, and encouragement throughout this process. Their perspectives and expertise have been instrumental in broadening my scientific thinking. I also thank my committee members, Drs. Shaneda Warren Anderson, and Megan Zuelsdorff, for their thoughtful feedback and guidance, which have greatly enriched the direction and quality of this dissertation.

I am also incredibly thankful to my mentors, peers, and friends at the Engelman Lab, Wang Lab, and the Waisman Center, whose expertise, thoughtful discussions, and shared enthusiasm made this work not only possible but deeply rewarding. I have learned so much from working alongside such talented, curious, and generous individuals. Your willingness to share ideas, provide feedback, and collaborate on challenging problems has been invaluable to my growth as a scientist. I especially appreciate the supportive and collegial environment you created—one that encouraged exploration, resilience, and continuous learning. Your encouragement, insights, and friendship have meant a great deal to me, both professionally and personally, and I carry those experiences with deep gratitude.

Additionally, I would like to express my sincere gratitude to the funding agencies that financially supported this work, including R21 NS128761, RF1 MH128695, R01 AG067025, the National Science Foundation CAREER Award (2144475), and a core grant to the Waisman Center from NICHD (P50 HD105353). I am also deeply grateful to the participants, investigators, and

teams from the WRAP and Wisconsin ADRC, whose contributions were invaluable to this research. This work was further supported by grants from the National Institutes of Health, including R01 AG27161, P30 AG062715, and RF1 AG054047. Finally, I would like to acknowledge the Waisman Center for their fellowship support through the Wisconsin Distinguished Graduate Fellowships, which provided critical funding and resources that greatly contributed to the advancement of my research.

At last, I am deeply grateful for the collective contributions of all those who have supported and inspired me throughout this process. Your guidance, encouragement, and collaboration have been essential in shaping this dissertation and my academic journey. I will always be thankful for the role each of you has played in helping me reach this milestone, and I carry your impact with me.

Introduction

Understanding Complex Biological Systems

Biological systems are inherently complex, involving intricate networks of molecular interactions that govern cellular functions and organismal behavior[1]. These systems operate across multiple scales, from individual molecules to entire organs, with a web of interconnected biological processes that ensure proper function. At the core of biological complexity lies the interplay between genes, proteins, metabolites, and other cellular components, each playing a vital role in regulating cellular states and responses to environmental signals[2].

To fully understand how biological systems function, we must consider not only individual components but also their dynamic interactions. This complexity is evident in numerous physiological processes, including development, immune responses, and disease mechanisms[3–5]. For example, the expression of genes in a specific cell type is not driven by isolated factors but is instead shaped by a coordinated network of regulatory mechanisms, such as transcription factors, epigenetic modifications, and chromatin remodeling[6]. Moreover, cellular responses are often governed by the integration of external stimuli and the internal state of cell, creating a continuously adaptive environment[7].

As the understanding of these systems deepens, it becomes clear that studying biological processes in isolation does not provide a comprehensive understanding of their roles in health and disease. Traditional, reductionist approaches have focused on individual molecules or pathways, but the complexity of biological systems demands a more holistic perspective[1]. This has led to the rise of multi-omics approaches, which enable the simultaneous study of several layers of biological information from genomics, transcriptomics, metabolomics, and epigenomics[8–12]. By integrating

data from these different sources, multi-omics can provide a more complete picture of the biological processes at play and how they contribute to various states, including diseases.

In the context of complex diseases, such as Alzheimer's disease, understanding these systems becomes even more critical. Diseases often involve disruptions at multiple levels, from genetic mutations to changes in gene expression, protein function, and cellular metabolism[13,14]. To understand these multifaceted processes and their contribution to disease progression, we must adopt a systems biology approach — one that integrates information across these diverse molecular layers.

Alzheimer's Disease

Alzheimer's disease (AD) is a prime example of a complex, multifactorial disorder, characterized by the progressive degeneration of cognitive functions, memory loss, and behavioral changes[10,14–16]. Its pathophysiology is influenced by genetic, epigenetic, and metabolic factors, which collectively contribute to the onset and progression of the disease[17,18]. At the molecular level, AD is marked by the accumulation of abnormal protein aggregates, such as amyloid-beta plaques and tau tangles, which disrupt normal cellular function[19–25]. However, these hallmark features alone do not fully explain the disease's complexity, and additional molecular factors, including changes in gene expression and metabolism, play a critical role in disease progression.

The genetic component of AD is crucial, with several risk genes, such as *APOE*, contributing to susceptibility[26–28]. Yet, genetic risk alone does not account for the entire disease process, suggesting the involvement of intricate molecular networks. Beyond genetic mutations, disruptions in cellular processes, such as synaptic plasticity, neurotransmission, mitochondrial

function, and inflammation also play significant roles in disease progression[29–35]. These changes occur across multiple levels, from molecular signaling to tissue and organ dysfunction, making AD a disease that transcends simple genetic or biochemical explanation.

Furthermore, the role of the brain's immune system in AD, particularly the activation of microglia and astrocytes, has become increasingly evident[34,36–41]. Chronic neuroinflammation, a key feature of AD, not only exacerbates neuronal damage but also alters the trajectory of disease progression[35,42]. Additionally, alterations in metabolism, including impaired glucose metabolism in the brain, have been linked to cognitive decline, further highlighting the multifaceted nature of AD[43,44].

To better understand AD and develop effective treatments, it is essential to adopt a holistic approach that focuses on the complex interplay between genomics, transcriptomics, metabolomics, and epigenomics. By integrating these molecular layers, we can uncover new insights into the mechanisms underlying the disease and identify potential therapeutic targets. This complexity highlights the need for advanced methodologies capable of analyzing vast amounts of omics data to provide a more comprehensive understanding of AD.

Multi-Omics Approaches in Biological Research

The study of biological systems has historically relied on single-omics approaches, such as genomics or transcriptomics, to examine isolated aspects of cellular function[45,46]. However, these approaches only offer a limited understanding of complex biological processes, as they capture only one layer of information in a multi-faceted network of biological interactions. To gain a more comprehensive understanding, multi-omics approaches have emerged as powerful tools to integrate and analyze data across several biological domains simultaneously. By incorporating

diverse types of omics data, such as genomics, transcriptomics, metabolomics, and epigenomics, we can explore the interconnectedness of genetic, molecular, and biochemical features within cells and tissues[11,47–49].

Each omic layer provides valuable insight into different biological processes: genomics reveals the genetic blueprint[50], transcriptomics identifies gene expression patterns[51], epigenomics explores chromatin accessibility and regulatory element activity[52], and metabolomics measures the small molecules involved in metabolic pathways[53]. When integrated, these datasets offer a more complete and dynamic picture of cellular states, biological functions, and disease mechanisms[8–12].

The power of multi-omics lies in its ability to link molecular features across different levels, providing insight into how genetic variants lead to altered gene expression, how this affects protein function, and how changes in protein activity ultimately influence cellular processes and disease outcomes[8–12]. By integrating multi-omics data, we can identify the connections between these molecular factors, gaining a deeper understanding of the mechanisms driving disease pathology and potential therapeutic targets.

However, integrating these diverse omics data presents its own set of challenges. The data from different omics layers can vary in their type, format, and dimensionality, making it difficult to combine them into a cohesive analysis. Additionally, the relationships between these different biological features are often complex and non-linear, requiring advanced computational tools to identify meaningful patterns and interactions. As a result, the need for novel computational frameworks and analytical tools that can handle multi-omics data has become increasingly evident.

Machine Learning for Multi-Omics Data Integration

Integrating multi-omics data presents significant challenges due to the varying types, formats, and dimensionalities of data across different omics layers[8,9]. These differences make it difficult to combine them into a unified analysis. Additionally, the complex, non-linear relationships between biological features necessitate advanced computational tools to identify meaningful patterns and interactions. Machine learning (ML) techniques, including deep learning, offer powerful solutions for overcoming these obstacles[54,55]. By leveraging these tools, we can integrate genomics, transcriptomics, metabolomics, and epigenomics data, providing a unified framework for analyzing complex biological systems. This integration enables the subsequent exploration of feature interactions, helping to uncover deeper insights into disease mechanisms and biological processes.

Through ML, we can develop predictive models that reveal hidden patterns and relationships in multi-omics data, identifying how various molecular features (such as genes, proteins, and metabolites) contribute to biological outcomes. By training models on large, multi-dimensional datasets, ML algorithms can identify complex associations that might be difficult to detect using traditional statistical methods[56]. For instance, deep neural networks can learn non-linear relationships and intricate interactions between genomic, transcriptomic, metabolomic, and epigenomic features, providing a holistic understanding of cellular processes[57].

In the context of epidemiological and bioinformatics research, ML models are particularly valuable for predicting disease outcomes and identifying biomarkers[58,59]. For example, in Alzheimer's disease, ML approaches can integrate genomic, transcriptomics metabolomic, and epigenomic data to predict disease progression or identify potential therapeutic targets[60,61]. By

integrating multi-omics data, these models can identify complex biological signatures that are associated with disease phenotypes, offering more accurate predictions than single-omics approaches.

Additionally, ML can help identify important feature interactions within multi-omics data.[62,63] These interactions, whether synergistic or antagonistic, are often crucial in understanding the underlying mechanisms of diseases. For example, certain gene mutations might have more significant effects when coupled with specific changes in protein expression or metabolite levels. Moreover, ML can assess the importance of individual features, highlighting those that have the greatest impact on disease outcomes[64,65]. Through techniques like feature importance ranking and interaction modeling, ML can uncover these critical relationships, providing a deeper understanding of disease mechanisms.

Despite the potential of integrating multi-omics data with machine learning, several challenges remain. The first challenge is the integration of multi-omics data itself. Different omics layers vary significantly in their structure, scale, and dimensionality. These differences make it difficult to combine the data into a cohesive model that can be easily interpreted by machine learning algorithms. Existing methods, like Cooperative Learning[66], DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays[67], and MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data[68], have been used to address this issue by representing each omic layer in a unified form. However, these techniques often struggle to fully capture complex, non-linear relationships across different data types, limiting their ability to model the intricate interactions that drive disease mechanisms. The second challenge is model interpretability. While machine learning models can make accurate

predictions, understanding the biological relationships driving those predictions is essential, particularly when working with complex, high-dimensional multi-omics data. This understanding is crucial for ensuring that the model's outcomes are not only accurate but also biologically meaningful. Existing interpretability methods, such as SHAP (Shapley Additive Explanations)[69] and LIME (Local Interpretable Model-Agnostic Explanations)[70], provide ways to explain how individual features contribute to model predictions. These techniques quantify the contribution of each feature, helping to uncover how specific genetic, epigenetic, or metabolic factors influence disease outcomes. However, these methods have limitations— SHAP is computationally expensive when dealing with large datasets and cannot be applied to complex machine learning models. Meanwhile, LIME lacks the capability to compute feature interactions, which are essential for understanding the relationships between variables in multi-omics data. Despite these challenges, both methods remain crucial in enhancing the biological relevance of the findings by providing insights into the underlying mechanisms driving predictions.

The application of ML to multi-omics data integration offers substantial promise for advancing our understanding of complex biological systems and predicting disease outcomes. However, the existing methods for integrating and interpreting multi-omics data face significant limitations, particularly in handling the heterogeneity of data and ensuring model interpretability. To fully harness the power of ML in biological research and clinical applications, it is essential to develop new methods that address these challenges, allowing for more accurate and biologically meaningful insights.

Gene Regulation and Transcription Factors in Oligodendrocytes

Oligodendrocytes are specialized glial cells within the central nervous system (CNS) responsible for the formation and maintenance of the myelin sheath, which insulates neuronal axons and

facilitates rapid electrical signaling[71–73]. This myelination process is essential for normal neural function, and defects in oligodendrocyte development or function can lead to a variety of neurological disorders, including AD[72–74]. Understanding the intricate gene regulatory networks that govern oligodendrocyte differentiation, maturation, and function is crucial for developing targeted therapies for these conditions.

Transcription factors (TFs) are proteins that bind to specific DNA sequences to regulate the expression of target genes (TGs)[75]. In oligodendrocytes, a precise balance of TF activity is required for proper gene expression during the differentiation of oligodendrocyte precursor cells (OPCs) into mature oligodendrocytes[76,77]. Several TFs have been identified as critical players in this process, including SOX10 and OLIG2, among others[78–84]. These transcription factors not only direct the expression of genes involved in oligodendrocyte differentiation but also can coordinate the expression of genes essential for myelination and maintaining oligodendrocyte function.

However, the role of transcription factors in gene regulation is not simply a matter of individual proteins acting in isolation. Increasing evidence suggests that TFs do not act alone but cooperate in complex TF-TF interactions within the regulatory regions of TGs[85,86]. This cooperation can involve direct protein-protein interactions or indirect effects through the modulation of chromatin structure and accessibility. For instance, SOX10 and OLIG2 are known to work together to regulate a set of genes critical for oligodendrocyte differentiation and myelination[78,81,87]. These cooperative interactions are critical for ensuring the temporal and spatial regulation of gene expression that supports the development and function of oligodendrocytes.

Understanding how TFs cooperate in oligodendrocytes to regulate gene expression is of great interest because disruptions in these regulatory networks can have profound effects on

oligodendrocyte function and CNS health[88,89]. Dysregulation of TF interactions can lead to defective myelination, contributing to various neurodegenerative diseases, including AD[90]. Moreover, recent studies suggest that the identification of cooperative TF pairs could provide valuable insights into the regulatory mechanisms underlying these diseases and might even offer new therapeutic targets[32,91,92].

To fully unravel these complex TF interactions, it is necessary to adopt advanced computational approaches that can integrate different types of genomic data, such as Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) and RNA sequencing (RNA-seq), to identify TF pairs that cooperate to regulate oligodendrocyte-specific genes[93,94]. These methods hold the potential to uncover new mechanisms of oligodendrocyte function and pathology, providing a deeper understanding of how transcriptional regulation contributes to CNS health and disease.

Challenges in Integrating Multi-Omics Data for Disease Prediction

Integrating multi-omics data is crucial for enhancing disease phenotype prediction, as it provides a comprehensive view of the molecular underpinnings of diseases. However, this integration is complex due to the variability in data types, formats, and scales. Genomic data provide static DNA sequences, transcriptomic data capture dynamic gene expression changes, and metabolomic data offers insights into protein abundance and metabolic activities. Standardizing these diverse data types is essential for effective integration.

Moreover, multi-omics datasets are often incomplete and noisy, due to technical limitations and variability in data collection methods[95]. Addressing missing values and minimizing noise through advanced imputation techniques and robust preprocessing methods is critical for accurate

analysis. The high dimensionality of these datasets, where the number of features often exceeds the number of samples, also increases the risk of overfitting[96]. Employing dimensionality reduction techniques and ensuring access to high-performance computing resources are essential for managing these challenges, as they help reduce computational complexity and enable the extraction of key features, such as using embeddings to represent high-dimensional data in a lower-dimensional space for more efficient and accurate analysis[97,98].

Understanding the complex relationships between the features in omics data is difficult due to the non-linear and multifaceted nature of biological systems[9]. To address this, sophisticated computational models are needed to capture these complexities and uncover meaningful patterns. Additionally, while ML models are effective in identifying patterns within complex datasets, their "black box" nature often limits interpretability[99]. Improving model transparency is particularly important in clinical settings, where understanding the rationale behind predictions can guide decision-making.

To tackle these challenges, novel computational frameworks and analytical tools are needed to effectively integrate diverse omics data. Overcoming these obstacles will significantly enhance disease phenotype predictions and help identify potential therapeutic targets.

Application of Multi-Omics and Machine Learning in Alzheimer's Disease

Despite significant research efforts, the etiology and pathogenesis of AD remain poorly understood. A deeper understanding of the molecular factors involved in the disease is essential to identify potential biomarkers for early diagnosis. The integration of multi-omics data and machine learning approaches offers promising strategies to reveal these underlying mechanisms and provide new insights into AD progression[10–12].

In the context of AD, the integration of multi-omics data with clinical information—such as cognitive assessments and clinical outcomes—further enhances the accuracy of predictive models. This holistic approach enables more personalized disease predictions and better-targeted therapeutic strategies[100,101]. By leveraging multi-omics data, machine learning algorithms can identify key molecular signatures, provide early diagnostic biomarkers, and predict disease phenotypes such as cognitive decline. This integrated strategy holds significant potential to advance our understanding of AD and improve patient care.

Bridging Biological Insights with Computational Approaches

While multi-omics technologies and ML have provided valuable insights into the molecular mechanisms of complex diseases, integrating these tools to deepen our understanding of disease biology remains a significant challenge. In this dissertation, I develop novel computational methods to enhance the integration of multi-omics data, specifically focusing on AD phenotype prediction. The second aim (Aim 2) introduces Cooperative Multi-view Integration and Scalable Interpretable Model Explainer (COSIME), a machine learning method that integrates multi-view omics data to improve disease phenotype prediction and assess feature importance and interactions, both within-view and across-view. The third aim (Aim 3) applies COSIME to identify key biomarkers and interactions between genomics and metabolomics, providing insights into AD risk factors and enhancing the understanding of disease progression. Collectively, these efforts advance bioinformatics and contribute to a systems-level understanding of the biological mechanisms underlying AD.

The integration of multi-omics data remains a significant challenge in understanding the molecular complexity of diseases like AD. This dissertation tackles this issue by developing ML

models that process diverse omics data types, including genomics, transcriptomics, and epigenomics. These models are designed to capture complex relationships between molecular layers, improving the accuracy of AD phenotype predictions and providing deeper insights into the disease's molecular mechanisms.

A key focus of this research is the development of algorithms that can effectively address challenges in multi-omics data, such as noise and high-dimensionality. The proposed computational frameworks integrate heterogeneous omics data while ensuring meaningful and interpretable biological insights. These ML methods not only enhance prediction accuracy but also help us better understand the regulatory dynamics of AD.

This dissertation also investigates the regulatory mechanisms in AD, specifically the role of transcription factors in gene regulation, through the integration of single-cell gene regulatory networks. By leveraging ML models to analyze chromatin accessibility and gene expression, the work delves into how these regulatory dynamics contribute to disease progression and offer new perspectives on genetic-environmental interactions in AD risk.

Ultimately, this research advances our understanding of AD by providing novel bioinformatics tools that integrate multi-omics data, enabling more accurate analysis of the disease's underlying biological mechanisms and identifying potential therapeutic targets.

Specific Aims

Studies of brain diseases, such as AD have implicated white matter alterations in their pathogenesis, and models of cell types, for instance, oligodendrocyte defects, have validated mechanistic connections with cognitive disorders[102]. Recent technological advances have enabled gene regulation detection through multi-omics (i.e., genomics, transcriptomics, proteomics). Particularly, emerging next-generation sequencing (NGS) technologies, such as single-cell RNA sequencing (scRNA-seq), allow us to study functional genomics and gene regulation at the cell-type level. Large collaborative projects, such as PsychENCODE[103] and Psych-AD[104], generate single-cell multi-omics data aiming to understand the molecular mechanisms of brain diseases, including AD. Metabolomics technology has emerged as a tool for studying small molecules influenced by factors, such as genetics and disease processes. Some plasma and cerebrospinal fluid (CSF) metabolites were identified as associated with AD[105,106] and the relationships between genomics, metabolomics, and AD phenotypes need to be comprehensively understood.

Our long-term goal is to integrate and analyze large-scale multi-omics data at the population level to shed light on complex relationships involved in AD pathology. The overall objective of this research is to integrate multi-omics data, such as scRNA-seq and single-cell Assay for Transposase-Accessible Chromatin sequencing (scATAC-seq), to identify the activity of the co-binding TFs in gene regulatory networks (GRNs). We then aim to develop a new flexible machine learning model that optimally integrates multi-omics data to accurately select AD risk factors and predict AD phenotypes. Lastly, we aim to apply the new method to AD epidemiological multi-omics data to identify novel AD risk factors. Our central hypothesis is that functional genetic variants, metabolites, and their interactions will influence changes in pre-clinical AD biomarkers and cognitive function. The rationale for our proposed research is that integration of genomic and

metabolomic data will enable the identification and statistical modeling of the complex interplay of genes and metabolites involved in AD pathology, which is necessary to achieve the goal of precision medicine for AD. We will test our central hypothesis by executing the following aims:

Aim 1: **Characterize the roles of cooperative transcription factor (TF) pairs in regulating target genes within the gene regulatory networks (GRNs) of oligodendrocytes, utilizing multi-omics data and deep learning approaches.**

H1a: scATAC-seq data will be used to identify co-binding TF pairs in oligodendrocyte-specific regulatory regions, and deep learning models trained on scRNA-seq data will predict the expression of target genes regulated by these cooperative TF pairs.

H1b**:** eQTLs will independently validate the regulatory roles of cooperative TF pairs, confirming their involvement in the regulation of genes associated with oligodendrocyte function and neurodegenerative diseases, such as AD.

Aim 2: Develop COSIME: Cooperative Multi-view Integration and Scalable Interpretable Model Explainer with applications to AD.

H2a: Developing a predictive model using an unsupervised neural network combined with learnable optimal transport methods to integrate multi-omics data will enhance the accuracy of AD phenotype prediction.

H2b: COSIME will provide interpretable outputs by identifying feature importance and elucidating feature interactions within and across different omics modalities when applied to AD phenotypes.

Aim 3: Identify genomic and metabolomic AD risk factors and their interaction effects on AD phenotypes by applying COSIME to data from the Wisconsin Registry for Alzheimer's Prevention (WRAP) and Wisconsin Alzheimer's Disease Research Center (ADRC) cohorts. H3: Genes and metabolites will be identified and replicated as potential AD risk factors by integrating multi-omics data using COSIME.

Through the proposed study, we expect to have an important positive impact because the identification of novel AD biomarkers, such as co-binding transcription factors, metabolic profiles, and interactions between those, using advanced bioinformatics tools will help understand AD pathology. This research will help prevent, diagnose, and treat dementia, and lays the groundwork for precision medicine.

Significance

scRNA-seq technology is a recently emerging tool to study functional genomics and gene regulation. Metabolomics technology is another tool that has emerged to study metabolites, small molecules influenced by factors, such as genetics and disease processes. Multi-omics offers a holistic view of human health and disease by providing an integrative perspective across multiple levels of biology (e.g., predicting GRNs and disease phenotype predictions). While several studies have identified transcription factors[107,108] or metabolites[109,110] associated with AD status, only a few focused on preclinical AD phenotypes[111,112], which may give insight into the pathophysiology of AD, point to therapeutic targets and discover genes and metabolites that can be utilized as early AD biomarkers.

The Wang lab has developed a computational pipeline of integrative multi-omics analyses for predicting cell-type specific disease genes and GRNs and a machine learning analysis found that cell-type specific disease genes improved clinical phenotype predictions, including those for AD.[113] The Engelman lab has integrated genetic, lifestyle, metabolomic, and AD biomarker data to better understand these associations[106]. What remains unclear is the comprehensive understanding of functional genomics and gene regulation, metabolomics, and their interactions in AD pathology. Developing and applying flexible and robust bioinformatics tools to optimally integrate multi-omics data and predict preclinical AD phenotypes is critical to prevent, diagnose, and treat AD.

This research will be significant because the identification of novel AD biomarkers, such as transcription factors, metabolic profiles, and interactions between those using advanced bioinformatics tools will help understand AD pathology. Machine learning approaches, including

deep learning, are expected to lead to identifying new pathways for targeting therapeutic agents and novel risk prediction and diagnostic tools for early AD pathology.

Innovation

The proposed research has several innovative aspects. First, it focuses on revealing the specific roles of co-binding TFs at regulatory elements and characteristics of regulatory hub enhancers. This is very novel because even though there are some studies about single-cell functional genomic research for oligodendrocytes—one of the major cell types for the central nervous system related to AD—as far as we know, no earlier studies predicting GRNs, TFs, gene-regulatory elements, and target genes using oligodendrocytes for AD exist.

The second unique aspect is multi-omics data integration and phenotype prediction. We are planning to develop a new machine learning model that will be flexible and suitable for handling the multi-omics data we use for phenotype prediction. Moreover, using a latent vector to find the best optimal transport plan to integrate multi-omics data and predict phenotypes is a new approach as far as we are aware.

Third, there are many AD phenotypes that are related to AD continuum categories of NIA-AA's A/T/N biomarker profiles[114] in WRAP. Using cognitive functions as a proxy for preclinical AD diagnosis and ptau217 as a proxy for amyloid deposition will give us a relatively large sample size.

Lastly, our research will discover novel pathways in AD pathology by providing novel AD risk genes and metabolites. There could be overlaps between our findings and the results in other existing studies, however, since we will consider interactions between genes and metabolites in integrating multi-omics data for machine learning phenotype prediction, we may find novel AD risk factors. Furthermore, validating our findings in WRAP in an additional cohort will strongly support our hypotheses.

References

1. Ma'ayan, A. Complex systems biology. *J. R. Soc. Interface.* **14**, 20170391 (2017).
2. Wolf, Y. I., Katsnelson, M. I. & Koonin, E. V. Physical foundations of biological complexity. *Proc. Natl. Acad. Sci. U.S.A.* **115**, (2018).
3. Adami, C., Ofria, C. & Collier, T. C. Evolution of biological complexity. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 4463–4468 (2000).
4. Chaplin, D. D. Overview of the immune response. *J Allergy Clin Immunol* **125**, S3-23 (2010).
5. Naylor, S. & Chen, J. Y. Unraveling human complexity and disease with systems biology and personalized medicine. *Per Med* **7**, 275–289 (2010).
6. Colvis, C. M. *et al.* Epigenetic mechanisms and gene networks in the nervous system. *J Neurosci* **25**, 10379–10389 (2005).
7. Poljšak, B. & Milisav, I. Clinical implications of cellular stress responses. *Bosn J Basic Med Sci* **12**, 122–126 (2012).
8. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol* **18**, 83 (2017).
9. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform Biol Insights* **14**, 1177932219899051 (2020).
10. Badhwar, A. *et al.* A multiomics approach to heterogeneity in Alzheimer's disease: focused review and roadmap. *Brain* **143**, 1315–1331 (2020).
11. Nativio, R. *et al.* An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease. *Nat Genet* **52**, 1024–1035 (2020).
12. Kapoor, M. *et al.* Multi-omics integration analysis identifies novel genes for alcoholism with potential overlap with neurodegenerative diseases. *Nat Commun* **12**, 5071 (2021).

13. Costa, V., Aprile, M., Esposito, R. & Ciccodicola, A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur J Hum Genet* **21**, 134–142 (2013).
14. Andrews, S. J. *et al.* The complex genetic architecture of Alzheimer's disease: novel insights and future directions. *eBioMedicine* **90**, 104511 (2023).
15. Kumar, A., Sidhu, J., Lui, F. & Tsao, J. W. Alzheimer Disease. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).
16. DeTure, M. A. & Dickson, D. W. The neuropathological diagnosis of Alzheimer's disease. *Mol Neurodegeneration* **14**, 32 (2019).
17. Wang, H., Yang, F., Zhang, S., Xin, R. & Sun, Y. Genetic and environmental factors in Alzheimer's and Parkinson's diseases and promising therapeutic intervention via fecal microbiota transplantation. *npj Parkinsons Dis.* **7**, 70 (2021).
18. Killin, L. O. J., Starr, J. M., Shiue, I. J. & Russ, T. C. Environmental risk factors for dementia: a systematic review. *BMC Geriatr* **16**, 175 (2016).
19. Gerrits, E. *et al.* Distinct amyloid-β and tau-associated microglia profiles in Alzheimer's disease. *Acta Neuropathol* **141**, 681–696 (2021).
20. Griciuc, A. *et al.* Alzheimer's disease risk gene CD33 inhibits microglial uptake of amyloid beta. *Neuron* **78**, 631–643 (2013).
21. Hampel, H. *et al.* The Amyloid-β Pathway in Alzheimer's Disease. *Mol Psychiatry* **26**, 5481–5503 (2021).
22. O'Brien, R. J. & Wong, P. C. Amyloid precursor protein processing and Alzheimer's disease. *Annu Rev Neurosci* **34**, 185–204 (2011).

23. Ramanan, V. K. *et al.* GWAS of longitudinal amyloid accumulation on^{18} F-florbetapir PET in Alzheimer's disease implicates microglial activation gene *IL1RAP*. *Brain* **138**, 3076–3088 (2015).

24. Sadigh-Eteghad, S. *et al.* Amyloid-beta: a crucial factor in Alzheimer's disease. *Med Princ Pract* **24**, 1–10 (2015).

25. Bloom, G. S. Amyloid-β and Tau: The Trigger and Bullet in Alzheimer Disease Pathogenesis. *JAMA Neurol* **71**, 505 (2014).

26. Ferrari-Souza, J. P. *et al. APOE* ε4 associates with microglial activation independently of Aβ plaques and tau tangles. *Sci. Adv.* **9**, eade1474 (2023).

27. Raulin, A.-C. *et al.* ApoE in Alzheimer's disease: pathophysiology and therapeutic strategies. *Mol Neurodegeneration* **17**, 72 (2022).

28. Yin, Z. *et al.* APOE4 impairs the microglial response in Alzheimer's disease by inducing TGFβ-mediated checkpoints. *Nat Immunol* **24**, 1839–1853 (2023).

29. D'Alessandro, M. C. B., Kanaan, S., Geller, M., Praticò, D. & Daher, J. P. L. Mitochondrial dysfunction in Alzheimer's disease. *Ageing Research Reviews* **107**, 102713 (2025).

30. Cai, Q. & Tammineni, P. Mitochondrial Aspects of Synaptic Dysfunction in Alzheimer's Disease. *J Alzheimers Dis* **57**, 1087–1103 (2017).

31. Skaper, S. D., Facci, L., Zusso, M. & Giusti, P. Synaptic Plasticity, Dementia and Alzheimer Disease. *CNSNDDT* **16**, 220–233 (2017).

32. Kandimalla, R. & Reddy, P. H. Therapeutics of Neurotransmitters in Alzheimer's Disease. *J Alzheimers Dis* **57**, 1049–1069 (2017).

33. Kinney, J. W. *et al.* Inflammation as a central mechanism in Alzheimer's disease. *A&D Transl Res & Clin Interv* **4**, 575–590 (2018).

34. MRC CFAS *et al.* Microglial immunophenotype in dementia with Alzheimer's pathology. *J Neuroinflammation* **13**, 135 (2016).

35. Zhang, W., Xiao, D., Mao, Q. & Xia, H. Role of neuroinflammation in neurodegeneration development. *Sig Transduct Target Ther* **8**, 267 (2023).

36. Chen, Y. & Colonna, M. Microglia in Alzheimer's disease at single-cell level. Are there common patterns in humans and mice? *J Exp Med* **218**, e20202717 (2021).

37. Hansen, D. V., Hanson, J. E. & Sheng, M. Microglia in Alzheimer's disease. *Journal of Cell Biology* **217**, 459–472 (2018).

38. McQuade, A. & Blurton-Jones, M. Microglia in Alzheimer's Disease: Exploring How Genetics and Phenotype Influence Risk. *Journal of Molecular Biology* **431**, 1805–1817 (2019).

39. Sadick, J. S. *et al.* Astrocytes and oligodendrocytes undergo subtype-specific transcriptional changes in Alzheimer's disease. *Neuron* **110**, 1788-1805.e10 (2022).

40. Verkhratsky, A., Olabarria, M., Noristani, H. N., Yeh, C.-Y. & Rodriguez, J. J. Astrocytes in Alzheimer's Disease. *Neurotherapeutics* **7**, 399–412 (2010).

41. González-Reyes, R. E., Nava-Mesa, M. O., Vargas-Sánchez, K., Ariza-Salamanca, D. & Mora-Muñoz, L. Involvement of Astrocytes in Alzheimer's Disease from a Neuroinflammatory and Oxidative Stress Perspective. *Front Mol Neurosci* **10**, 427 (2017).

42. Cherry, J. D., Olschowka, J. A. & O'Banion, M. K. Neuroinflammation and M2 microglia: the good, the bad, and the inflamed. *J Neuroinflammation* **11**, 98 (2014).

43. Butterfield, D. A. & Halliwell, B. Oxidative stress, dysfunctional glucose metabolism and Alzheimer disease. *Nat Rev Neurosci* **20**, 148–160 (2019).

44. Ryu, W.-I. *et al.* Brain cells derived from Alzheimer's disease patients have multiple specific innate abnormalities in energy metabolism. *Mol Psychiatry* **26**, 5702–5714 (2021).

45. Malone, A. F. *et al.* Harnessing Expressed Single Nucleotide Variation and Single Cell RNA Sequencing To Define Immune Cell Chimerism in the Rejecting Kidney Transplant. *JASN* **31**, 1977–1986 (2020).

46. Barbu, M. C. *et al.* Epigenetic prediction of major depressive disorder. *Mol Psychiatry* **26**, 5112–5123 (2021).

47. Desai, N. *et al.* Temporal and spatial heterogeneity of host response to SARS-CoV-2 pulmonary infection. *Nat Commun* **11**, 6319 (2020).

48. Miles, L. A. *et al.* Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature* **587**, 477–482 (2020).

49. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

50. Theissinger, K. *et al.* How genomics can help biodiversity conservation. *Trends in Genetics* **39**, 545–559 (2023).

51. Dinh, H. Q. *et al.* Single-cell transcriptomics identifies gene expression networks driving differentiation and tumorigenesis in the human fallopian tube. *Cell Reports* **35**, 108978 (2021).

52. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**, 207–220 (2019).

53. Fessenden, M. Metabolomics: Small molecules, single cells. *Nature* **540**, 153–155 (2016).

54. Reel, P. S., Reel, S., Pearson, E., Trucco, E. & Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances* **49**, 107739 (2021).

55. Feldner-Busztin, D. *et al.* Dealing with dimensionality: the application of machine learning to multi-omics data. *Bioinformatics* **39**, btad021 (2023).

56. Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N. & Fanos, V. Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. *Medicina (Kaunas)* **56**, 455 (2020).

57. Leng, D. *et al.* A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biol* **23**, 171 (2022).

58. Ng, S., Masarone, S., Watson, D. & Barnes, M. R. The benefits and pitfalls of machine learning for biomarker discovery. *Cell Tissue Res* **394**, 17–31 (2023).

59. Park, D. J. *et al.* Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Sci Rep* **11**, 7567 (2021).

60. Chu, S. H. *et al.* Integration of Metabolomic and Other Omics Data in Population-Based Study Designs: An Epidemiological Perspective. *Metabolites* **9**, 117 (2019).

61. Horgusluoglu, E. *et al.* Integrative metabolomics-genomics approach reveals key metabolic pathways and regulators of Alzheimer's disease. *Alzheimers Dement* **18**, 1260–1278 (2022).

62. Song, W. *et al.* AutoInt: Automatic Feature Interaction Learning via Self-Attentive Neural Networks. in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* 1161–1170 (ACM, Beijing China, 2019). doi:10.1145/3357384.3357925.

63. Tsang, M., Rambhatla, S. & Liu, Y. How does this interaction affect me? Interpretable attribution for feature interactions. Preprint at https://doi.org/10.48550/ARXIV.2006.10965 (2020).

64. Rengasamy, D. *et al.* Feature importance in machine learning models: A fuzzy information fusion approach. *Neurocomputing* **511**, 163–174 (2022).

65. Saarela, M. & Jauhiainen, S. Comparison of feature importance measures as explanations for classification models. *SN Appl. Sci.* **3**, 272 (2021).

66. Ding, D. Y., Li, S., Narasimhan, B. & Tibshirani, R. Cooperative learning for multiview analysis. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2202113119 (2022).

67. Singh, A. *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **35**, 3055–3062 (2019).

68. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* **21**, 111 (2020).

69. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Preprint at https://doi.org/10.48550/ARXIV.1705.07874 (2017).

70. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. Preprint at https://doi.org/10.48550/ARXIV.1602.04938 (2016).

71. Elbaz, B. & Popko, B. Molecular Control of Oligodendrocyte Development. *Trends in Neurosciences* **42**, 263–277 (2019).

72. Ettle, B., Schlachetzki, J. C. M. & Winkler, J. Oligodendroglia and Myelin in Neurodegenerative Diseases: More Than Just Bystanders? *Mol Neurobiol* **53**, 3046–3062 (2016).

73. Simons, M. & Nave, K.-A. Oligodendrocytes: Myelination and Axonal Support. *Cold Spring Harb Perspect Biol* **8**, a020479 (2016).

74. Quan, L., Uyeda, A. & Muramatsu, R. Central nervous system regeneration: the roles of glial cells in the potential molecular mechanism underlying remyelination. *Inflamm Regener* **42**, 7 (2022).

75. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

76. Tiane, A. *et al.* From OPC to Oligodendrocyte: An Epigenetic Journey. *Cells* **8**, 1236 (2019).

77. Marton, R. M. *et al.* Differentiation and maturation of oligodendrocytes in human three-dimensional neural cultures. *Nat Neurosci* **22**, 484–491 (2019).

78. Liu, Z. *et al.* Induction of oligodendrocyte differentiation by Olig2 and Sox10: Evidence for reciprocal interactions and dosage-dependent mechanisms. *Developmental Biology* **302**, 683–693 (2007).

79. Lopez-Anido, C. *et al.* Differential Sox10 genomic occupancy in myelinating glia. *Glia* **63**, 1897–1914 (2015).

80. Pozniak, C. D. *et al.* Sox10 directs neural stem cells toward the oligodendrocyte lineage by decreasing Suppressor of Fused expression. *Proc Natl Acad Sci U S A* **107**, 21795–21800 (2010).

81. Sock, E. & Wegner, M. Using the lineage determinants Olig2 and Sox10 to explore transcriptional regulation of oligodendrocyte development. *Developmental Neurobiology* **81**, 892–901 (2021).

82. Turnescu, T. *et al.* Sox8 and Sox10 jointly maintain myelin gene expression in oligodendrocytes. *Glia* **66**, 279–294 (2018).

83. Wang, J. *et al. Olig2* Ablation in Immature Oligodendrocytes Does Not Enhance CNS Myelination and Remyelination. *J. Neurosci.* **42**, 8542–8555 (2022).

84. Yu, Y. *et al.* Olig2 targets chromatin remodelers to enhancers to initiate oligodendrocyte differentiation. *Cell* **152**, 248–261 (2013).

85. Oksuz, O. *et al.* Transcription factors interact with RNA to regulate genes. *Mol Cell* **83**, 2449-2463.e13 (2023).

86. Göös, H. *et al.* Human transcription factor protein interaction networks. *Nat Commun* **13**, 766 (2022).

87. Küspert, M., Hammer, A., Bösl, M. R. & Wegner, M. Olig2 regulates Sox10 expression in oligodendrocyte precursors through an evolutionary conserved distal enhancer. *Nucleic Acids Res* **39**, 1280–1293 (2011).

88. He, D. *et al.* Chd7 cooperates with Sox10 and regulates the onset of CNS myelination and remyelination. *Nat Neurosci* **19**, 678–689 (2016).

89. Santiago, C. & Bashaw, G. J. Transcription factors and effectors that regulate neuronal morphology. *Development* **141**, 4667–4680 (2014).

90. Meng, G. & Mei, H. Transcriptional Dysregulation Study Reveals a Core Network Involving the Progression of Alzheimer's Disease. *Front Aging Neurosci* **11**, 101 (2019).

91. Radaeva, M., Ton, A.-T., Hsing, M., Ban, F. & Cherkasov, A. Drugging the 'undruggable'. Therapeutic targeting of protein–DNA interactions with the use of computer-aided drug discovery methods. *Drug Discovery Today* **26**, 2660–2679 (2021).

92. Talukdar, P. D. & Chatterji, U. Transcriptional co-activators: emerging roles in signaling pathways and potential therapeutic targets for diseases. *Sig Transduct Target Ther* **8**, 427 (2023).

93. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

94. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**, 293–304 (2024).

95. Flores, J. E. *et al.* Missing data in multi-omics integration: Recent advances through artificial intelligence. *Front Artif Intell* **6**, 1098308 (2023).

96. Charilaou, P. & Battat, R. Machine learning models and over-fitting considerations. *World J Gastroenterol* **28**, 605–607 (2022).

97. Kabir, M. F., Chen, T. & Ludwig, S. A. A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. *Healthcare Analytics* **3**, 100125 (2023).

98. Jia, W., Sun, M., Lian, J. & Hou, S. Feature dimensionality reduction: a review. *Complex Intell. Syst.* **8**, 2663–2693 (2022).

99. Hassija, V. *et al.* Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cogn Comput* **16**, 45–74 (2024).

100. Clark, C., Dayon, L., Masoodi, M., Bowman, G. L. & Popp, J. An integrative multi-omics approach reveals new central nervous system pathway alterations in Alzheimer's disease. *Alz Res Therapy* **13**, 71 (2021).

101. Meng, L. *et al.* Multi-omics analysis reveals the key factors involved in the severity of the Alzheimer's disease. *Alz Res Therapy* **16**, 213 (2024).

102. Nasrabady, S. E., Rizvi, B., Goldman, J. E. & Brickman, A. M. White matter changes in Alzheimer's disease: a focus on myelin and oligodendrocytes. *acta neuropathol commun* **6**, 22 (2018).

103. Emani, P. S. *et al.* Single-cell genomics and regulatory networks for 388 human brains. *Science* **384**, eadi5199 (2024).

104. Psych-AD: Neuropsychiatric Symptoms in Alzheimer's Disease. https://adknowledgeportal.synapse.org/Explore/Programs/DetailsPage?Program=Psych-AD.

105. Vogt, N. M. *et al.* The gut microbiota-derived metabolite trimethylamine N-oxide is elevated in Alzheimer's disease. *Alz Res Therapy* **10**, 124 (2018).

106. Darst, B. F., Lu, Q., Johnson, S. C. & Engelman, C. D. Integrated analysis of genomics, longitudinal metabolomics, and Alzheimer's risk factors among 1,111 cohort participants. *Genetic Epidemiology* **43**, 657–674 (2019).

107. Kodam, P., Sai Swaroop, R., Pradhan, S. S., Sivaramakrishnan, V. & Vadrevu, R. Integrated multi-omics analysis of Alzheimer's disease shows molecular signatures associated with disease progression and potential therapeutic targets. *Sci Rep* **13**, 3695 (2023).

108. Kapoor, M. *et al.* Multi-omics integration analysis identifies novel genes for alcoholism with potential overlap with neurodegenerative diseases. *Nat Commun* **12**, 5071 (2021).

109. Nho, K. *et al.* Altered bile acid profile in mild cognitive impairment and Alzheimer's disease: Relationship to neuroimaging and CSF biomarkers. *Alzheimer's & Dementia* **15**, 232–244 (2019).

110. Wang, J. *et al.* Peripheral serum metabolomic profiles inform central cognitive impairment. *Sci Rep* **10**, 14059 (2020).

111. Casanova, R. *et al.* Blood metabolite markers of preclinical Alzheimer's disease in two longitudinally followed cohorts of older individuals. *Alzheimer's & Dementia* **12**, 815–822 (2016).

112. Rai, S. N. *et al.* Therapeutic Potential of Vital Transcription Factors in Alzheimer's and Parkinson's Disease With Particular Emphasis on Transcription Factor EB Mediated Autophagy. *Front. Neurosci.* **15**, 777347 (2021).

113. Jin, T. *et al.* scGRNom: a computational pipeline of integrative multi-omics analyses for predicting cell-type disease genes and regulatory networks. *Genome Med* **13**, 95 (2021).

114. Jack, C. R. *et al.* A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers. *Neurology* **87**, 539–547 (2016).

CHAPTER 2: COSIME: Cooperative multi-view integration and Scalable and Interpretable Model Explainer

COSIME: Cooperative multi-view integration and Scalable and Interpretable Model Explainer

Jerome J. Choi[1,2], Noah Cohen Kalafut[2,4], Tim Gruenloh[3], Corinne D. Engelman[1], Tianyuan Lu[1,3,◇] & Daifeng Wang[2,3,4,◇,*]

[1]*Department of Population Health Sciences, University of Wisconsin-Madison School of Medicine and Public Health*
[2]*Waisman Center, University of Wisconsin-Madison*
[3]*Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison School of Medicine and Public Health*
[4]*Department of Computer Sciences, University of Wisconsin-Madison*

[◇]*Joint senior authors.*
[*]*Corresponding author. daifeng.wang@wisc.edu*

Abstract

Single-omics approaches often provide a limited view of complex biological systems, whereas multi-omics integration offers a more comprehensive understanding by combining diverse data views. However, integrating heterogeneous data types and interpreting the intricate relationships between biological features—both within and across different data views—remains a bottleneck. To address these challenges, we introduce COSIME (Cooperative Multi-view Integration and Scalable Interpretable Model Explainer). COSIME uses backpropagation of Learnable Optimal Transport (LOT) to deep neural networks, enabling the learning of latent features from multiple views to predict disease phenotypes. In addition, COSIME incorporates Monte Carlo sampling to efficiently estimate Shapley values and Shapley-Taylor indices, enabling the assessment of both feature importance and their pairwise interactions—synergistically or antagonistically—in predicting disease phenotypes. We applied COSIME to both simulated data and real-world datasets, including single-cell transcriptomics, single-cell spatial transcriptomics, epigenomics, and metabolomics, specifically for Alzheimer's disease-related phenotypes. Our results demonstrate that COSIME significantly improves prediction performance while offering enhanced interpretability of feature relationships. For example, we identified that synergistic interactions between microglia and astrocyte genes associated with Alzheimer's

disease are more likely to be active at the edges of the middle temporal gyrus as indicated by spatial locations. Finally, COSIME is open-source and available for general use.

Introduction

Single-omics approaches, while valuable for providing insights into individual biological layers, often provide a limited view of complex biological systems. Focusing on a single molecular layer—such as genomics, transcriptomics, or metabolomics—provides a limited view of cellular processes, as these layers do not function in isolation. Multi-view data integration overcomes this limitation by combining information from multiple omic layers, offering a more holistic view of biology. This comprehensive approach allows us to identify novel biomarkers, understand the underlying biology of diseases, and improve predictions for disease phenotypes[1, 2]. Machine learning algorithms can efficiently process and analyze large-scale multi-view data, identifying patterns and relationships that are often complex for traditional methods to discern. Moreover, in biological systems, features—especially those from different omic layers—may interact in complex ways that jointly influence phenotypes. Machine learning can help overcome these challenges by modeling and quantifying complex feature interactions, including those across different omic layers.

There are a few studies that provide multi-omics data integration and prediction. Cooperative learning[3] is a supervised learning method with multiple sets of features and it combines the usual squared-error loss of predictions with an agreement penalty that encourages input datasets to be similar. However, the strength of the agreement penalty is controlled by a fixed parameter, which is typically set through a tuning, and is not learned dynamically during training. This means that the model does not deeply integrate features in a way that fully captures their complex interdependencies. Moreover, this approach is less flexible in adapting to complex cross-view relationships since it does not learn interactions end-to-end during training. Data Integration Analysis for Biomarker discovery using Latent components (DIABLO)[4] is a supervised learning method for multi-omics integration that seeks for common information across different data types through the selection of a subset of molecular features. Nonetheless, DIABLO assumes that the different omic layers share some level of homogeneity in terms of their relationship to the outcome. In practice, this may not always hold, especially when the omic views measure fundamentally different biological processes that are not strongly linked to each other. Multi-Omics Factor Analysis v2 (MOFA+)[5] is a method for multi-omics integration based on factor analysis and assumes that the latent structure of data can be captured with linear models, designed to analyze and integrate multiple types of omics data in an unsupervised fashion.

However, MOFA+ does not have a built-in mechanism for analyzing feature interactions or for providing direct insights into how individual features contribute to the final prediction. Importantly, none of these methods are designed to handle non-linear relationships or feature interactions. Non-linear interactions are especially relevant in biological data, where the effects of one variable (e.g., a gene) on an outcome (e.g., disease) might not be constant and might vary depending on other factors. Furthermore, deep learning and optimal transport approaches have not been as widely adopted in multi-omics integration as traditional statistical methods.

Several approaches have been developed for the interpretability of machine learning models. SHAP (SHapley Additive exPlanations)[6] provides functions to compute feature importance for a particular prediction in machine learning models. While SHAP has optimized algorithms for certain types of models (like Tree explainer for decision trees), it can still be computationally expensive, especially for large datasets and complex models. Moreover, the feature interaction matrix is only available as an output for tree-based models. Additionally, dependence plots display the relationship between only two features, making it difficult to capture the overall patterns of feature interactions in predictions. Local Interpretable Model-agnostic Explanations (LIME)[7] offers localized interpretability, ideal for understanding individual predictions in simple models. LIME works by approximating the model locally around a specific data point using a simple interpretable model, such as linear regression. This means that LIME generates explanations for individual predictions rather than providing a global understanding of the model's behavior. It may miss patterns that are important across the entire dataset, leading to misleading or incomplete explanations if the behavior of the model varies significantly across the feature space. Additionally, LIME does not account for feature interactions, limiting its ability to capture more complex relationships between features.

To tackle these challenges, this study introduces Cooperative Multiview Integration and Scalable and Interpretable Model Explainer (COSIME). COSIME features two key components. First, it integrates multi-view data leveraging deep neural network encoders (deep encoders) and Learnable Optimal Transport (LOT) techniques, combining both unsupervised and supervised learning; Second, COSIME implements a mechanism for assessing feature importance within each view, as well as quantifying both within-view and across-view interactions by estimating Shapley values and Shapley-Taylor indices. Through extensive evaluations, we demonstrate the utility of COSIME using both simulated and real-world multi-view datasets. On simulated data, we assessed binary classification and continuous prediction tasks under varying signal levels, using both early and late fusion strategies[8]. For real-world applications, COSIME was applied to Alzheimer's disease diagnosis using transcriptomics-metabolomics and transcriptomics-epigenomics datasets, as well as to predicting Alzheimer's disease progression scores using transcriptomics-spatial

transcriptomics data. These analyses highlight the flexibility of COSIME, offering a comprehensive solution for multi-view data analysis.

Results

Overview of COSIME

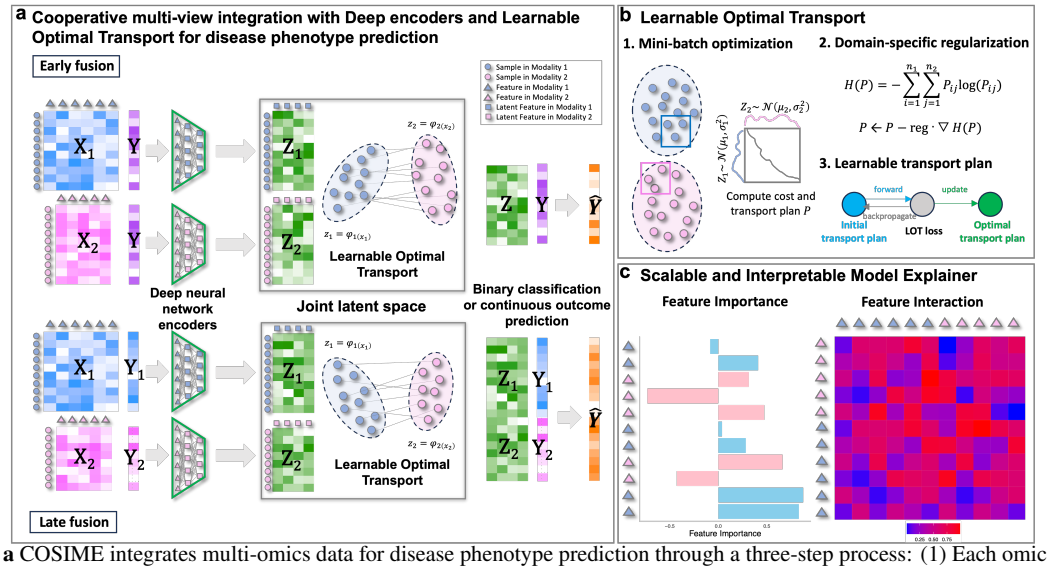
An overview of COSIME is illustrated in **Fig. 1**. The first component involves integrating multi-view data for disease phenotype prediction by leveraging deep learning-based encoders, enabling the model to address both linear and non-linear relationships, offering superior flexibility compared to traditional methods that typically rely on linear assumptions. COSIME effectively captures the complex, multi-layered interactions between different omic modalities—such as transcriptomics and epigenomics, transcriptomics and spatial transcriptomics, and transcriptomics and metabolomics—while preserving the distinct features of each data type. Learnable Optimal Transport (LOT) offers three main advantages over conventional alignment methods. First, it uses a learnable transport plan, which allows for dynamic adaptation to complex, heterogeneous data distributions, ensuring more accurate and flexible alignment. Second, LOT supports mini-batch processing, making it scalable for large datasets while maintaining computational efficiency. Finally, LOT can incorporate domain-specific regularization, enabling the model to better handle view-specific differences and improve alignment in specialized contexts. These features make LOT more robust than traditional methods, which often struggle with data misalignment and view-specific challenges.

The second component of COSIME focuses on computing feature importance values for each view, as well as pairwise interaction values for both within-view and across-view interactions. To achieve this, COSIME employs the Scalable and Interpretable Model Explainer, which leverages the Shapley values and Shapley-Taylor indices[9] to compute both feature importance and pairwise feature interactions, respectively. By applying Monte Carlo sampling and batch processing, COSIME efficiently estimates feature importance and pairwise feature interactions, enabling faster and more scalable computation while enhancing model interpretability. Additionally, it allows for the identification of the directionality of interactions—whether they exhibit synergism (complementary effects) or antagonism (conflicting effects)—offering deeper insights into how features interact in the model.

COSIME enhances the prediction accuracy by improving the modeling of across-view relationships within a shared latent space. It effectively integrates heterogeneous data, enabling more accurate predictions across different biological layers. Moreover, it also provides a powerful interpretability by identifying feature

importance and interactions, allowing for deeper insights into how individual features and their combinations contribute to the model's predictions. While these two main components of COSIME can be used together for both enhanced prediction and interpretability, they are also designed to be used independently depending on the specific needs of the analysis. This combination paves the way for identifying key biomarkers and understanding disease mechanisms. By integrating multi-omics data and assessing feature relationships, COSIME addresses the challenges of heterogeneous data and complex feature interactions, paving the way for identifying key biomarkers and understanding disease mechanisms.

Fig 1: Cooperative multi-view integration and Scalable and Interpretable Model Explainer (COSIME).



a COSIME integrates multi-omics data for disease phenotype prediction through a three-step process: (1) Each omic dataset is passed through separate deep neural network encoders (deep encoders). (2) Learnable Optimal Transport (LOT) aligns and merges these features into a joint latent space. (3) The integrated latent representation is then used to predict disease phenotypes. **b** LOT integrates heterogeneous datasets by offering three key advantages: (1) a learnable transport plan that adapts during training, (2) mini-batch processing for scalability, and (3) domain-specific regularization to enhance alignment between source and target distributions. **c** Scalable and Interpretable Model Explainer interprets model predictions by identifying feature importance and interactions. It quantifies both individual feature attributions and pairwise feature interactions, revealing synergistic or antagonistic effects on the model's output.

In the following sections, the performance metrics were compared across different methods using the mean and ±1.96 times the standard deviation from 5-fold cross-validation. For the best-trained models across the different multi-view datasets, we computed feature importance for each data view and interaction

values both within and across views. All results for the multi-view predictions using COSIME models and other methods are provided in **Supplementary Data 1** and **2**.

Simulation study

Multi-view data were generated by different signal levels (high and low) and types of outcome variables (binary and continuous).

Compared to three benchmarking methods—Cooperative Learning (CL)[3], Data Integration Analysis for Biomarker discovery using Latent components (DIABLO)[4], and Multi-Omics Factor Analysis v2 (MOFA+)[5] with logistic regression, COSIME early fusion performed best for binary outcome classification with high-signal multi-view datasets (AUROC: 0.845 ± 0.026, AUPRC: 0.854 ± 0.029, accuracy: 0.754 ± 0.054). COSIME late fusion (AUROC: 0.828 ± 0.021, AUPRC: 0.853 ± 0.021, accuracy: 0.761 ± 0.019) outperformed the three benchmarking methods (**Fig. 2a**). For binary outcome classification with low-signal multi-view datasets, COSIME late fusion achieved the best performance (AUROC: 0.737 ± 0.017, AUPRC: 0.736 ± 0.025, accuracy: 0.640 ± 0.052), while COSIME early fusion (AUROC: 0.615 ± 0.038, AUPRC: 0.662 ± 0.039, accuracy: 0.573 ± 0.027) outperformed CL, DIABLO, and MOFA+ with logistic regression (**Fig. 2a**).

We computed feature interaction values for each view using COSIME early fusion for binary outcome prediction with high-signal multi-view datasets. Pairwise feature interactions were categorized into three groups: (Group 1) the interaction terms artificially introduced during data generation, (Group 2) interactions involving latent features, and (Group 3) other interactions. The scaled pairwise feature interaction values for the first 50 features are shown in **Fig. 2b**. Pairwise Wilcoxon rank-sum two-sided tests were performed between these three groups, and the results demonstrate that COSIME successfully captured significant differences in interaction values across the groups. In view A, Group 1 exhibits significantly higher interaction values than both Group 2 (Pairwise Wilcoxon rank-sum two-sided test p–value < 0.05) and Group 3 (Pairwise Wilcoxon rank-sum two-sided test p–value < 0.0001). Additionally, Group 2 shows significantly higher interaction values compared to Group 3 (Pairwise Wilcoxon rank-sum two-sided test p–value < 0.0001). Similarly, in view B, Group 1 shows significantly higher interaction values than both Group 2 (Pairwise Wilcoxon rank-sum two-sided test p–value < 0.001) and Group 3 (Pairwise Wilcoxon rank-sum two-sided test p–value < 0.0001), while Group 2 also has significantly higher interaction values than Group 3 (Pairwise Wilcoxon rank-sum two-sided test p–value < 0.0001).

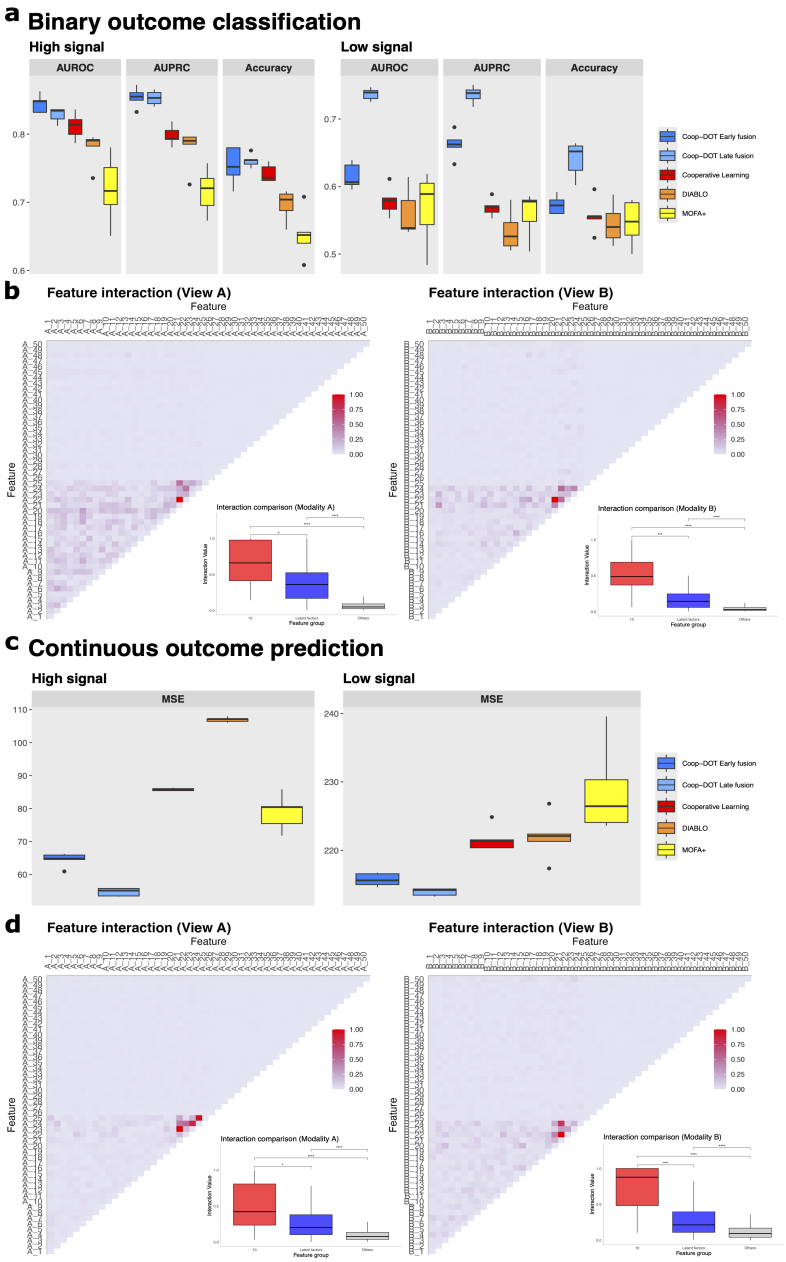
For continuous outcome prediction with high-signal multi-view datasets, COSIME delivered the most

accurate results (MSE: 54.701 ± 2.391). In comparison, COSIME early fusion outperformed CL, DIABLO, and MOFA+ with regression (MSE: 64.490 ± 4.130) (**Fig. 2c**). When applied to low-signal multi-view datasets, COSIME early fusion achieved the best performance (MSE: 215.707 ± 1.833), while COSIME late fusion showed superior results over the benchmarking methods (MSE: 218.425 ± 1.099) (**Fig. 2c**).

Feature interaction values for each view were also calculated using COSIME early fusion for continuous outcome prediction with high-signal data. The pairwise feature interactions were divided into three categories: (Group 1) the interaction terms artificially added during data generation, (Group 2) interactions involving latent features, and (Group 3) all other interactions. The scaled pairwise feature interaction values for the first 50 features are shown in **Fig. 2d**. Pairwise Wilcoxon rank-sum two-sided tests were conducted between the three groups, and the findings confirm that COSIME effectively identified significant differences in interaction values across these categories. In view A, Group 1 showed significantly higher interaction values than both Group 2 (Pairwise Wilcoxon rank-sum two-sided test p–value < 0.05) and Group 3 (Pairwise Wilcoxon rank-sum two-sided test p–value < 0.0001), with Group 2 also having significantly higher interaction values than Group 3 (Pairwise Wilcoxon rank-sum two-sided test p–value < 0.0001). In view B, Group 1 displayed significantly higher interaction values than both Group 2 (Pairwise Wilcoxon rank-sum two-sided test p–value < 0.001) and Group 3 (Pairwise Wilcoxon rank-sum two-sided test p–value < 0.0001), while Group 2 also had significantly higher interaction values compared to Group 3 (Pairwise Wilcoxon rank-sum two-sided test p–value < 0.0001).

Overall, both COSIME early fusion and late fusion models outperformed the benchmarking methods in predicting the outcomes. Additionally, COSIME accurately captured the pairwise feature interactions purposely introduced during data generation, validating its ability to identify interactions present in the data.

Fig 2: Prediction performance and feature interaction from simulated data.



a Prediction performance for binary outcome classification using COSIME early fusion, COSIME late fusion, CL, DIABLO, and MOFA+ with logistic regression on both high-signal and low-signal multi-view datasets. **b** Heatmaps depicting the pairwise interactions of the first 50 features from views A and B for binary outcome prediction, along with box plots and pairwise Wilcoxon rank-sum two-sided tests for three interaction groups: (Group 1) 10 interaction terms artificially introduced during data generation, (Group 2) 1,790 interactions involving latent features, and (Group 3) 3,150 other interactions. The box plots display the distribution of interaction values, including the first quartile, median, and third quartile. Asterisks indicate statistical significance: * for p–value < 0.05 and **** for p–value < 0.0001. **c** Prediction performance for continuous outcome prediction using COSIME early fusion, COSIME late fusion, CL, DIABLO, and MOFA+ with regression on both high-signal and low-signal multi-view datasets. **d** Heatmaps depicting the pairwise interactions of the first 50 features from views A and B for continuous outcome prediction, along with box plots and pairwise Wilcoxon rank-sum two-sided tests for three interaction groups: (Group 1) 10 interaction terms artificially introduced during data generation, (Group 2) 1,790 interactions involving latent features, and (Group 3) 3,150 other interactions. The box plots display the distribution of interaction values, including the first quartile, median, and third quartile. Asterisks indicate statistical significance: *** p–value < 0.001 and **** for p–value < 0.0001.

Classifying cognitive diagnosis for Alzheimer's disease from transcriptomic (astrocytes) and metabolomic data

Multi-view data from scRNA-seq of astrocytes and metabolomics were used to test the prediction performance of COSIME models. COSIME early fusion with matched samples (n=2,286) achieved the best performance for binary classification of Alzheimer's disease (AD) cognitive diagnosis using multi-view datasets (astrocytes and metabolomics) (AUROC: 0.842 ± 0.011, AUPRC: 0.864 ± 0.019, accuracy: 0.773 ± 0.018). To leverage all available samples (4292 metacells in transcriptomics and 2286 samples in metabolomics) across views, we also applied COSIME late fusion for unmatched samples, which yielded comparable results (AUROC: 0.804 ± 0.007, AUPRC: 0.829 ± 0.014, accuracy: 0.732 ± 0.016) (**Fig. 3a**).

Feature importance values were used to prioritize features. The top 20 most important genes (based on absolute values) from the transcriptomic data are displayed in **Fig. 3b**, and enrichment analysis reveals that these genes are strongly associated with AD. Notably, processes such as long-term synaptic potentiation, amyloid-beta ($A\beta$) formation, and amyloid precursor protein (APP) catabolism are strongly implicated in AD. These processes that involve synaptic function, $A\beta$ production, and APP metabolism represent crucial aspects of AD pathophysiology[10, 11].

Fig 3c shows the top 20 most important metabolites (ranked by absolute feature importance values)

from the metabolomic data. Enrichment analysis of these metabolites reveals metabolic pathways such as pantothenate and CoA biosynthesis, arginine biosynthesis, nicotinate and nicotinamide metabolism, and histidine metabolism. These pathways may contribute to AD by exacerbating the brain's energy deficit, promoting oxidative damage, and driving neuroinflammation[12].

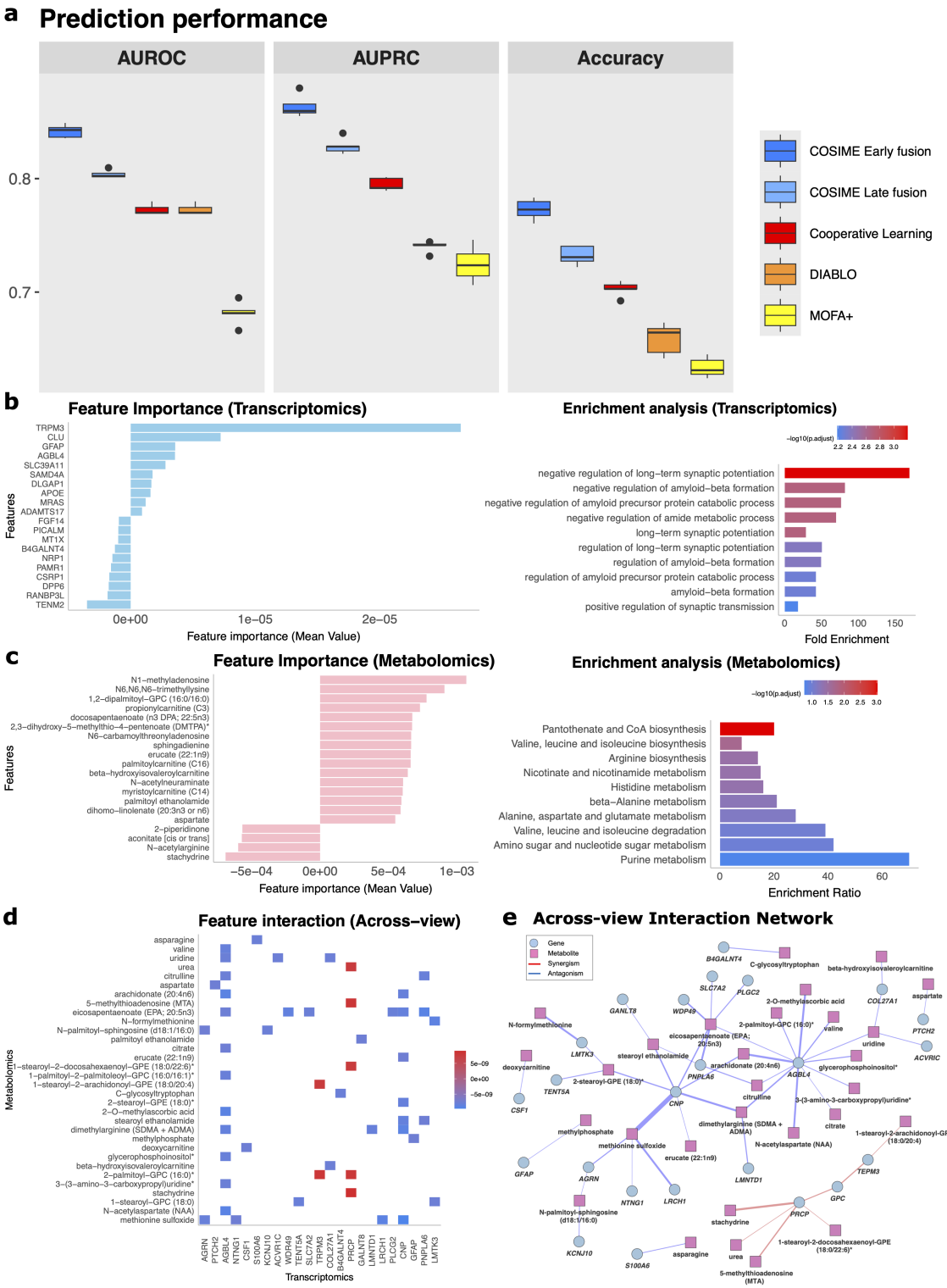
Pairwise feature interaction values were computed, and the top 50 across-view feature interaction values (based on absolute feature interaction values) are shown in **Fig. 3d**. *PRCP* has synergistic effects with metabolites such as urea and 5-methylthioadenosine (MTA). *PRCP* is involved in peptide degradation, particularly those peptides related to inflammatory responses [13, 14]. Urea is a byproduct of protein metabolism, typically excreted via the kidneys. It plays a role in nitrogen metabolism and helps in the detoxification of ammonia in the brain [15, 16]. MTA is a byproduct of the methionine salvage pathway and plays a role in maintaining cellular methylation capacity. It is involved in the recycling of S-adenosylmethionine, which is crucial for methylation reactions, including those related to neurotransmitters and DNA [17]. Erucate (22:1n9) is a monounsaturated fatty acid, which can influence lipid metabolism, and metabolic homeostasis. The synergism between *PRCP* and MTA suggests a relationship between peptide degradation, methylation capacity, and inflammatory pathways. Given that *PRCP* is involved in degrading inflammatory peptides and MTA plays a role in maintaining methylation for neurotransmitter synthesis and DNA methylation, this interaction may indicate a broader involvement of *PRCP* in neuroinflammatory processes and cellular maintenance in AD. Similarly, the interaction between *PRCP* and Erucate (22:1n9) links peptide degradation with lipid metabolism. Erucate (22:1n9), a key fatty acid in lipid metabolism, may interact with *PRCP* to modulate cellular homeostasis, suggesting that *PRCP* might influence both neuroinflammatory and lipid metabolic pathways critical for maintaining brain cell integrity in AD.

AGBL4 shows antagonistic effects with several metabolites such as arachidonate (20:4n6), 2-O-methylascorbic acid, and N-acetylaspartate (NAA) (**Fig. 3d**). *AGBL4* encodes a protein that is involved in regulating actin dynamics and cell signaling, particularly in processes such as endocytosis, vesicle trafficking, and cell motility[18]. Arachidonate (20:4n6) is a polyunsaturated fatty acid and is a precursor for various eicosanoids, including prostaglandins, leukotrienes, and thromboxanes. These molecules play crucial roles in inflammation, immune responses, and cell signaling[19]. 2-O-methylascorbic acid is a derivative of ascorbic acid (vitamin C), and is known for its antioxidant properties[20]. NAA is a metabolite primarily found in neurons and is considered a marker of neuronal integrity and function[21, 22]. *AGBL4* and arachidonate (20:4n6) may exhibit antagonistic interactions, as *AGBL4* may disrupt pro-inflammatory pathways triggered by arachidonate (20:4n6). By inhibiting actin remodeling and vesicular trafficking, which are essential for eicosanoid function, *AGBL4* may affect immune responses and inflammation, pro-

cesses central to neurodegeneration. The antagonism between *AGBL4* and 2-O-methylascorbic acid may arise from the involvement of *AGBL4* in promoting cellular processes such as inflammation and oxidative stress, which counteract the antioxidant effects of 2-O-methylascorbic acid. *AGBL4* may influence actin dynamics or cellular responses in a way that increases the need for antioxidants or reduces their efficacy. Additionally, the antagonistic relationship between *AGBL4* and NAA may reflect the impact of *AGBL4* on neuronal integrity. *AGBL4* may alter pathways involved in NAA synthesis or utilization, potentially impairing neuronal health and neurotransmitter balance, as evidenced by reduced NAA levels. Collectively, these interactions may contribute to neurodegenerative processes.

Furthermore, *PRCP* and *AGBL4* may play an important role as hub genes in the across-view network (**Fig. 3e**). These two genes are likely involved in astrocyte function under both normal and pathological conditions, particularly through their roles in inflammation, actin remodeling, and neurovascular interactions[23, 24]. As hub genes, they may influence across-view networks, connecting pathways that involve peptide signaling, cytoskeletal dynamics, neuroinflammation, and synaptic plasticity. These interactions are crucial for understanding the role of astrocytes in diseases such as AD, where neuroinflammation, astrocyte reactivity, and synaptic dysfunction are key features[25, 26]. Astrocyte reactivity, often triggered by neuroinflammation, is associated with the increased production of pro-inflammatory cytokines like TNF-α, IL-1β, and IL-6, all of which can worsen neuronal loss and synaptic dysfunction[27, 28]. In this context, *AGBL4* might shift the balance toward a more damaging inflammatory state, thereby increasing the risk of AD.

Fig 3: Classifying cognitive diagnosis for Alzheimer's disease from transcriptomic (astrocytes) and metabolomic data.



a Prediction performance using COSIME early fusion, COSIME late fusion, CL, DIABLO, and MOFA+ with logistic regression. **b** Bar plots showing the top 20 prioritized genes (absolute feature importance values) and enrichment analysis for those genes. **c** Bar plots showing the top 20 prioritized metabolites (absolute feature importance values) and enrichment analysis for those metabolites. **d** Heatmap showing the top 50 pairwise across-view (gene-metabolite) interactions. **e** Across-view interaction network showing relationships between the top 50 pairwise across-view (gene-metabolite) interactions.

Classifying cognitive status (dementia) from transcriptomic (oligodendrocyte) and epigenomic (oligodendrocyte) data

Multi-view data from both scRNA-seq and scATAC-seq of oligodendrocytes were utilized to evaluate the predictive performance of COSIME models. COSIME early fusion performed the best for binary classification of cognitive status (dementia) using multi-view datasets (oligodendrocytes) (AUROC: 0.874 ± 0.011, AUPRC: 0.805 ± 0.013, accuracy: 0.794 ± 0.041) (**Fig. 4a**).

We investigated pairwise interactions between the ten known key TFs involved in oligodendrocytes[29]. For instance, synergistic interactions were identified between SOX10 and MAZ as well as between MYRF and MAZ. Specifically, SOX10 is critical for oligodendrocyte differentiation and can be involved in reactive gliosis or attempts to repair myelin in AD and dementia[30]. MAZ is involved in regulating genes related to cell proliferation, differentiation, and survival[31]. In dementia, even though MAZ is not oligodendrocyte-specific, the presence of both SOX10 and MAZ may indicate synergism which is a disruption of normal oligodendrocyte functions and myelin loss. MYRF is another key player in myelination that works closely with driving factors of oligodendrocyte differentiation and myelin gene expression[32]. For similar reasons, synergism between MYRF and MAZ may indicate that oligodendrocyte differentiation is promoted in an attempt to repair myelin loss, a process that is common in dementia cases due to demyelination and oligodendrocyte dysfunction (**Fig. 4b**).

Meanwhile, antagonistic interaction was identified between SOX8 and HIF1A. SOX8 is critical in regulating cell fate decisions, development, and differentiation across various tissues[33] while HIF1A, a stress and inflammatory-related factor[34], could suppress differentiation and antagonistically contributing to the pathology of AD and dementia. SOX8 is also antagonistically interacting with PBX1, which is important for the development and differentiation of neural cells, including oligodendrocytes[35, 36]. PBX1 could be involved in AD pathology by disrupting oligodendrocyte differentiation and myelin integrity, while SOX8 might counteract these effects and support oligodendrocyte survival. Their antagonistic interaction

may suggest that an imbalance between these two TFs may contribute to the progression of the dementia (**Fig. 4b**).

Based on the top fifty across-view absolute interaction values (TF expression levels and ATAC peak counts), we identified several TFs that are either synergistically or antagonistically interacting with multiple oligodendrocyte-specific regulatory regions. We found synergistic interactions between MAZ and ATAC peak regions that map to multiple genes, including *PTPRF*[37], *ERBB3*[38], *GRID1*[39], *LIMCH1*[40], *FLNC*[41], and *DUSP7*[42] (based on genomic annotations). Since increased chromatin openness likely promotes the expression of these genes associated with neurodegeneration, myelin repair, and cell stress responses, the identified synergistic interactions further support the potential role of MAZ in dementia progression. Furthermore, ZNF135 is a zinc finger TF known for its involvement in regulating chromatin structure and transcriptional activation, particularly in response to cellular stress[43]. Synergistic interactions were identified between ZNF135 and ATAC peak regions that map to genes involved in neural development, synaptic signaling, axon guidance, and neurotransmitter regulation, such as *NKX2-2*[44], *PTPRF*[37], *CNTN2*[45], *PSD2*[46], and *SLC6A9*[47], which may reflect the activation of disease-related pathways during the progression of neurodegeneration (**Fig. 4c**).

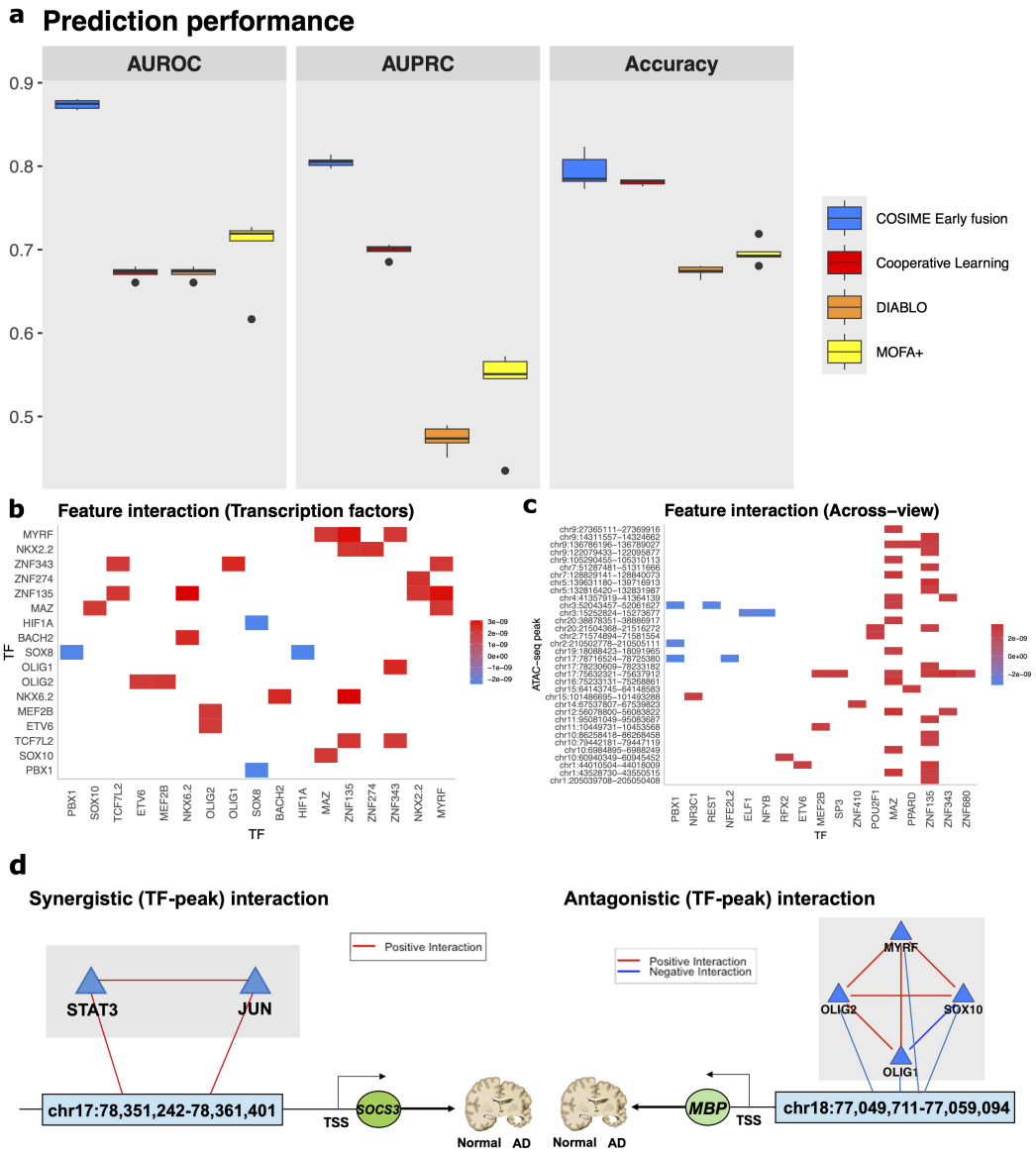
In contrast, PBX1 acts antagonistically with several open chromatin regions, as its upregulation may reduce chromatin accessibility, which is critical for oligodendrocyte function and myelin maintenance. This antagonistic action may prevent the activation of protective genes such as *CPS1*[48], and *CYTH1*[49] in those open chromatic regions, leading to impaired neuroinflammatory responses and exacerbating neurodegenerative processes. The ability of PBX1 to close chromatin at disease-associated regions may contribute to a negative predictive signal for dementia, as it correlates with the loss of protective mechanisms in oligodendrocytes (**Fig. 4c**).

It is crucial to examine the regulatory mechanisms underlying the interactions between TFs, ATAC peak regions, target genes, and dementia. Both STAT3 and JUN are synergistic TFs, and they also exhibit synergistic activity in the genomic region chr17:78,351,242-78,361,401, which contains the *SOCS3* gene. STAT3 is a critical TF involved in the JAK-STAT signaling pathway, regulating immune responses, inflammation, cell survival, and differentiation[50, 51]. JUN, a component of the AP-1 transcription factor complex, regulates genes involved in cell proliferation, apoptosis, and stress responses, particularly in response to inflammatory signals[52, 53]. STAT3 and JUN are known to regulate *SOCS3*, a gene that is upregulated in dementia. *SOCS3* plays a crucial role in regulating the JAK-STAT pathway by inhibiting STAT3 activation[54, 55]. These connections suggest that the coordinated activity of STAT3 and JUN in this region may contribute to the dysregulation of *SOCS3*, linking it to the neuroinflammatory and

neurodegenerative pathways associated with dementia (**Fig. 4d**).

The four key TFs involved in oligodendrocyte function — SOX10[30], MYRF[30], OLIG1[56], and OLIG2[56]— exhibit synergistic relationships, except for SOX10 and OLIG1, which interact antagonistically. SOX10, MYRF, OLIG1, and OLIG2 are essential TFs that regulate oligodendrocyte differentiation, myelination, and myelin maintenance. In the context of dementia, their synergistic interactions suggest that the activities of these TFs contribute to increased myelin damage and oligodendrocyte dysfunction, which accelerates the progression of neurodegeneration. Notably, all four of these TFs antagonistically interact with a ATAC peak in a specific genomic region, chr18:77,049,711-77,059,094, where the gene *MBP* is located. These interactions suggest that these TFs may suppress the expression of *MBP*, which is essential for the formation and maintenance of the myelin sheath in oligodendrocytes[57] (**Fig. 4d**). Taken together, the characterization of these interactions provides valuable insights into the underlying mechanisms that may contribute to the pathogenesis and progression of dementia.

Fig 4: Classifying dementia from transcriptomic (oligodendrocyte) and epigenomic (oligodendrocyte) data



a Prediction performance using COSIME early fusion, CL, DIABLO, and MOFA+ with logistic regression. **b** Heatmap showing the top 50 pairwise TF interactions. **c** Heatmap showing the top 50 pairwise across-view (TF-ATAC peak) interactions. **d** Diagrams illustrating examples of synergistic and antagonistic relationships between TF-ATAC peak-AD phenotype.

Predicting Alzheimer's disease progression scores from transcriptomic (microglia) and spatial transcriptomic MERFISH (astrocyte) data

To assess the predictive capability of COSIME models, multi-view data combining scRNA-seq of microglia and single-cell spatial transcriptomics (MERFISH) of astrocytes were applied. COSIME early fusion performed better than the other benchmarking methods for continuous outcome prediction of Alzheimer's disease progression scores using multi-view datasets (microglia and astrocytes) (MSE:0.033 ± 0.005) (**Fig. 5a**).

Fig. 5b shows the top 20 important genes (ranked by mean absolute feature importance values) in each view. In microglia, there are genes closely related to AD through their roles in neuroinflammation, $A\beta$ clearance, and microglial activation, among those with a positive feature importance. *SPP1* is a well-established microglial marker and plays a critical role in the activation and inflammatory response of microglia. It has been shown to regulate phagocytosis of amyloid plaques and can contribute to both neuroprotection and neurodegeneration[58]. *APOE* genotype is a major genetic risk factor for AD, particularly the E4 allele, where microglial *APOE* is involved in $A\beta$ clearance. It has been shown that *APOE* e4 (the risk allele) impairs this function and enhances neuroinflammation[59, 60]. *C1QB* encodes one of the subunits of C1q, which is part of the complement system and a major player in immune responses[61, 62]. C1q marks apoptotic or damaged neurons for phagocytosis by microglia. These genes are likely upregulated in AD and often reflect microglial dysfunction, contributing to both increased neuroinflammation and impaired $A\beta$ clearance. On the other hand, the genes that have negative feature importance values may be downregulated in microglia and they are crucial to the immune response, microglial surveillance, and synaptic pruning in AD. For example, *BIN1* has been implicated in AD and is involved in regulating synaptic function and endocytosis[63].

In the astrocyte MERFISH data, the genes that have negative feature importance values may suggest dysfunctional astrocytes with impaired synaptic regulation, reduced neuroprotection, and failure to clear neurotoxic debris in AD. For instance, *RYR3* is involved in calcium signaling in astrocytes, which plays a role in synaptic regulation and glutamate uptake[64, 65]. Downregulation of *RYR3* could impair calcium signaling and neurotransmitter homeostasis, contributing to synaptic dysfunction in AD. In contrast, the genes that have positive feature importance values are involved in neuroinflammation, astrocyte differentiation, synaptic function, and neurodegeneration, and their upregulation in AD suggests activated astrocytes contributing to neuroinflammatory responses and synaptic damage. *DGKG* is involved in signal transduction through lipid metabolism and inflammatory response. Its upregulation in astrocytes may promote

increased neuroinflammation in AD. Also, *DGKG* has been linked to glutamate receptor signaling, which could disrupt synaptic communication, a hallmark of AD (**Fig. 5b**)[66]. Enrichment analyses for the top 20 prioritized genes (absolute feature importance values) for both microglia and astrocytes were implemented (**Supplementary Figure 6**). The top 10 categories in each analysis show that most of them are closely related to AD.

We investigated the spatial distribution of cells with the expression levels and feature importance values of *RYR3* (**Fig. 5c**). As a result, cells with the highest 25% of feature importance were more likely to be located in the central regions of the middle temporal gyrus, where the astrocyte MERFISH data were collected, compared to the remaining 75% (T-statistic: 18.783, two-tailed t-test p–value < 0.0001), based on the Euclidean distance from each cell to the midpoint of the bounding box. Similarly, when comparing the lowest 25% of feature importance values to the remaining 75% for *DGKG*, we found that cells with highly negative feature importance values were more likely to be located at the center of the middle temporal gyrus. The two-tailed t-test (T-statistic: -2.861 and two-tailed t-test p–value < 0.01) revealed a significant difference, suggesting that genes with high negative effects are preferentially active in central regions. These findings further support the observation that upregulated genes associated with AD tend to be more active at the edges of the middle temporal gyrus, while downregulated genes are concentrated in the central regions (**Fig. 5c**).

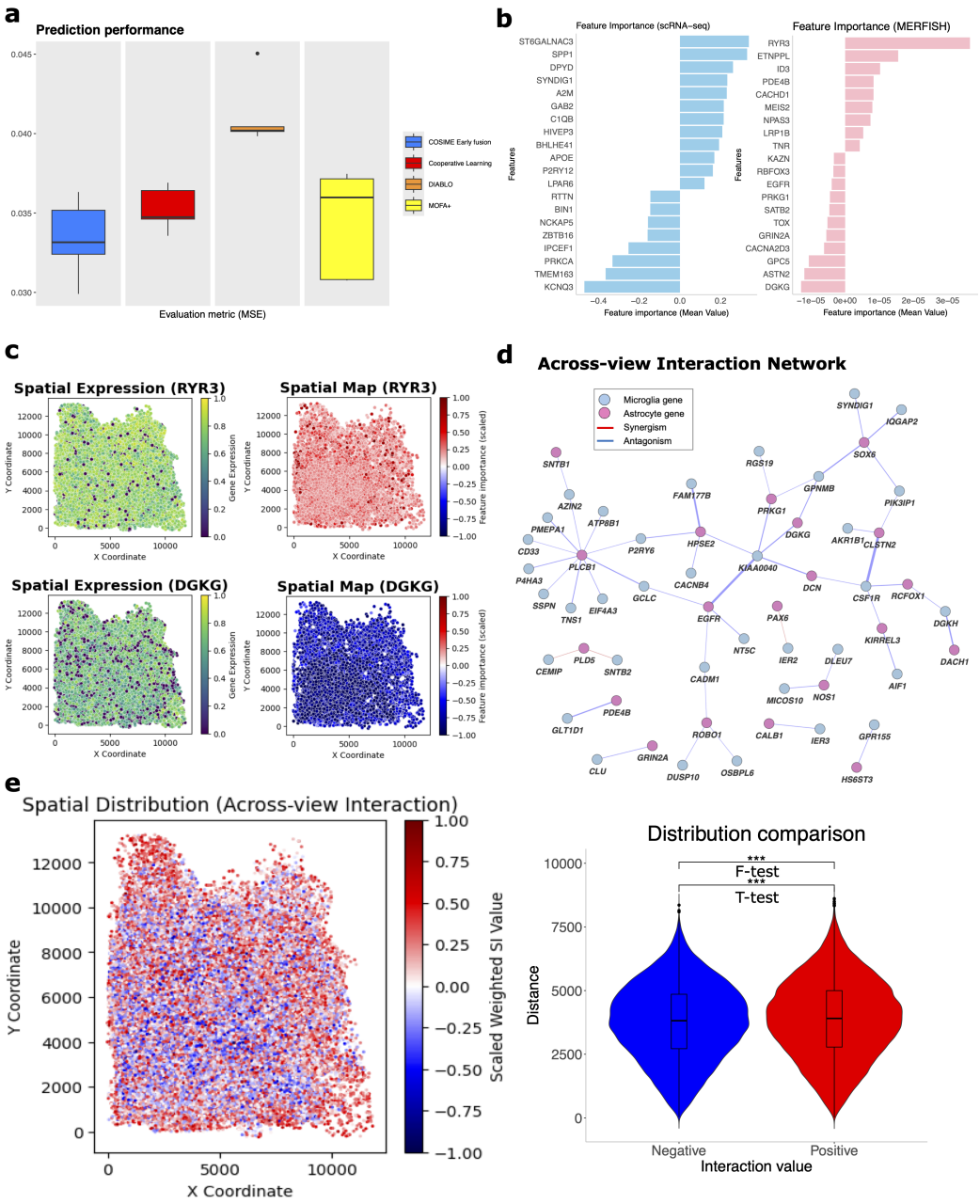
We constructed a gene-gene interaction network based on the top 50 absolute across-view interaction values. Several astrocyte-specific MERFISH genes, such as *PLCB1* and *HPSE2*, were identified as key hub genes, linking multiple microglia genes. *PLCB1*, in particular, is involved in critical signaling pathways that regulate cellular functions such as inflammation, synaptic transmission, and cell differentiation. As an astrocyte-specific gene, *PLCB1* plays a central role in intracellular signaling, including phosphatidylinositol signaling, which could modulate microglial activity and other immune-related processes in the context of AD[67]. Similarly, *HPSE2*, another astrocyte-associated gene, regulates heparan sulfate metabolism, a process involved in cellular signaling, extracellular matrix remodeling, and inflammation[68]. Its interaction with microglial genes suggests that *HPSE2* may influence neuroinflammatory responses and contribute to the pathogenesis of AD (**Fig. 5d**).

A spatial map was generated using all across-view gene-gene interaction values (**Fig. 5e**). Similar to the approach used in the analyses for **Fig. 5c**, we computed the Euclidean distance from each cell to the midpoint of the bounding box and performed a two-tailed t-test to assess the null hypothesis that the distances between cells in positive (synergistic) interaction values and the midpoint of the bounding box are equal to the distances between negative (antagonistic) interaction values) and the midpoint of the bounding

box were the same. We also conducted a two-tailed F-test to assess the null hypothesis that the variances of the two groups were equal. The null hypothesis was rejected in both tests (T-statistic: 18.783 (two-tailed t-test p–value < 0.001) and F-statistic: 1.242 (two-tailed t-test p–value < 0.001)) (**Fig. 5e**). These results suggest that synergistic interactions between microglia and astrocyte genes associated with AD are more likely to be active at the edges of the middle temporal gyrus, while antagonistic interactions tend to be localized more toward the center.

In AD, it has been shown that certain brain regions are more vulnerable to neurodegeneration, including parts of the temporal lobe, which includes the middle temporal gyrus[69]. The middle temporal gyrus is an important region for memory processing, semantic processing, and cognitive functions[70]. It is particularly vulnerable to early pathological changes in AD, such as amyloid plaque deposition and tau tangles[71]. The edges of middle temporal gyrus, like the cortical periphery, can be more exposed to various factors such as inflammation, glial activation, and synaptic dysfunction, which are key drivers of neurodegeneration in AD[72]. As a result, it is plausible that genes at the edges of the middle temporal gyrus are involved in triggering AD progression, especially through their potential role in glial activation (e.g., neuroinflammation), synaptic dysfunction, and neurodegeneration. These regions are often early hotspots for AD pathology, and the genes expressed there may contribute to the progression of AD.

Fig 5: Predicting Alzheimer's disease progression scores from transcriptomic (microglia) and spatial transcriptomic MERFISH (astrocyte) data.



a Prediction performance using COSIME early fusion, CL, DIABLO, and MOFA+ with regression. **b** Bar plots showing the top 20 prioritized genes (absolute feature importance values) in each view. **c** Spatial plots showing expression level and importance values across cells for *RYR3* and *DGKG*. **d** Across-view interaction network showing relationships between the top 50 pairwise across-view (gene-gene) interactions. **e** Spatial distribution of all across-view interaction values colored by the directionality and the distribution of 80,072 positive and 68,619 negative interaction values, with statistical tests (two-tailed t-test and two-tailed F-test) applied to assess differences between two groups. The violin plots including box plots display the distribution of interaction values in the two groups. Asterisk indicate statistical significance: *** for p–value < 0.001.

Discussion

We introduce COSIME (Cooperative Multiview Integration and Scalable Interpretable Model Explainer) and apply it to various multi-view data sets, including simulated data, transcriptomics-metabolomics, transcriptomics-epigenomics, and transcriptomics-spatial transcriptomics. COSIME has two key components: first, it integrates multi-view data for disease phenotype prediction by leveraging deep encoders and LOT techniques. COSIME combines both unsupervised and supervised learning. Second, it computes feature importance values for each view and pairwise feature interaction values for both within-view and across-view.

COSIME addresses several challenges in the existing literature. First, COSIME can handle both linear and nonlinear relationships as well as feature interactions within the data. The deep encoders transform the data into meaningful embeddings without relying on linear assumptions. Once encoded, these embeddings can be used with either linear or nonlinear methods for disease phenotype prediction. Second, COSIME accommodates both matched and unmatched samples across two data views. Each view is processed through a separate deep encoder. For matched samples, COSIME supports either early fusion or late fusion, while for unmatched samples, it uses late fusion. LOT optimally aligns embeddings from the two views into a joint latent space. Third, LOT is a novel technique that aligns the latent space distributions of two distinct data views by treating the transport plan as a learnable parameter, which is dynamically optimized during training. This allows LOT to adapt more flexibly to complex data structures, align latent spaces in a shared representation space, and adjust its alignment strategy to the specific characteristics of the data, making it particularly effective for downstream tasks such as prediction and classification. Fourth, COSIME efficiently computes feature importance scores for individual features and their pairwise interactions through Monte Carlo sampling and batch processing in any complex machine learing models. Lastly, COSIME can identify

the directionality of feature interactions, distinguishing between synergism and antagonism. This enhances model interpretability and can be applied to a wide range of machine learning models. COSIME outputs feature importance and interaction matrices that are both straightforward and intuitive.

To the best of our knowledge, COSIME is the first machine learning method that incorporates the LOT concept to integrate two heterogeneous multi-view datasets for predictive modeling. Given that the relationships in omics data—such as transcriptomics, metabolomics, epigenomics, and spatial transcriptomics—are often nonlinear, its capacity to manage such complexity is crucial. Existing methods such as Cooperative Learning[3], DIABLO[4], and MOFA+[5] are not designed to capture nonlinear relationships. We have demonstrated that COSIME models (both early fusion and late fusion) outperform these methods (Cooperative learning, DAILBO, and MOFA+ with regression models) in predicting disease phenotypes. Moreover, unlike other existing methods, COSIME is able to output both feature importance and interaction values from any complex machine learning models. SHAP[6] has a function to output a feature interaction matrix only available for tree-based models and LIME[7] does not account for feature interactions and is more suitable for interpreting individual predictions in simpler models, such as linear regression. COSIME is user-friendly, allowing users to either train a model using their data, compute feature importance and interaction values with a pre-trained model and their data, or do both. Moreover, users can choose a fusion method (early or late) based on their research objectives and data, and can also choose to compute feature importance, feature interactions, or both.

We demonstrate that COSIME outperforms existing well-known multi-omics models in integrating and predicting outcomes, while also providing valuable feature importance and interaction information. However, there are several areas where COSIME can be further improved and updated. First, while COSIME currently computes pairwise feature interaction values, there may be meaningful interactions involving more than two features. Expanding the model to account for higher-order feature interactions could enhance predictive power but would require more sophisticated algorithms and optimizations. Second, determining the optimal number of hidden layers in the deep encoders, along with the size of the latent space, is a complex task. Although the dimensionality of the joint latent space has been fine-tuned, exploring the number of hidden layers will be addressed in future work to further optimize model performance. To keep pace with rapidly advancing technologies in single-cell sequencing, omics data generation, and the development of state-of-the-art methods, we will continue to update COSIME in future work. Lastly, COSIME can be applied to additional brain-related diseases, such as neuropsychiatric disorders[73, 74], which were not explored in this study, once large-scale multi-omics data become available in the future. Moreover, COSIME can be integrated with other methods[75–77] when multi-view imputation for missing

data is needed, expanding its versatility in handling diverse datasets.

Methods

COSIME overview

Cooperative Multiview Integration and Scalable and Interpretable Model Explainer (COSIME) is a machine learning model that integrates multi-view data for disease phenotype prediction and computes feature importance and interaction scores. By leveraging deep learning-based encoders, COSIME effectively captures the complex, multi-layered interactions between different omic modalities while preserving the unique characteristics of each data type. The integration of LOT techniques aligns and merges heterogeneous datasets, improving the accuracy of modeling across-view relationships in the joint latent space. In addition, COSIME leverages the Shapley-Taylor Interaction Index[9] to compute feature importance and interaction values, allowing for a deeper understanding of how individual features and their interactions contribute to the predictions (**Fig. 1**).

Simulated Data

We generated synthetic data by creating two distinct datasets, $\mathbf{X}_1$ and $\mathbf{X}_2$, each with a specified number of features, to simulate a scenario for multi-view analysis. The function `generate_data` was used to simulate these datasets. Latent factors, which are unobserved variables that affect the data, were introduced with specified standard deviations (std, std_1, and std_2) and were either correlated or uncorrelated depending on the experimental design. The strength of the influence of these latent factors on the multi-view datasets was controlled by the scaling factors scale_1 and scale_2. Additionally, interaction effects between the latent factors of $\mathbf{X}_1$ and $\mathbf{X}_2$ were optionally included to introduce non-linear relationships, with the weight of these interactions controlled by the interaction_weight parameter.

Latent Factors and Dataset Generation

We generate the latent factors for the datasets $\mathbf{X}_1$ and $\mathbf{X}_2$ based on the correlation setting. If the latent factors are correlated, we use a shared latent factor matrix $\mathbf{U}$ while if the latent factors are uncorrelated, we use separate latent factor matrices $\mathbf{U}_1$ and $\mathbf{U}_2$.

Latent Factor Generation: The latent factors for both $\mathbf{X}_1$ and $\mathbf{X}_2$ are drawn from normal distributions:

$$\mathbf{U}_1 \sim \mathcal{N}(0, \sigma_1^2)_{n \times f_{\text{latent}}} \tag{1}$$

$$\mathbf{U}_2 \sim \mathcal{N}(0, \sigma_2^2)_{n \times f_{\text{latent}}} \tag{2}$$

Where:

- n is the number of samples.
- f_{latent} is the number of latent factors.
- σ_1^2, σ_2^2 are the variances of the latent factors affecting $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively.

Correlated Latent Factors: If the latent factors in the two data views are correlated, we use a shared latent factor matrix $\mathbf{U}$, and both datasets $\mathbf{X}_1$ and $\mathbf{X}_2$ are generated as:

$$\mathbf{X}_1 = \mathbf{U} \cdot \beta_1 + \epsilon_1 \tag{3}$$

$$\mathbf{X}_2 = \mathbf{U} \cdot \beta_2 + \epsilon_2 \tag{4}$$

Where:

- $\mathbf{U}$ is the shared latent factor matrix.
- β_1, β_2 are the weights for $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively.
- ϵ_1, ϵ_2 represent the noise for $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively, and follow:

$$\epsilon_1 \sim \mathcal{N}(0, \sigma_{\epsilon_1}^2) \tag{5}$$

$$\epsilon_2 \sim \mathcal{N}(0, \sigma_{\epsilon_2}^2) \tag{6}$$

Uncorrelated Latent Factors: If the latent factors in the two data views are uncorrelated, we generate separate latent factor matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ for $\mathbf{X}_1$ and $\mathbf{X}_2$, and the datasets are generated as:

$$\mathbf{X}_1 = \mathbf{U}_1 \cdot \beta_1 + \epsilon_1 \tag{7}$$

$$\mathbf{X}_2 = \mathbf{U}_2 \cdot \beta_2 + \epsilon_2 \tag{8}$$

Where:

- $\mathbf{U}_1$ and $\mathbf{U}_2$ are the uncorrelated latent factor matrices for $\mathbf{X}_1$ and $\mathbf{X}_2$.
- β_1, β_2 are the weights for $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively.
- ϵ_1, ϵ_2 represent the noise for $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively, and follow:

$$\epsilon_1 \sim \mathcal{N}(0, \sigma_{\epsilon_1}^2) \tag{9}$$

$$\epsilon_2 \sim \mathcal{N}(0, \sigma_{\epsilon_2}^2) \tag{10}$$

Interaction Effects

Correlated Latent Factors: When the latent factors in the two data views are correlated, the interaction term is generated by considering pairwise combinations of the latent factors. Specifically, the interaction term $\mathbf{z}_{\text{interaction}}$ is calculated as:

$$\mathbf{z}_{\text{interaction}} = \sum_{i=f_{\text{latent}}+1}^{f_{\text{latent}}+5} \sum_{j=f_{\text{latent}}+1}^{f_{\text{latent}}+5} w \cdot U_i \cdot U_j \tag{11}$$

Where:

- U_i and U_j are the elements of the latent factor matrix $\mathbf{U}$.
- The interaction term involves the product of latent factors U_i and U_j for each pair.
- The sum is performed over the indices i and j, which range from $f_{\text{latent}} + 1$ to $f_{\text{latent}} + 5$, where the interaction is modeled between latent factors.
- w is the interaction weight, a constant that determines the strength of the interaction between latent factors.

The interaction term $\mathbf{z}_{\text{interaction}}$, which is given by equation 11, is then added to the feature matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ as follows:

$$x_{1,i} = x_{1,i} + \mathbf{z}_{\text{interaction}} \tag{12}$$

$$x_{2,j} = x_{2,j} + \mathbf{z}_{\text{interaction}} \tag{13}$$

Where equations 12 and 13 show how the interaction term $\mathbf{z}_{\text{interaction}}$ is added to the features x_1 and x_2 in the correlated case.

Outcome Variable

The outcome variables $\mathbf{y}_1$ and $\mathbf{y}_2$ are generated based on the feature matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, taking into account whether interaction terms are included or not.

Correlated Latent Factors For the case where the latent factors are correlated (i.e., a shared latent factor matrix $\mathbf{U}$ is used for both $\mathbf{X}_1$ and $\mathbf{X}_2$), the outcome generation is as follows:

- **With Interaction:** When interaction terms are included, the expected outcome μ_{all} is computed as the linear combination of the latent factors $\mathbf{U}$ and the latent factor weights β, along with the pairwise interaction terms between the additional latent factors. The interaction term $\mathbf{z}_{\text{interaction}}$, given by equation 11 is used, which is defined as:

$$\mu_{\text{all}} = \mathbf{U}\beta + \mathbf{z}_{\text{interaction}} \tag{14}$$

- **Without Interaction:** When no interaction terms are included, the expected outcome is simply the linear combination of the latent factors $\mathbf{U}$ weighted by β:

$$\mu_{\text{all}} = \mathbf{U}\beta \tag{15}$$

Uncorrelated Latent Factors For the case where the latent factors are uncorrelated (i.e., separate latent factor matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ are used for $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively), the outcome generation is as follows:

- The expected outcome is the sum of the linear combinations of $\mathbf{U}_1$ and $\mathbf{U}_2$ with their respective weights $\beta_{\mathbf{U}_1}$ and $\beta_{\mathbf{U}_2}$:

$$\mu_{\text{all}} = \mathbf{U}_1\beta_{\mathbf{U}_1} + \mathbf{U}_2\beta_{\mathbf{U}_2} \tag{16}$$

Outcome Variable Generation: The outcome variables $\mathbf{y}_1$ and $\mathbf{y}_2$ are generated using the expected outcomes μ_{all} computed above, along with Gaussian noise for continuous outcomes or using a logistic function for binary outcomes.

- **Continuous Outcomes:** For continuous outcomes, the outcome variables $\mathbf{y}_1$ and $\mathbf{y}_2$ are generated by adding Gaussian noise to the expected values:

$$\mathbf{y}_1 = \mu_{\text{all}} + \epsilon_1, \quad \epsilon_1 \sim \mathcal{N}(0, \sigma^2) \tag{17}$$

$$\mathbf{y}_2 = \mu_{\text{all}} + \epsilon_2, \quad \epsilon_2 \sim \mathcal{N}(0, \sigma^2) \tag{18}$$

Where ϵ_1 and ϵ_2 are Gaussian noise terms with mean 0 and variance σ^2.

- **Binary Outcomes:** For binary outcomes, the predicted probabilities $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ are computed using the logistic function:

$$\hat{\mathbf{y}}_1 = \frac{1}{1 + \exp(-\mu_{\text{all}})} \tag{19}$$

$$\hat{\mathbf{y}}_2 = \frac{1}{1 + \exp(-\mu_{\text{all}})} \tag{20}$$

The outcome variables are then generated based on Bernoulli trials with added Gaussian noise to the predicted probabilities:

$$\mathbf{y}_1 \sim \text{Bernoulli}(p = \hat{\mathbf{y}}_1 + \epsilon_3), \quad \epsilon_3 \sim \mathcal{N}(0, \sigma^2) \tag{21}$$

$$\mathbf{y}_2 \sim \text{Bernoulli}(p = \hat{\mathbf{y}}_2 + \epsilon_4), \quad \epsilon_4 \sim \mathcal{N}(0, \sigma^2) \tag{22}$$

Where ϵ_3 and ϵ_4 are Gaussian noise terms with mean 0 and variance σ^2.

Experimental Conditions

We generated a total of eight experimental conditions, divided between binary and continuous outcome variables. Each condition included two distinct datasets to facilitate multi-view analysis. The conditions were based on two factors: strength of signal in the data (high vs. low) and the fusion method (early vs. late). For each dataset, we used different random seeds to introduce variability while keeping all other parameters consistent, particularly the latent factor structure $\mathbf{U}$. In all datasets, 10 interaction terms were introduced within each view to capture the dependencies between selected features. These interactions model the relationships between certain features within the same view.

The experimental conditions were as follows:

- **Signal Strength and Fusion Type**

549 – **High Signal (Early Fusion)**: Data were generated with a strong latent factor influence, including
550 interaction effects between features. The latent factor structure **U** and interaction terms remained
551 consistent across both datasets.

552 – **High Signal (Late Fusion)**: The same strong latent factor influence was used as in the early
553 fusion condition. However, the datasets were generated with different samples, while maintaining
554 consistency in latent factor structure **U** and interaction effects.

555 – **Low Signal (Early Fusion)**: Data were generated with a weaker latent factor influence. Inter-
556 action effects were included, and the latent factor structure **U** and interaction terms remained
557 consistent across both datasets.

558 – **Low Signal (Late Fusion)**: Data were generated with weak latent factor influence. The two
559 datasets were generated with different samples, while maintaining consistency in latent factor
560 structure **U** and interaction effects.

561 • **Outcome Type**

562 – **Binary Outcomes**: Interaction effects were included, and the latent factor structure **U** and
563 interaction terms remained consistent across datasets, based on the signal strength and fusion
564 type.

565 – **Continuous Outcomes**: Similar to binary outcomes, but the outcome variables were generated
566 using Gaussian noise as detailed in the main text.

567 **Parameters used to generate the datasets:**

568 The following are the parameters used to generate the datasets, based on the notations introduced in the
569 previous sections.

570 • Sample size: $n = 1000$

571 • Number of features: $f_{\mathbf{X}_1} = f_{\mathbf{X}_2} = 100$

572 • Number of latent factors: $f_{\text{latent}} = 25$ (Including five for interacting features)

573 • Scaling factors: $\text{scale}_1 = \text{scale}_2 = 5$ for binary high signal and continuous high signal; $\text{scale}_1 =$
574 $\text{scale}_2 = 10$ for binary low signal and continuous low signal

- Latent factor standard deviations: std = 1 (shared standard deviation for latent factors affecting both views if correlated), $\text{std}_1 = 1$ (standard deviation for latent factors influencing $\mathbf{X}_1$ if uncorrelated), $\text{std}_2 = 1$ (standard deviation for latent factors influencing $\mathbf{X}_2$ if uncorrelated)
- Factor strength: latent_strength = 2 for high signal, latent_strength = 1 for low signal
- Noise standard deviation: $\sigma = 5$ for strong signal conditions, $\sigma = 10$ for weak signal conditions (binary outcomes), $\sigma = 15$ for weak signal conditions (continuous outcomes)
- Correlation between two datasets: TRUE if $\mathbf{X}_1$ and $\mathbf{X}_2$ are correlated, FALSE if they are uncorrelated
- Data type: Binary or continuous outcomes for $\mathbf{y}_1$ and $\mathbf{y}_2$
- Interaction effects: TRUE if interactions between features of $\mathbf{X}_1$ and $\mathbf{X}_2$ are included, with interaction weight $w = 10$
  - Interaction effects between the 21st and 25th features within each view ($\mathbf{X}_1$ and $\mathbf{X}_2$) were artificially introduced.

Where:

– $\mathbf{X}_1$ and $\mathbf{X}_2$ represent the feature matrices for the two datasets.

– $\mathbf{U}$ is the shared latent factor matrix when the latent factors are correlated, and $\mathbf{U}_1$ and $\mathbf{U}_2$ are the separate latent factor matrices for the uncorrelated case.

– β, $\beta_{\mathbf{U}_1}$, and $\beta_{\mathbf{U}_2}$ are the latent factor weights used to generate the expected outcome variables.

– latent_strength indicates the strength of the latent factors' influence on the outcome variables, with higher values corresponding to stronger signal conditions.

– σ denotes the standard deviation of the noise added to the outcome variables.

– w is the interaction weight, a constant that controls the influence of interaction terms between features from $\mathbf{X}_1$ and $\mathbf{X}_2$, particularly for the 21st to 25th features in each view.

Real Data

Transcriptomics-Metabolomics (ROSMAP)

The *Religious Orders Study* and *Rush Memory and Aging Project* (ROSMAP) are ongoing longitudinal clinical-pathologic cohort studies of aging and AD [78]. The ROSMAP data include a wealth of clinical,

cognitive, neuroimaging, genetic, and neuropathological information from older adults, providing invaluable insights into the early stages of AD, cognitive decline, and related neurodegenerative processes.

Transcriptomics Single-cell RNA-seq (scRNA-seq) for astrocytes was downloaded from a published study [79]. From 149,558 cells and 17,817 genes, we projected 4,292 metacells and 16,718 genes for AD pathologic diagnosis using *Metacell-2* [80]. Then, we performed differential expression testing for Alzheimer's disease (AD) pathological diagnosis using *Seurat* and selected the top 280 differentially expressed genes (p-value adjusted < 0.1 and $\log_2$(fold change) > 0.2). Additionally, 25 genes that are either AD-risk or astrocyte-specific (*APOE*[59, 60], *CLU*[81], *SORL1*[82], *TREM2*[82], *ABCA7*[82], *PICALM*[82], *BIN1*[63], *PLD3*[82], *CD33*[83], *CD2AP*[82], *EPHA1*[82], *INPP5D*[82], *FERMT2*[82], *CR1*[82], *PTK2B*[82], *SLC24A4*[82], *CASS4*[82], *ACE*[83], *SORCS1*[82], *GAB2*[84], *CDK5R1*[85], *PRNP*[86], *IL1RAP*[87], *CNP*[68], *TOMM40*[88]) were also selected.

Metabolomics ROSMAP metabolomic data processed by *Metabolon HD4* was downloaded from *AD knowledge Portal* - backend (syn10235592). The dataset comprises a total of 1,055 biochemicals, 971 compounds of known identity (named biochemicals) and 84 compounds of unknown structural identity (unnamed biochemicals) for 514 brain samples. We selected 305 biochemicals that do not have any missingness in their samples.

Multi-view Data For early fusion, we merged transcriptomic and metabolomic data and achieved 2,286 metacells (samples) and 305 genes and 305 metabolites in each view. Metabolites were matched to the donors for metacells in transcriptomic data. For late fusion, we used full samples for each view whose AD pathologic diagnosis is known. 4,292 metacells and 2,286 samples for transcriptomics and metabolomics, respectively were used.

Transcriptomics-Epigenomics (SEA-AD)

The *Seattle Alzheimer's Disease Brain Cell Atlas* (SEA-AD) consortium studies deep molecular and cellular understanding of the early pathogenesis of AD [89]. By integrating neuropathological data, single-cell and spatial genomics, and longitudinal clinical metadata, SEA-AD serves as a unique resource for exploring the mechanisms underlying Alzheimer's disease and related dementias.

Transcriptomics scRNA-seq data for oligodendrocytes from middle temporal gyrus (111,194 cells) was downloaded from the *Registry of Open Data on AWS* as AnnData objects (h5ad format). 3,927 metacells

and 17,368 genes were projected for cognitive status (dementia). We selected expressions for the 206 transcription factors that are known to co-bind or interact each other in oligodendrocytes[29].

Epigenomics Single-cell ATAC-seq (scATAC-seq) data for all nuclei from middle temporal gyrus was downloaded from the *Registry of Open Data on AWS* as AnnData objects (h5ad format). There were 35,925 cells and 218,882 peaks for oligodendrocytes. Among those peaks, we chose 702 peaks that are overlapped with the known oligodendrocyte-specific peaks[29].

Multi-view Data We merged the scRNA-seq and scATAC-seq datasets by cell. The 35,925 cells from the scATAC-seq data were then grouped according to the metacells they belonged to in the scRNA-seq data, and the average counts for each metacell were computed.

scRNA-seq-Single-cell Spatial Transcriptomic (MERFISH) (SEA-AD)

scRNA-seq scRNA-seq data for microglia was downloaded from *CellxGene* [90] for a published study [89]. From 40,000 cells and 18,279 genes, we projected 1,009 metacells and 17,348 genes for AD progression score using *Metacell-2*. Then, we implemented differential expression testing for AD progression score using the Differential Gene Expression analysis based on the negative binomial distribution (*DESeq2*) [91] and selected 400 differentially expressed genes (p-value adjusted < 0.05 and $\log_2$(fold change) > 0.75). Additionally, 28 genes that are either AD-risk or microglia-specific were identified. (*CSF1R*[92], *TREM2*[82], *SORL1*[82], *CD68*[92], *TNF*[93], *CLEC7A*[92], *IL6*[**108**], *APOE*[59, 60], *C1QA*[92], *C1QB*[92], *C1QC*[92], *SPP1*[92], *MAPK8*[94], *PTGS2*[95], *VEGFA*[96], *FOS*[92], *CLU*[81], *CR1*[82], *PICALM*[82], *CD33*[83], *SORCS1*[82], *GAB2*[84], *CASS4*[82], *CDK5R1*[92], *PRNPL*[86], *IL1RAP*[92], *CNP*[68], *TOMM40*[88]) were also selected.

Spatial transcriptomic (MERFISH) Astrocyte single-cell MERFISH data for the whole taxonomy collected from the middle temporal gyrus was obtained through the *Open Data Registry on AWS* as AnnData objects (h5ad format) for a published study [89]. From 1,321,191 cells and 135 genes, 8,693 metacells and 134 genes were projected for AD progression score using *Metacell-2*. We 1,016 metacells were for astrocytes.

Multi-view Data Among the 1,016 metacells in the MERFISH data, we randomly selected 1,009 metacells. The AD progression score was highly correlated between the scRNA-seq and MERFISH datasets (r

= 0.986). For each dataset, we computed the mean AD progression score, which was highly correlated with the corresponding scores in both datasets (r = 0.997 for scRNA-seq and r = 0.996 for MERFISH).

Model Design

Let $\mathbf{X}_1 \in \mathbb{R}^{n \times p_{\mathbf{X}_1}}$ and $\mathbf{X}_2 \in \mathbb{R}^{n \times p_{\mathbf{X}_2}}$ represent two data views, and $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^n$ be real-valued labels that represent the same disease phenotype corresponding to $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. COSIME inputs each data view into a deep neural network encoder (deep encoder), performs dimensionality reduction to learn K-dimensional features, and creates embeddings in an unsupervised fashion. Then, the Learnable Optimal Transport (LOT) aligns the distributions of two latent spaces in a joint latent space. We use the joint latent vector to classify or predict disease phenotypes.

Deep Neural Network Encoders (Deep encoders)

COSIME is designed to process dual-view data using separate probabilistic deep encoders for each view. Each view $x_1 \in X_1$ and $x_2 \in X_2$ is passed through its own deep encoder to learn a latent representation. These encoders approximate the posterior distributions $p(z_1|x_1)$ and $p(z_2|x_2)$, where z_1 and z_2 are the latent variables corresponding to views 1 and 2, respectively. COSIME utilizes backpropagation of Learnable Optimal Transport (LOT) to these deep neural networks, enabling the learning of latent features from multiple views, which is then used for predicting disease phenotypes.

The deep encoders output the mean (μ_i) and log-standard deviation ($\log(\sigma_i)$) for the posterior distribution of each view. The encoder for each view consists of multiple fully connected layers with LeakyReLU activations. Batch Normalization is applied to stabilize learning and accelerate convergence, while Dropout is used as a regularization technique to mitigate overfitting. The final output of each deep encoder is a latent vector z_i that represents the compressed information from the corresponding input data of a view.

To regularize the learned posterior distributions and prevent overfitting, the model incorporates the Kullback-Leibler divergence (KLD) loss during the training of the deep encoders. This loss encourages the posterior distributions of each view to remain close to a standard Gaussian prior distribution $p(z_i) = \mathcal{N}(0, I)$, which helps improve the ability of the model to generalize and prevents overfitting. The KLD loss for each view is computed as:

$$\mathcal{L}_{\text{KLD}_1} = D_{\text{KLD}_1}(\psi_1(z_1|x_1) \parallel p(z_1)) \tag{23}$$

$$\mathcal{L}_{\text{KLD}_2} = D_{\text{KLD}_2}(\psi_2(z_2|x_2) \parallel p(z_2)) \tag{24}$$

where ψ_1 and ψ_2 are the encoder functions for views A and B, respectively, D_{KLD_A} and D_{KLD_B} are the KLD measures for the latent representations of views A and B, respectively, and $p(z_1)$ and $p(z_2)$ are the prior distributions for the latent variables z_1 and z_2 corresponding to views A and B. The KLD loss for each view i is defined as:

$$D_{\mathrm{KLD}_i}(\psi_i(z_i|x_i) \parallel p(z_i)) = \frac{1}{2}\left(\sigma_i^2 + \mu_i^2 - \log\sigma_i^2 - 1\right) \tag{25}$$

where μ_i and σ_i are the mean and standard deviation of the posterior distribution for view i.

Learnable Optimal Transport (LOT)

COSIME utilizes Learnable Optimal Transport (LOT), a novel method designed to align the latent space distributions of two distinct data views. There are several key advantages of LOT.

First, unlike traditional optimal transport methods, Learnable Optimal Transport (LOT) treats the transport plan as a learnable parameter, optimized dynamically during training rather than relying on a fixed or pre-defined plan. This allows LOT to adapt more flexibly to complex data structures and align latent spaces in a joint latent space. By learning the transport plan iteratively, LOT can adjust its alignment strategy to the specific characteristics of the data, making it particularly effective for downstream tasks such as prediction and classification.

Second, LOT can scale effectively with large datasets. By incorporating mini-batch Sinkhorn[97] updates, LOT accelerates the optimization process, reduces memory usage, and ensures efficient computation. This makes LOT particularly well-suited for high-dimensional data from multiple sources, enabling it to handle large-scale tasks without sacrificing performance.

Third, LOT integrates domain-specific regularization through the coupling matrix, which allows the model to incorporate prior knowledge about the relationship between the views. This is particularly beneficial when working with heterogeneous data in the multi-view analysis, where domain-specific relationships can enhance the quality of the learned latent space. We apply entropy regularization to stabilize the optimization process, ensuring smooth convergence and better-converging embeddings, which are crucial for downstream tasks.

Learnable Optimal Transport (LOT) Process LOT treats the transport plan as a learnable parameter, which is iteratively updated through backpropagation. This enables LOT to adaptively align the latent spaces of different views in a shared representation space.

Both the source and target distributions in LOT are modeled as Gaussian distributions, each parameterized by their means (μ) and standard deviations (σ). The transport plan is learned by minimizing the pairwise transport cost between these distributions. The cost reflects the distance between their latent vectors, and the transport plan is updated during training to reduce this cost, aligning the distributions in a shared latent space.

The process involves the following steps:

1. **Step 1: Mini-Batching Processing**

   Before performing the computation, the source and target distributions are split into mini-batches to ensure efficient processing, especially for large datasets. Let the total number of source samples be N_{src} and the total number of target samples be N_{tgt}. In the mini-batching procedure, we split the source distribution $\{\mu_{\text{src}}, \sigma_{\text{src}}\}$ and the target distribution $\{\mu_{\text{tgt}}, \sigma_{\text{tgt}}\}$ into smaller batches of size B_{src} and B_{tgt}, respectively. Denote the i-th mini-batch of the source and target distributions as:

$$\mathcal{B}^i_{\text{src}} = \left\{\mu^i_{\text{src}}, \sigma^i_{\text{src}}\right\}, \quad \mathcal{B}^i_{\text{tgt}} = \left\{\mu^i_{\text{tgt}}, \sigma^i_{\text{tgt}}\right\} \tag{26}$$

   for $i = 1, 2, \ldots, N_{\text{src}}/B_{\text{src}}$ and $j = 1, 2, \ldots, N_{\text{tgt}}/B_{\text{tgt}}$, respectively. This mini-batch processing reduces memory consumption and allows the transport plan to be optimized iteratively.

   Each mini-batch is processed separately, and the transport plan is updated incrementally based on these smaller subsets of the source and target distributions.

2. **Step 2: Pairwise Distance Computation**

   The pairwise distances between the means μ^i_{src} and μ^j_{tgt} and the standard deviations σ^i_{src} and σ^j_{tgt} of the mini-batches are computed using the Euclidean distance:

$$\text{dist}_\mu(i, j) = \|\mu^i_{\text{src}} - \mu^j_{\text{tgt}}\|_2, \quad \text{dist}_\sigma(i, j) = \|\sigma^i_{\text{src}} - \sigma^j_{\text{tgt}}\|_2 \tag{27}$$

   where i and j refer to the samples from the source and target mini-batches, respectively.

   These pairwise distances are used to construct the cost matrix for each mini-batch, which will then be used in the subsequent steps.

3. **Step 3: Cost Matrix Construction**

After computing the pairwise distances, the total cost matrix for each mini-batch is constructed as follows:

$$\text{cost}^i_{total}(i,j) = \text{dist}_\mu(i,j) + \text{dist}_\sigma(i,j) + \epsilon \tag{28}$$

where ϵ is a small constant for numerical stability.

4. **Step 4: Sinkhorn Algorithm (Learnable Transport Plan)**

The transport matrix $\mathbf{P}^i$ for each mini-batch is computed iteratively using the Sinkhorn algorithm. The transport plan is initialized as a learnable parameter, $\mathbf{P}^i$, where each element of $\mathbf{P}^i$ represents the transport probability between a sample from the source and a sample from the target. The transport plan is updated during each iteration to minimize the total transportation cost.

The update rule for the transport plan $\mathbf{P}^i(i,j)$ for each mini-batch is:

$$\mathbf{P}^i(i,j) = \frac{\exp\left(-\frac{\text{cost}^i_{total}(i,j)}{\text{reg}}\right)}{Z^i_{\text{src}} \cdot Z^i_{\text{tgt}}} \tag{29}$$

where Z^i_{src} and Z^i_{tgt} are the normalizing factors for the source and target distributions, respectively:

$$Z^i_{\text{src}} = \sum_j \exp\left(-\frac{\text{cost}^i_{total}(i,j)}{\text{reg}}\right), \quad Z^i_{\text{tgt}} = \sum_i \exp\left(-\frac{\text{cost}^i_{total}(i,j)}{\text{reg}}\right) \tag{30}$$

Here, the learnable transport plan allows gradients to flow through the transport matrix, enabling the optimization process using backpropagation.

Handling Different Sample Sizes The Sinkhorn algorithm can effectively handle source and target distributions with differing sample sizes, making it particularly useful for COSIME late fusion. To ensure that transport plans remain valid, we apply normalization to the weights of the source and target distributions during each mini-batch update as follows:

$$\mathbf{A}(i) = \frac{\mathbf{A}(i)}{\sum_i \mathbf{A}(i)}, \quad \mathbf{B}(j) = \frac{\mathbf{B}(j)}{\sum_j \mathbf{B}(j)} \tag{31}$$

This adjustment ensures that the transport plan remains meaningful even when the source and target distributions contain different numbers of samples.

The Sinkhorn algorithm iteratively refines the transport plan $\mathbf{P}^i$ until convergence or until a predefined number of iterations.

5. **Step 5: Dual Variables and Marginal Constraints**

Dual variables $\mathbf{u}^i$ and $\mathbf{v}^i$ are introduced to enforce the marginal constraints of the transport problem. These variables ensure that the transport plan satisfies the required marginal distributions for the source and target distributions. The dual variables are updated iteratively for each mini-batch:

$$\mathbf{u}^i(i) = \frac{\mathbf{A}(i)}{Z^i_{\text{src}}} \cdot \exp\left(-\frac{\text{cost}^i_{total}(i,j)}{\text{reg}}\right) \tag{32}$$

$$\mathbf{v}^i(j) = \frac{\mathbf{B}(j)}{Z^i_{\text{tgt}}} \cdot \exp\left(-\frac{\text{cost}^i_{total}(i,j)}{\text{reg}}\right) \tag{33}$$

where the dual variables ensure that the transport plan matches the marginal distributions of the source and target distributions.

6. **Step 6: Optimizing the Transport Plan**

During each mini-batch, the transport plan is optimized by performing backpropagation to minimize the total loss. This involves computing the gradients of the transport plan with respect to the loss function and updating the transport plan parameters via gradient descent.

7. **Step 7: Final LOT Loss Calculation**

After optimizing the transport plan, the final LOT loss is computed as the sum of the product of the cost matrix and the transport plan:

$$\mathcal{L}_{\text{LOT}} = \sum_{i,j} \text{cost}_{total}(i,j) \cdot \mathbf{P}(i,j) \tag{34}$$

This loss function quantifies the minimum cost required to transform the source distribution into the target distribution, thus aligning the two views in a joint latent space.

Phenotype Prediction/Classification

COSIME employs the aligned latent representations from each view, obtained via separate probabilistic deep encoders. After the deep encoders generate the latent vectors for each view, LOT aligns these embeddings into a joint latent space. The model then uses these aligned embeddings for phenotype prediction.

Two fusion strategies are employed to combine the embeddings from the two views: early fusion and late fusion. Both fusion strategies are applied for binary outcome classification and continuous outcome prediction tasks.

Early Fusion The fused representation is obtained by taking the mean of the representations $z_1 \in \mathbb{R}^d$ and $z_2 \in \mathbb{R}^d$ from each view:

$$z_{\text{fusion}} = \frac{z_1 + z_2}{2} \tag{35}$$

Late Fusion The latent representations $z_1 \in \mathbb{R}^d$ and $z_2 \in \mathbb{R}^d$ from each view are vertically concatenated to form the fused representation z_{fusion}:

$$z_{\text{fusion}} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \tag{36}$$

where $z_{\text{fusion}} \in \mathbb{R}^{2d}$ is the fused vector obtained by stacking z_1 and z_2 vertically.

Binary Outcome Prediction The fused representation z_{fusion} is then passed through a linear layer followed by a sigmoid activation function to compute the predicted probability $\hat{p}$:

$$\hat{p} = \sigma(W_{\text{binary}} \cdot z_{\text{fusion}} + b) \tag{37}$$

where $\sigma(\cdot)$ is the sigmoid function, and W_{binary} and b are the learned weights and bias parameters of the prediction layer.

Loss function

The **BCEWithLogitsLoss** is computed for the early fusion setup, which combines the sigmoid activation and binary cross-entropy in a single, more numerically stable operation. This loss function is defined as:

$$\mathcal{L}_{\text{BCE}}^{\text{early}} = -\frac{1}{n}\sum_{i=1}^{n} [y_i \log(\sigma(\hat{p}_i)) + (1 - y_i) \log(1 - \sigma(\hat{p}_i))] \tag{38}$$

where $\hat{p}_i$ is the predicted probability of the i-th sample belonging to the positive class (i.e., the probability output by the model), and y_i is the true binary label for the i-th sample, where $y_i \in \{0, 1\}$.

Continuous Phenotype Prediction For continuous phenotype prediction, the fused representation z_{fusion} is passed through a linear transformation:

$$\hat{y} = W_{\text{reg}} \cdot z_{\text{fusion}} + b \tag{39}$$

where W_{reg} and b are the learned weights and bias of the regression layer.

Loss for Regression The Mean Squared Error (MSE) loss for continuous outcome prediction using a linear model is computed as:

$$\mathcal{L}_{\text{MSE}}^{\text{early}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{40}$$

where $\hat{y}_i$ is the predicted continuous value for the i-th sample, representing the model's prediction, and y_i is the true continuous value for the i-th sample, which is the actual target value the model aims to predict. The loss measures the squared difference between the predicted and true values for each sample and averages this over all N samples.

Loss for Neural Network A neural network (NN) model can be used for continuous outcome prediction, the output $\hat{y}_i$ is predicted by the neural network, and the loss function remains MSE:

$$\mathcal{L}_{\text{MSE}}^{\text{NN, early}} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \hat{y}_i^{\text{NN}} \right)^2 \tag{41}$$

where $\hat{y}_i^{\text{NN}}$ is the continuous value predicted by the neural network for the i-th sample, which is the output of the neural network model for that particular input, and y_i is the true continuous value for the i-th sample, which is the actual target value that the model is attempting to predict. The loss measures the squared difference between the predicted value and the true value for each sample, and the average squared error is computed over all N samples.

Prediction Loss: The prediction loss term $\mathcal{L}_{\text{pred}}$ is defined based on the task at hand:

$$\mathcal{L}_{\text{pred}} = \begin{cases} \mathcal{L}_{\text{BCE}}^{\text{early}}, & \text{for binary outcome classification (early fusion)} \\ \mathcal{L}_{\text{MSE}}^{\text{early}}, & \text{for continuous outcome prediction (early fusion, regression)} \\ \mathcal{L}_{\text{MSE}}^{\text{NN, early}}, & \text{for continuous outcome prediction (early fusion, NN)} \\ \mathcal{L}_{\text{BCE}}^{\text{late}}, & \text{for binary outcome classification (late fusion)} \\ \mathcal{L}_{\text{MSE}}^{\text{late}}, & \text{for continuous outcome prediction (late fusion, regression)} \\ \mathcal{L}_{\text{MSE}}^{\text{NN, late}}, & \text{for continuous outcome prediction (late fusion, NN)} \end{cases}$$

Total Weighted Loss

During the training phase, the contributions of each loss term can vary in magnitude, depending on the relative scale of the individual losses. To balance the influence of each term on the total loss, we normalize the loss terms dynamically. This dynamic normalization ensures that no individual loss term has an outsized influence on the total loss, allowing for a more balanced optimization process.

We first compute the inverse of each loss term:

$$\text{Inv}_{\mathcal{L}_{\text{KLD}_{\mathcal{A}}}} = \frac{1}{\mathcal{L}_{\text{KLD}_A}(\theta) + 1e^{-8}} \tag{42}$$

$$\text{Inv}_{\mathcal{L}_{\text{KLD}_{\mathcal{B}}}} = \frac{1}{\mathcal{L}_{\text{KLD}_B}(\theta) + 1e^{-8}} \tag{43}$$

$$\text{Inv}_{\mathcal{L}_{\text{LOT}}} = \frac{1}{\mathcal{L}_{\text{LOT}}(\theta) + 1e^{-8}} \tag{44}$$

$$\text{Inv}_{\mathcal{L}_{\text{pred}}} = \frac{1}{\mathcal{L}_{\text{pred}}(\theta) + 1e^{-8}} \tag{45}$$

Next, we compute the total inverse magnitude:

$$\text{Total\_Inverse} = \text{Inv}_{\mathcal{L}_{\text{KLD}_{\mathcal{A}}}} + \text{Inv}_{\mathcal{L}_{\text{KLD}_{\mathcal{B}}}} + \text{Inv}_{\mathcal{L}_{\text{LOT}}} + \text{Inv}_{\mathcal{L}_{\text{pred}}} \tag{46}$$

Now, we normalize the inverse magnitudes to get their relative adjusted weights:

$$\text{Norm}_{\mathcal{L}_{\text{KLD}_{\mathcal{A}}}} = \frac{\text{Inv}_{\mathcal{L}_{\text{KLD}_{\mathcal{A}}}}}{\text{Total\_Inverse}} \tag{47}$$

$$\text{Norm}_{\mathcal{L}_{\text{KLD}_{\mathcal{B}}}} = \frac{\text{Inv}_{\mathcal{L}_{\text{KLD}_{\mathcal{B}}}}}{\text{Total\_Inverse}} \tag{48}$$

$$\text{Norm}_{\mathcal{L}_{\text{LOT}}} = \frac{\text{Inv}_{\mathcal{L}_{\text{LOT}}}}{\text{Total\_Inverse}} \tag{49}$$

$$\text{Norm}_{\mathcal{L}_{\text{pred}}} = \frac{\text{Inv}_{\mathcal{L}_{\text{pred}}}}{\text{Total\_Inverse}} \tag{50}$$

Finally, the total weighted loss $\mathcal{L}_{\text{Total\_weighted}}(\theta)$ is computed as:

$$\begin{aligned}\mathcal{L}_{\text{Total\_weighted}}(\theta) = {} & KLD_{A_{\text{weight}}} \cdot \text{Norm}_{\mathcal{L}_{\text{KLD}_{\mathcal{A}}}} \cdot \mathcal{L}_{\text{KLD}_A}(\theta) + \\ & KLD_{B_{\text{weight}}} \cdot \text{Norm}_{\mathcal{L}_{\text{KLD}_{\mathcal{B}}}} \cdot \mathcal{L}_{\text{KLD}_B}(\theta) + \\ & OT_{\text{weight}} \cdot \text{Norm}_{\mathcal{L}_{\text{LOT}}} \cdot \mathcal{L}_{\text{LOT}}(\theta) + \\ & CL_{\text{weight}} \cdot \text{Norm}_{\mathcal{L}_{\text{pred}}} \cdot \mathcal{L}_{\text{pred}}(\theta)\end{aligned} \tag{51}$$

where λ_{KLD_A}, λ_{KLD_B}, λ_{LOT}, and λ_{pred} are the weights that control the importance of each normalized loss term, specifically $\text{Norm}_{\mathcal{L}_{\text{KLD}_{\mathcal{A}}}}$, $\text{Norm}_{\mathcal{L}_{\text{KLD}_{\mathcal{B}}}}$, $\text{Norm}_{\mathcal{L}_{\text{LOT}}}$, and $\text{Norm}_{\mathcal{L}_{\text{pred}}}$, respectively. These normalized loss terms are the contributions of each individual loss term $\mathcal{L}_{\text{KLD}_A}(\theta)$, $\mathcal{L}_{\text{KLD}_B}(\theta)$, $\mathcal{L}_{\text{LOT}}(\theta)$, and $\mathcal{L}_{\text{pred}}(\theta)$ to the total loss, which are normalized to ensure that no single loss dominates the optimization process.

To optimize the model parameters, we minimize the total weighted loss:

$$\theta^* = \underset{\theta}{\text{argmin}}\, \mathcal{L}_{\text{Total\_weighted}}(\theta) \tag{52}$$

This method prevents overemphasis on any one loss component, allowing the model to effectively balance view alignments and prediction accuracy. It enables flexible adjustment of task importance, which is crucial when working with multi-view data, where different objectives (e.g., alignment, prediction) may compete for optimization.

Model training process

The input data for each view was split into 75% for training and 25% for testing (holdout). 5-fold cross-validation was applied to the training set. The best-performing model was selected based on the minimum classification (or prediction) loss during the training phase and was then evaluated on the testing set. The **Algorithm 1: COSIME Model Training** is detailed in the **Supplementary Information**.

The prediction performance of the models was evaluated using holdout data with metrics such as Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), and accuracy for binary outcome classification, and Mean Squared Error (MSE) for continuous

outcome prediction. The performance metrics were compared across different methods using the mean and ±1.96 times the standard deviation. For the best-trained models across the different multi-view datasets, we computed feature importance for each data view and interaction values both within and across views.

Hyperparameter tuning

We tuned several hyperparameters such as the learning rate, learning rate decay (gamma), loss weights, dropout rate, joint latent space dimensionality, early stopping patience, minimum change for early stopping, and weight decay in the optimizer. Hyperparameter optimization was performed using *Ray Tune* [98] with Distributed Asynchronous Hyperparameter Optimization (*Hyperopt*) [99]. The Adam optimizer was employed to minimize the total loss, with a maximum of 300 epochs. Early stopping was used to prevent overfitting, based on performance improvements on the validation set in the training phase.

Optimization and Loss Minimization

During the training phase, the model updates the hyper-parameters by minimizing the total loss. The optimization process involves computing gradients with respect to the total loss function and updating the parameters using gradient descent or its variants.

By minimizing the total loss, COSIME simultaneously learns to:

- Regularize the posterior distributions of the latent representations for each view through the KLD losses,
- Align the embeddings from both views into a shared latent space using LOT,
- Predict or classify the disease phenotype by minimizing classification (or prediction) loss.

Evaluating the Performance of the Model Predictions

To evaluate the performance of the best model selected in the training phase for both binary outcome classification and continuous outcome prediction, several standard metrics are employed:

Binary Outcome Classification

1. **Area Under the Receiver Operating Characteristic Curve (AUROC):** AUROC measures the ability of the model to discriminate between the positive and negative classes. It is calculated by assessing the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold

settings. The AUROC value ranges from 0 to 1, where 1 indicates perfect classification, and 0.5 is the expected performance of a random classifier. The formula for AUROC is:

$$\text{AUROC} = \int_0^1 \text{TPR}(f)\, d\text{FPR}(f) \tag{53}$$

where $\text{TPR}(f)$ and $\text{FPR}(f)$ are the True Positive Rate and False Positive Rate at each decision threshold f.

2. **Area Under the Precision-Recall Curve (AUPRC):** AUPRC is another useful metric for evaluating binary classification performance, particularly when dealing with imbalanced datasets. It measures the area under the curve created by plotting precision against recall at various threshold settings. The formula for AUPRC is:

$$\text{AUPRC} = \int_0^1 \text{Precision}(r)\, d\text{Recall}(r) \tag{54}$$

where $\text{Precision}(r)$ and $\text{Recall}(r)$ are the precision and recall at each decision threshold r.

3. **Accuracy:** Accuracy is the proportion of correctly classified instances (both true positives and true negatives) out of the total number of instances. It is calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{55}$$

where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives.

Continuous Outcome Prediction

Mean Squared Error (MSE): MSE measures the average squared difference between the true values and the predicted values. It is a common loss function for regression tasks and is used to quantify the error in continuous prediction tasks. The formula for MSE is:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{56}$$

where y_i is the true continuous label for the i-th sample, $\hat{y}_i$ is the predicted value for the i-th sample, and n is the total number of samples.

These evaluation metrics assess the effectiveness of the COSIME model in both binary and continuous phenotype prediction tasks. Higher AUROC and AUPRC values, along with higher accuracy and lower MSE, indicate better model performance.

Feature Importance and Interaction

COSIME introduces an advanced approach for computing both feature importance and feature interactions, building upon the Shapley-Taylor framework[9]. By using Monte Carlo sampling and a batching process, COSIME efficiently approximates Shapley values and Shapley-Taylor indices, making it highly scalable and interpretable for large, high-dimensional datasets while maintaining memory efficiency. This approach ensures COSIME is well-suited for complex real-world applications.

COSIME employs Monte Carlo sampling to efficiently approximate both Shapley values and feature interactions. This approach significantly enhances its scalability, interpretability, and flexibility, especially when handling large datasets and high-dimensional feature spaces. Instead of relying on exact calculations, which become computationally expensive as the number of features or data points increases, Monte Carlo sampling allows COSIME to estimate Shapley values and Shapley-Taylor indices efficiently. This method also offers precise and interpretable calculations of pairwise feature interactions by directly simulating feature combinations. By estimating these interactions in this way, COSIME avoids complex approximations or indirect assumptions, providing clearer insights into how combinations of features jointly affect predictions. Furthermore, COSIME is better equipped to handle complex feature interactions and nonlinear relationships more effectively.

Additionally, COSIME is memory-efficient due to its batching process, which breaks the data into smaller subsets to minimize memory usage. This enables COSIME to handle large datasets on systems with limited memory, making it suitable for real-world applications with resource constraints. The batching process ensures that COSIME can scale efficiently while maintaining performance across high-dimensional feature spaces.

COSIME also provides the ability to calculate signed feature interactions, which indicate whether pairs of features interact synergistically (positive interaction) or antagonistically (negative interaction). This is achieved by directly simulating feature combinations during the Monte Carlo sampling process, allowing COSIME to estimate both the magnitude and direction of interactions. The sign of an interaction is determined by comparing the joint effect of the features to their individual effects, revealing whether their

combined influence on the prediction. This enhancement adds a significant layer of interpretability, enabling users to understand not only the strength but also the directionality of feature interactions.

Finally, COSIME allows users to customize the Monte Carlo sample size and maximum memory usage, providing flexibility in the computational process. When the memory usage is specified, COSIME automatically determines the optimal batch size, making the method both user-friendly and adaptable to different system configurations. The details of **Algorithm 2: COSIME Feature Importance and Interaction Computation** are provided in the **Supplementary Information**.

Shapley Value (Feature Importance Computation)

The Shapley value ϕ_i for a given feature i is computed by considering the marginal contribution of feature i across all possible combinations of other features. This is approximated using Monte Carlo sampling over multiple iterations. The Shapley value for each feature i and each sample s is computed as:

$$\phi_p(s) = \frac{1}{M} \sum_{m=1}^{M} \left(f(\mathbf{x}_s^{(p-1)}) - f(\mathbf{x}_s^{(p)}) \right) \tag{57}$$

where:

- $f(\mathbf{x}_s)$ represents the prediction using all features,
- $f(\mathbf{x}_s^{(p)})$ represents the prediction when feature p is masked,
- $f(\mathbf{x}_s^{(p-1)})$ represents the prediction when all features except for feature p are masked.

The Shapley values for all features are stored in a matrix $\mathcal{S} \in \mathbb{R}^{N \times F}$, where N is the number of samples and F is the number of features.

$$\mathcal{S}_{ij} = \phi_j(s_i) \tag{58}$$

where i indexes the i-th sample ($i \in \{1, 2, \ldots, N\}$) and j indexes the j-th feature ($j \in \{1, 2, \ldots, F\}$), and $\phi_j(s_i)$ represents the Shapley value for feature j with respect to sample i.

Feature Interaction Effects

The interaction effect between two features p and q quantifies the joint contribution of features p and q compared to when they are individually absent. The interaction effect $\mathcal{I}_{pq}$ is computed as:

$$\mathcal{I}_{pq}(s) = f(\mathbf{x}_s) - f(\mathbf{x}_s^{(p)}) - f(\mathbf{x}_s^{(q)}) + f(\mathbf{x}_s^{(pq)}) \tag{59}$$

where:

- $f(\mathbf{x}_s)$: the prediction of the model using all features,
- $f(\mathbf{x}_s^{(p)})$ and $f(\mathbf{x}_s^{(q)})$: the predictions of the model with features p and q individually masked,
- $f(\mathbf{x}_s^{(pq)})$: the prediction of the model when both features p and q are masked.

For pairwise interactions, the interaction effect is averaged over M Monte Carlo iterations:

$$\mathcal{I}_{ij}(s) = \frac{1}{M} \sum_{m=1}^{M} \left[f(\mathbf{x}_s) - f(\mathbf{x}_s^{(i)}) - f(\mathbf{x}_s^{(j)}) + f(\mathbf{x}_s^{(ij)}) \right] \tag{60}$$

The interaction effects for all pairs of features are stored in the interaction matrix $\mathcal{I} \in \mathbb{R}^{F \times F}$, where each entry $\mathcal{I}_{ij}$ represents the interaction between features i and j.

$$\mathcal{I}_{ij} = \mathcal{I}_{ji} \tag{61}$$

Monte Carlo Approach and Interaction Algorithm

Given the model f, input data $X \in \mathbb{R}^{N \times F}$ (where N is the number of samples and F is the number of features), and target labels $y \in \mathbb{R}^N$, we compute the Shapley values and interaction effects as follows:

- **Shapley Values**: For each feature i, compute the marginal contribution to each sample over M Monte Carlo iterations. Store the results in the matrix $\mathcal{S}$.
- **Interaction Effects**: For each pair of features (i, j), compute the interaction effect using the formula above, averaging over M Monte Carlo iterations. Store the results in the interaction matrix $\mathcal{I}$.

The algorithm processes the data in batches to manage memory efficiently. If the batch size is not provided, it is dynamically computed based on the available memory. The number of batches B is given by:

$$B = \left\lceil \frac{N}{\text{batch size}} \right\rceil \tag{62}$$

The batch processing ensures that we do not exceed the maximum memory usage during computation. Considering the available resources and estimated computing time, we used 100 Monte Carlo iterations for the simulated multi-view data and 10 iterations for the real multi-view data. For each multi-view dataset, the test (holdout) data was used.

Memory Management

To ensure that the memory usage remains within a specified limit, we compute the memory required for each batch as:

$$\text{Batch memory (GB)} = \frac{\text{batch size} \times \text{element size of X} \times F}{1 \times 10^9} \tag{63}$$

If the calculated batch memory exceeds the available memory, the batch size is adjusted accordingly. The dynamic batch size calculation ensures that the computation does not exceed the memory limit.

Output

The Shapley values are stored in a matrix $\mathcal{S}$ with dimensions $N \times F$, and the interaction effects are stored in a matrix $\mathcal{I}$ with dimensions $F \times F$. These matrices represent the individual contributions of each feature and their pairwise interactions in the prediction process.

In summary, our new approach provides a comprehensive methodology for assessing both individual feature importance (Shapley values) and pairwise feature interactions in machine learning models. By leveraging these tools, we enable a deeper understanding of how features individually and collectively influence model predictions. The computational framework we present, along with the accompanying code, facilitates the application of this methodology to various prediction tasks, enhancing model interpretability and providing valuable insights into feature relationships.

Downstream analyses

Gene Enrichment Analysis

The function `enrichGO` from the `clusterProfiler` [100] R package, a universal enrichment tool for interpreting omics data, was used to perform enrichment analyses for the 20 prioritized genes by absolute feature importance (FI) values in each view of real multi-view datasets. The top 10 categories were displayed by $\log_{10}$(FDR) (**Fig. 3b** and **Supplementary Figure 1**).

Metabolite Enrichment Analysis

`MetaboAnalyst 6.0` [101] was used to implement a metabolite set enrichment analysis for 20 prioritized metabolites by absolute feature FI values in the metabolomic data for classifying cognitive diagnosis for Alzheimer's disease (**Fig. 3c**).

Interaction Network

For the top 50 across-view interaction values, the edges and nodes for the interacting genes (and genes and metabolites) were formatted in R, and the network was generated using `Cytoscape 3.10.3` [102] (**Fig. 3e** and **Fig. 5d**).

Spatial mapping and statistical tests

Spatial expression plots for *RYR3* and *DGKG* were generated in Python, employing `matplotlib` and `seaborn` (**Fig. 5c**). To assess whether there are significant spatial differences between cells with high FI values and low feature importance values for each gene, the following steps were taken:

- Compute the midpoint of the bounding box: It is calculated by finding the midpoint of the bounding box defined by the x and y coordinates. Specifically, the center is determined by averaging the maximum and minimum values of both the x and y coordinates:

$$\text{center}_x = \frac{\max(x) + \min(x)}{2}, \quad \text{center}_y = \frac{\max(y) + \min(y)}{2} \tag{64}$$

  This central point serves as the reference from which the distances of each point are measured.

- Compute the Euclidean distance from each point to the midpoint of the bounding box: Once the midpoint of the bounding box is defined, the Euclidean distance from each point (x_i, y_i) to the computed midpoint of the bounding box $(\text{center}_x, \text{center}_y)$ is calculated using the formula:

$$\text{distance}_i = \sqrt{(x_i - \text{center}_x)^2 + (y_i - \text{center}_y)^2} \tag{65}$$

  This step computes a scalar value for each point that quantifies how far each point is from the center.

- Perform a two-sample two-tailed t-test: The dataset is divided into two groups based on the feature importance:
  - Group 1: Points with FI values greater than or equal to the 75th percentile (i.e., the top 25%).
  - Group 2: Points with FI values below the 75th percentile.

  We performed a two-sample two-tailed t-test to evaluate whether the distances from the midpoint of the bounding box differed between the top 25% of cells with the highest feature importance values

and the remaining 75% of cells. The null hypothesis (H_0) for the two-tailed t-test is:

$$H_0 : \mu_{\text{Group1}_{\text{FI}}} = \mu_{\text{Group2}_{\text{FI}}}$$

where $\mu_{\text{Group1}_{\text{FI}}}$ and $\mu_{\text{Group2}_{\text{FI}}}$ represent the means of the distances for the high and low FI groups, respectively. The alternative hypothesis is:

$$H_1 : \mu_{\text{Group1}_{\text{FI}}} \neq \mu_{\text{Group2}_{\text{FI}}}$$

The test statistic t and the associated p–value are calculated to determine whether the difference in means is statistically significant.

Similarly, the spatial distribution of across-view interaction values was visualized using the same tools (**Fig. 5e**). The following steps were taken to generate the spatial distribution of across-view interaction values and assess whether there are significant spatial differences between cells with positive feature interaction values and negative feature interaction values, the subsequent steps were carried out:

- Distance calculation: The distance of each cell from the center of the tissue sample was calculated using its spatial coordinates (X, Y). This distance quantifies how far each cell is from the central point of the tissue section. The Euclidean distance from the center is computed as:

$$\text{Distance to center} = \sqrt{(X - X_{\text{center}})^2 + (Y - Y_{\text{center}})^2} \tag{66}$$

  where X_{center} and Y_{center} are the coordinates of the center of the tissue.

- Cell-level interaction score: For each cell, the interaction score was computed using the MERFISH gene expression levels and pairwise across-view interaction values. Let $\mathbf{E}^{(i)}_{\text{MERFISH}}$ denote the gene expression vector for cell i, which is of length G_{MERFISH}. The interaction matrix $\mathbf{M}$ has dimensions $G_{\text{MERFISH}} \times G_{\text{scRNA-seq}}$, where each element M_{g_1,g_2} represents the pairwise interaction between MERFISH gene g_1 and scRNA-seq gene g_2. The cell-level interaction score $S^{(i)}$ for each cell i was calculated as the dot product of the gene expression vector for that cell and the interaction matrix:

$$S^{(i)} = \mathbf{E}^{(i)}_{\text{MERFISH}} \cdot \mathbf{M} \tag{67}$$

In expanded form, the score for cell i with respect to scRNA-seq gene j is:

$$S_j^{(i)} = \sum_{g_1=1}^{G_{\text{MERFISH}}} \mathbf{E}_{\text{MERFISH}}^{(i)}[g_1] \cdot M_{g_1,j} \tag{68}$$

This score quantifies the interaction between the MERFISH gene expression profile of cell i and the scRNA-seq gene expression profile, enabling grouping of cells based on their interaction scores. Cells were then grouped based on the sign of their interaction scores: those with positive scores were assigned to the Positive group, while those with negative scores were assigned to the Negative group. This grouping allowed for further comparison of spatial distributions and other characteristics between the two groups.

- Perform a two-sample two-tailed t-test: To determine whether there were significant differences in the distances of cells from the center between the two groups, an independent two-sample t-test was performed. The hypotheses tested were:
  - Null hypothesis (H_0): The mean distance from the center of the Positive and Negative groups are equal, i.e.,
  $$H_0 : \mu_{\text{Positive}} = \mu_{\text{Negative}}$$
  - Alternative hypothesis (H_1): The mean distance from the center of the Positive and Negative groups are not equal, i.e.,
  $$H_1 : \mu_{\text{Positive}} \neq \mu_{\text{Negative}}$$
- Perform a two-tailed F-test: In addition to comparing the means, an F-test was performed to compare the variances of the two groups. The hypotheses tested were:
  - Null hypothesis (H_0): The variances of the Positive and Negative groups are equal, i.e.,
  $$H_0 : \sigma^2_{\text{Positive}} = \sigma^2_{\text{Negative}}$$
  - Alternative hypothesis (H_1): The variances of the Positive and Negative groups are not equal, i.e.,
  $$H_1 : \sigma^2_{\text{Positive}} \neq \sigma^2_{\text{Negative}}$$

Benchmarking methods

Cooperative Learning

Cooperative learning is a framework for fitting machine learning models to multi-view data[3]. It presents a loss function that minimizes the prediction error across all views, while encouraging the predictions from the views to agree. This framework with linear regression, which is what we use for benchmarking.

We implemented the Cooperative Learning framework using the `multiview` R package. Following the examples provided on the Cooperative Learning GitHub repository, we preprocessed the input data using the *preProcess* function to center and scale it. For model evaluation, we conducted 5-fold cross-validation using the `cv.multiview` function, with a maximum of 300 iterations. For binary outcome classification, we set `family = binomial()` and `type.measure = "class"`. For continuous outcome prediction, we used the default `family = gaussian()` and `type.measure = "mse"`. The hyperparameter values for the regularization parameter were chosen from the set $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$, while the lambda values were left at their default settings.

Data Integration Analysis for Biomarker discovery using Latent components (DIABLO)

Data Integration Analysis for Biomarker discovery using Latent components (DIABLO) is a supervised extension of canonical correlation analysis[4]. It aims to identify linear combinations of variables from each dataset that maximize the correlation between datasets.

We implemented DIABLO using the `mixOmics` R package. Following the tutorial on the mixOmics website, we created a data block using the matrix function. For model training, we performed 5-fold cross-validation using the `block.splsda` function. The number of variables was tested within the set $\{10, 20, 30, 40, 50\}$.

Multi-Omics Factor Analysis v2 (MOFA+)

Multi-Omics Factor Analysis v2 (MOFA+) is an unsupervised matrix decomposition method that can be viewed as a generalization of sparse principal components analysis6.

We used the `MOFA2` R package to obtain latent representations. Following the tutorial on the MOFA+ downstream analysis in R, we trained a MOFA model using the `run_mofa` function. To project unseen data from a new view into the latent space Z_i, we computed the pseudo-inverse of the weight matrix W_i. The latent representations Z_i for all views were then concatenated. For prediction, we applied logistic regression for binary outcomes and linear regression for continuous outcomes. To classify and predict new, unseen

samples, we used the concatenated latent representation matrix Z in PyTorch (Python 3.10). We evaluated different latent factor dimensions from the set $\{1, 5, 10, 15\}$.

Data Availability

Simulated data and all data supporting the results are included in Supplementary Data 1–9 and are publicly available on GitHub (https://github.com/daifengwanglab/COSIME). ROSMAP scRNA-seq was sourced from a website (https://compbio.mit.edu/ad$_{a}ging_{b}rain/\#loading-the-raw-data$) and metabolomic data can be downloaded from the AD Knowledge Portal - backend (syn10235592). SEA-AD scATAC-seq, scRNA-seq, and MERFISH data were obtained from SEA-AD: Seattle Alzheimer's Disease Brain Cell Atlas (https://cellxgene.cziscience.com/collections/1ca90a2d-2943-483d-b678-b809bf464c30).

Code Availability

The entire COSIME code is implemented in Python. The COSIME Python package and examples can be accessed publicly at https://github.com/daifengwanglab/COSIME.

References

1. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. en. *Genome Biology* **18,** 83. ISSN: 1474-760X. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1215-1 (2024) (Dec. 2017).

2. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. eng. *Bioinformatics and Biology Insights* **14,** 1177932219899051. ISSN: 1177-9322 (2020).

3. Ding, D. Y., Li, S., Narasimhan, B. & Tibshirani, R. Cooperative learning for multiview analysis. en. *Proceedings of the National Academy of Sciences* **119,** e2202113119. ISSN: 0027-8424, 1091-6490. https://pnas.org/doi/full/10.1073/pnas.2202113119 (2024) (Sept. 2022).

4. Singh, A. *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. en. *Bioinformatics* **35** (ed Birol, I.) 3055–3062. ISSN: 1367-4803, 1367-4811. https://academic.oup.com/bioinformatics/article/35/17/3055/5292387 (2024) (Sept. 2019).

5. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. en. *Genome Biology* **21,** 111. ISSN: 1474-760X. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02015-1 (2024) (Dec. 2020).

6. Lundberg, S. & Lee, S.-I. *A Unified Approach to Interpreting Model Predictions* Version Number: 2. 2017. https://arxiv.org/abs/1705.07874 (2024).

7. Ribeiro, M. T., Singh, S. & Guestrin, C. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier* Version Number: 3. 2016. https://arxiv.org/abs/1602.04938 (2024).

8. Gadzicki, K., Khamsehashari, R. & Zetzsche, C. *Early vs Late Fusion in Multimodal Convolutional Neural Networks* in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)* (IEEE, Rustenburg, South Africa, July 2020), 1–6. ISBN: 978-0-578-64709-8. https://ieeexplore.ieee.org/document/9190246/ (2024).

9. Dhamdhere, K., Agarwal, A. & Sundararajan, M. *The Shapley Taylor Interaction Index* Version Number: 2. 2019. https://arxiv.org/abs/1902.05622 (2024).

10. O'Brien, R. J. & Wong, P. C. Amyloid precursor protein processing and Alzheimer's disease. eng. *Annual Review of Neuroscience* **34,** 185–204. ISSN: 1545-4126 (2011).

11. Hampel, H. *et al.* The Amyloid- Pathway in Alzheimer's Disease. en. *Molecular Psychiatry* **26,** 5481–5503. ISSN: 1359-4184, 1476-5578. https://www.nature.com/articles/s41380-021-01249-0 (2024) (Oct. 2021).

12. Kinney, J. W. *et al.* Inflammation as a central mechanism in Alzheimer's disease. en. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **4,** 575–590. ISSN: 2352-8737, 2352-8737. https://alz-journals.onlinelibrary.wiley.com/doi/10.1016/j.trci.2018.06.014 (2024) (Jan. 2018).

13. Mallela, J., Yang, J. & Shariat-Madar, Z. Prolylcarboxypeptidase: A cardioprotective enzyme. en. *The International Journal of Biochemistry & Cell Biology* **41,** 477–481. ISSN: 13572725. https://linkinghub.elsevier.com/retrieve/pii/S1357272508001003 (2024) (Mar. 2009).

14. Ngo, M.-L., Mahdi, F., Kolte, D. & Shariat-Madar, Z. Upregulation of prolylcarboxypeptidase (PRCP) in lipopolysaccharide (LPS) treated endothelium promotes inflammation. eng. *Journal of Inflammation (London, England)* **6,** 3. ISSN: 1476-9255 (Jan. 2009).

15. Weiner, I. D., Mitch, W. E. & Sands, J. M. Urea and Ammonia Metabolism and the Control of Renal Nitrogen Excretion. eng. *Clinical journal of the American Society of Nephrology: CJASN* **10,** 1444–1458. ISSN: 1555-905X (Aug. 2015).

16. Gropman, A. L., Summar, M. & Leonard, J. V. Neurological implications of urea cycle disorders. eng. *Journal of Inherited Metabolic Disease* **30,** 865–879. ISSN: 1573-2665 (Nov. 2007).

17. Moreno, B. *et al.* Methylthioadenosine reverses brain autoimmune disease. eng. *Annals of Neurology* **60,** 323–334. ISSN: 0364-5134 (Sept. 2006).

18. Ye, B. *et al.* Klf4 glutamylation is required for cell reprogramming and early embryonic development in mice. en. *Nature Communications* **9,** 1261. ISSN: 2041-1723. https://www.nature.com/articles/s41467-018-03008-2 (2024) (Mar. 2018).

19. Wang, B. *et al.* Metabolism pathways of arachidonic acids: mechanisms and potential therapeutic targets. en. *Signal Transduction and Targeted Therapy* **6,** 94. ISSN: 2059-3635. https://www.nature.com/articles/s41392-020-00443-w (2024) (Feb. 2021).

20. Gegotek, A. & Skrzydlewska, E. en. in *Vitamins and Hormones* 247–270 (Elsevier, 2023). ISBN: 978-0-443-15768-4. https://linkinghub.elsevier.com/retrieve/pii/S0083672922000838 (2024).

21. Moffett, J. R., Ross, B., Arun, P., Madhavarao, C. N. & Namboodiri, A. M. A. N-Acetylaspartate in the CNS: from neurodiagnostics to neurobiology. eng. *Progress in Neurobiology* **81,** 89–131. ISSN: 0301-0082 (Feb. 2007).

22. Rebelos, E. *et al.* Circulating N-Acetylaspartate does not track brain NAA concentrations, cognitive function or features of small vessel disease in humans. en. *Scientific Reports* **12,** 11530. ISSN: 2045-2322. https://www.nature.com/articles/s41598-022-15670-0 (2024) (July 2022).

23. Lima, F. R. S. *et al.* Cellular prion protein expression in astrocytes modulates neuronal survival and differentiation. en. *Journal of Neurochemistry* **103,** 2164–2176. ISSN: 0022-3042, 1471-4159. https://onlinelibrary.wiley.com/doi/10.1111/j.1471-4159.2007.04904.x (2024) (Dec. 2007).

24. Salter, M. *et al.* Initial Identification of a Blood-Based Chromosome Conformation Signature for Aiding in the Diagnosis of Amyotrophic Lateral Sclerosis. en. *EBioMedicine* **33,** 169–184. ISSN: 23523964. https://linkinghub.elsevier.com/retrieve/pii/S2352396418302226 (2024) (July 2018).

25. Price, B. R., Johnson, L. A. & Norris, C. M. Reactive astrocytes: The nexus of pathological and clinical hallmarks of Alzheimer's disease. eng. *Ageing Research Reviews* **68,** 101335. ISSN: 1872-9649 (July 2021).

26. González-Reyes, R. E., Nava-Mesa, M. O., Vargas-Sánchez, K., Ariza-Salamanca, D. & Mora-Muñoz, L. Involvement of Astrocytes in Alzheimer's Disease from a Neuroinflammatory and Oxidative Stress Perspective. eng. *Frontiers in Molecular Neuroscience* **10,** 427. ISSN: 1662-5099 (2017).

27. Lawrence, J. M., Schardien, K., Wigdahl, B. & Nonnemacher, M. R. Roles of neuropathology-associated reactive astrocytes: a systematic review. eng. *Acta Neuropathologica Communications* **11,** 42. ISSN: 2051-5960 (Mar. 2023).

28. Zhang, W., Xiao, D., Mao, Q. & Xia, H. Role of neuroinflammation in neurodegeneration development. en. *Signal Transduction and Targeted Therapy* **8,** 267. ISSN: 2059-3635. https://www.nature.com/articles/s41392-023-01486-5 (2024) (July 2023).

29. Choi, J. J., Svaren, J. & Wang, D. *Single-cell multi-omics analysis reveals cooperative transcription factors for gene regulation in oligodendrocytes* en. June 2024. http://biorxiv.org/lookup/doi/10.1101/2024.06.19.599799 (2024).

30. Hornig, J. *et al.* The Transcription Factors Sox10 and Myrf Define an Essential Regulatory Network Module in Differentiating Oligodendrocytes. en. *PLoS Genetics* **9** (ed Barres, B. A.) e1003907. ISSN: 1553-7404. https://dx.plos.org/10.1371/journal.pgen.1003907 (2024) (Oct. 2013).

31. Triner, D. *et al.* Myc-Associated Zinc Finger Protein Regulates the Proinflammatory Response in Colitis and Colon Cancer via STAT3 Signaling. eng. *Molecular and Cellular Biology* **38,** e00386–18. ISSN: 1098-5549 (Nov. 2018).

32. Bujalka, H. *et al.* MYRF Is a Membrane-Associated Transcription Factor That Autoproteolytically Cleaves to Directly Activate Myelin Genes. en. *PLoS Biology* **11** (ed ffrench-Constant, C.) e1001625. ISSN: 1545-7885. https://dx.plos.org/10.1371/journal.pbio.1001625 (2024) (Aug. 2013).

33. Schmidt, K., Glaser, G., Wernig, A., Wegner, M. & Rosorius, O. Sox8 Is a Specific Marker for Muscle Satellite Cells and Inhibits Myogenesis. en. *Journal of Biological Chemistry* **278,** 29769–29775. ISSN: 00219258. https://linkinghub.elsevier.com/retrieve/pii/S0021925820842434 (2024) (Aug. 2003).

34. Sarkar, A. & Hochedlinger, K. The sox family of transcription factors: versatile regulators of stem and progenitor cell fate. eng. *Cell Stem Cell* **12,** 15–30. ISSN: 1875-9777 (Jan. 2013).

35. McGettrick, A. F. & O'Neill, L. A. The Role of HIF in Immunity and Inflammation. en. *Cell Metabolism* **32,** 524–536. ISSN: 15504131. https://linkinghub.elsevier.com/retrieve/pii/S1550413120304150 (2024) (Oct. 2020).

36. Hau, A.-C. *et al.* Transcriptional cooperation of PBX1 and PAX6 in adult neural progenitor cells. eng. *Scientific Reports* **11,** 21013. ISSN: 2045-2322 (Oct. 2021).

37. Lee, J.-R. Protein tyrosine phosphatase PTPRT as a regulator of synaptic formation and neuronal development. eng. *BMB reports* **48,** 249–255. ISSN: 1976-670X (May 2015).

38. Sithanandam, G. & Anderson, L. M. The ERBB3 receptor in cancer and cancer gene therapy. eng. *Cancer Gene Therapy* **15,** 413–448. ISSN: 1476-5500 (July 2008).

39. Fossati, M. *et al.* Trans-Synaptic Signaling through the Glutamate Receptor Delta-1 Mediates Inhibitory Synapse Formation in Cortical Pyramidal Neurons. eng. *Neuron* **104,** 1081–1094.e7. ISSN: 1097-4199 (Dec. 2019).

40. Lin, Y.-H. *et al.* LIMCH1 regulates nonmuscle myosin-II activity and suppresses cell migration. eng. *Molecular Biology of the Cell* **28,** 1054–1065. ISSN: 1939-4586 (Apr. 2017).

41. Zhou, A.-X., Hartwig, J. H. & Akyürek, L. M. Filamins in cell signaling, transcription and organ development. en. *Trends in Cell Biology* **20,** 113–123. ISSN: 09628924. https://linkinghub.elsevier.com/retrieve/pii/S0962892409002876 (2024) (Feb. 2010).

42. Kondoh, K. & Nishida, E. Regulation of MAP kinases by MAP kinase phosphatases. en. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1773,** 1227–1237. ISSN: 01674889. https://linkinghub.elsevier.com/retrieve/pii/S0167488906004538 (2024) (Aug. 2007).

43. Lomniczi, A. *et al.* Epigenetic regulation of puberty via Zinc finger protein-mediated transcriptional repression. en. *Nature Communications* **6,** 10195. ISSN: 2041-1723. https://www.nature.com/articles/ncomms10195 (2024) (Dec. 2015).

44. Briscoe, J. *et al.* Homeobox gene Nkx2.2 and specification of neuronal identity by graded Sonic hedgehog signalling. en. *Nature* **398,** 622–627. ISSN: 0028-0836, 1476-4687. https://www.nature.com/articles/19315 (2024) (Apr. 1999).

45. Chatterjee, M. *et al.* Contactin-2, a synaptic and axonal protein, is reduced in cerebrospinal fluid and brain tissue in Alzheimer's disease. eng. *Alzheimer's Research & Therapy* **10,** 52. ISSN: 1758-9193 (June 2018).

46. Citri, A. & Malenka, R. C. Synaptic Plasticity: Multiple Forms, Functions, and Mechanisms. en. *Neuropsychopharmacology* **33,** 18–41. ISSN: 0893-133X, 1740-634X. https://www.nature.com/articles/1301559 (2024) (Jan. 2008).

47. Marques, B. L. *et al.* Neurobiology of glycine transporters: From molecules to behavior. en. *Neuroscience & Biobehavioral Reviews* **118,** 97–110. ISSN: 01497634. https://linkinghub.elsevier.com/retrieve/pii/S0149763420304826 (2024) (Nov. 2020).

48. McGowan, M. *et al.* The Application of Neurodiagnostic Studies to Inform the Acute Management of a Newborn Presenting With Sarbamoyl Shosphate Synthetase 1 Deficiency. eng. *Child Neurology Open* **8,** 2329048X20985179. ISSN: 2329-048X (2021).

49. Rak, J. *et al.* Cytohesin 1 regulates homing and engraftment of human hematopoietic stem and progenitor cells. eng. *Blood* **129,** 950–958. ISSN: 1528-0020 (Feb. 2017).

50. Steelman, A. J. *et al.* Activation of oligodendroglial Stat3 is required for efficient remyelination. eng. *Neurobiology of Disease* **91,** 336–346. ISSN: 1095-953X (July 2016).

51. Jain, M. *et al.* Role of JAK/STAT in the Neuroinflammation and its Association with Neurological Disorders. eng. *Annals of Neurosciences* **28,** 191–200. ISSN: 0972-7531 (July 2021).

52. Zhou, C. *et al.* JUN is a key transcriptional regulator of the unfolded protein response in acute myeloid leukemia. eng. *Leukemia* **31,** 1196–1205. ISSN: 1476-5551 (May 2017).

53. Caprariello, A. V., Mangla, S., Miller, R. H. & Selkirk, S. M. Apoptosis of oligodendrocytes in the central nervous system results in rapid focal demyelination. eng. *Annals of Neurology* **72,** 395–405. ISSN: 1531-8249 (Sept. 2012).

54. Yan, M. *et al.* SOCS modulates JAK-STAT pathway as a novel target to mediate the occurrence of neuroinflammation: Molecular details and treatment options. en. *Brain Research Bulletin* **213,** 110988. ISSN: 03619230. https://linkinghub.elsevier.com/retrieve/pii/S0361923024001217 (2024) (July 2024).

55. Dominguez, E., Mauborgne, A., Mallet, J., Desclaux, M. & Pohl, M. SOCS3-mediated blockade of JAK/STAT3 signaling pathway reveals its major contribution to spinal cord neuroinflammation and mechanical allodynia after peripheral nerve injury. eng. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* **30,** 5754–5766. ISSN: 1529-2401 (Apr. 2010).

56. Zhou, Q. & Anderson, D. J. The bHLH Transcription Factors OLIG2 and OLIG1 Couple Neuronal and Glial Subtype Specification. en. *Cell* **109,** 61–73. ISSN: 00928674. https://linkinghub.elsevier.com/retrieve/pii/S0092867402006773 (2024) (Apr. 2002).

57. Galiano, M. *et al.* Myelin basic protein functions as a microtubule stabilizing protein in differentiated oligodendrocytes. en. *Journal of Neuroscience Research* **84,** 534–541. ISSN: 0360-4012, 1097-4547. https://onlinelibrary.wiley.com/doi/10.1002/jnr.20960 (2024) (Aug. 2006).

58. Mendiola, A. S. *et al.* Defining blood-induced microglia functions in neurodegeneration through multiomic profiling. en. *Nature Immunology* **24,** 1173–1187. ISSN: 1529-2908, 1529-2916. https://www.nature.com/articles/s41590-023-01522-0 (2024) (July 2023).

59. Ferrari-Souza, J. P. *et al. APOE* 4 associates with microglial activation independently of A plaques and tau tangles. en. *Science Advances* **9,** eade1474. ISSN: 2375-2548. https://www.science.org/doi/10.1126/sciadv.ade1474 (2024) (Apr. 2023).

60. Yin, Z. *et al.* APOE4 impairs the microglial response in Alzheimer's disease by inducing TGF- mediated checkpoints. en. *Nature Immunology* **24,** 1839–1853. ISSN: 1529-2908, 1529-2916. https://www.nature.com/articles/s41590-023-01627-6 (2024) (Nov. 2023).

61. Schäfer, M. K.-H. *et al.* Complement C1q Is Dramatically Up-Regulated in Brain Microglia in Response to Transient Global Cerebral Ischemia. en. *The Journal of Immunology* **164,** 5446–5452. ISSN: 0022-1767, 1550-6606. https://journals.aai.org/jimmunol/article/164/10/5446/32714/Complement-C1q-Is-Dramatically-Up-Regulated-in (2024) (May 2000).

62. Fraser, D. A., Pisalyaput, K. & Tenner, A. J. C1q enhances microglial clearance of apoptotic neurons and neuronal blebs, and modulates subsequent inflammatory cytokine production. eng. *Journal of Neurochemistry* **112,** 733–743. ISSN: 1471-4159 (Feb. 2010).

63. Voskobiynyk, Y. *et al.* Alzheimer's disease risk gene BIN1 induces Tau-dependent network hyperexcitability. eng. *eLife* **9,** e57354. ISSN: 2050-084X (July 2020).

64. Matyash, M., Matyash, V., Nolte, C., Sorrentino, V. & Kettenmann, H. Requirement of functional ryanodine receptor type 3 for astrocyte migration. en. *The FASEB Journal* **16,** 1–25. ISSN: 0892-6638, 1530-6860. https://onlinelibrary.wiley.com/doi/10.1096/fj.01-0380fje (2024) (Jan. 2002).

65. Zhang, Y. *et al.* Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. eng. *Neuron* **89,** 37–53. ISSN: 1097-4199 (Jan. 2016).

66. Massart, J. & Zierath, J. R. Role of Diacylglycerol Kinases in Glucose and Energy Homeostasis. en. *Trends in Endocrinology & Metabolism* **30,** 603–617. ISSN: 10432760. https://linkinghub.elsevier.com/retrieve/pii/S1043276019301201 (2024) (Sept. 2019).

67. Lorente-Gea, L., García, B., Martín, C., Quirós, L. M. & Fernández-Vega, I. Heparan sulfate proteoglycans and heparanases in Alzheimer's disease: current outlook and potential therapeutic targets. eng. *Neural Regeneration Research* **12,** 914–915. ISSN: 1673-5374 (June 2017).

68. Sadick, J. S. *et al.* Astrocytes and oligodendrocytes undergo subtype-specific transcriptional changes in Alzheimer's disease. en. *Neuron* **110,** 1788–1805.e10. ISSN: 08966273. https://linkinghub.elsevier.com/retrieve/pii/S0896627322002446 (2024) (June 2022).

69. Pandya, V. A. & Patani, R. Region-specific vulnerability in neurodegeneration: lessons from normal ageing. eng. *Ageing Research Reviews* **67,** 101311. ISSN: 1872-9649 (May 2021).

70. Davey, J. *et al.* Exploring the role of the posterior middle temporal gyrus in semantic cognition: Integration of anterior temporal lobe with executive processes. eng. *NeuroImage* **137,** 165–177. ISSN: 1095-9572 (Aug. 2016).

71. Chen, S. *et al.* Spatially resolved transcriptomics reveals genes associated with the vulnerability of middle temporal gyrus in Alzheimer's disease. en. *Acta Neuropathologica Communications* **10,** 188. ISSN: 2051-5960. https://actaneurocomms.biomedcentral.com/articles/10.1186/s40478-022-01494-6 (2024) (Dec. 2022).

72. Harrison, N. A., Doeller, C. F., Voon, V., Burgess, N. & Critchley, H. D. Peripheral inflammation acutely impairs human spatial memory via actions on medial temporal lobe glucose metabolism. eng. *Biological Psychiatry* **76,** 585–593. ISSN: 1873-2402 (Oct. 2014).

73. Emani, P. S. *et al.* Single-cell genomics and regulatory networks for 388 human brains. en. *Science* **384,** eadi5199. ISSN: 0036-8075, 1095-9203. https://www.science.org/doi/10.1126/science.adi5199 (2025) (May 2024).

74. Gupta, C. *et al. Network-based drug repurposing for psychiatric disorders using single-cell genomics* en. Dec. 2024. http://medrxiv.org/lookup/doi/10.1101/2024.12.01.24318008 (2025).

75. Chandrashekar, P. B. *et al.* DeepGAMI: deep biologically guided auxiliary learning for multimodal integration and imputation to improve genotype–phenotype prediction. en. *Genome Medicine* **15,** 88. ISSN: 1756-994X. https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-023-01248-6 (2025) (Oct. 2023).

76. Cohen Kalafut, N., Huang, X. & Wang, D. Joint variational autoencoders for multimodal imputation and embedding. en. *Nature Machine Intelligence* **5,** 631–642. ISSN: 2522-5839. https://www.nature.com/articles/s42256-023-00663-z (2025) (May 2023).

77. Alatkar, S. A. & Wang, D. CMOT: Cross-Modality Optimal Transport for multimodal inference. en. *Genome Biology* **24,** 163. ISSN: 1474-760X. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-02989-8 (2025) (July 2023).

78. Bennett, D. A. *et al.* Religious Orders Study and Rush Memory and Aging Project. eng. *Journal of Alzheimer's disease: JAD* **64,** S161–S189. ISSN: 1875-8908 (2018).

79. Mathys, H. *et al.* Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. en. *Cell* **186,** 4365–4385.e27. ISSN: 00928674. https://linkinghub.elsevier.com/retrieve/pii/S009286742300973X (2024) (Sept. 2023).

80. Ben-Kiki, O., Bercovich, A., Lifshitz, A. & Tanay, A. Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis. en. *Genome Biology* **23,** 100. ISSN: 1474-760X. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02667-1 (2024) (Dec. 2022).

81. Endo, F. *et al.* Molecular basis of astrocyte diversity and morphology across the CNS in health and disease. en. *Science* **378,** eadc9020. ISSN: 0036-8075, 1095-9203. https://www.science.org/doi/10.1126/science.adc9020 (2024) (Nov. 2022).

82. Karch, C. M., Cruchaga, C. & Goate, A. M. Alzheimer's Disease Genetics: From the Bench to the Clinic. en. *Neuron* **83,** 11–26. ISSN: 08966273. https://linkinghub.elsevier.com/retrieve/pii/S0896627314004851 (2024) (July 2014).

83. Malik, M. *et al.* CD33 Alzheimer's Risk-Altering Polymorphism, CD33 Expression, and Exon 2 Splicing. en. *The Journal of Neuroscience* **33,** 13320–13325. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.1224-13.2013 (2024) (Aug. 2013).

84. Reiman, E. M. *et al.* GAB2 Alleles Modify Alzheimer's Risk in APOE 4 Carriers. en. *Neuron* **54,** 713–720. ISSN: 08966273. https://linkinghub.elsevier.com/retrieve/pii/S0896627307003790 (2024) (June 2007).

85. Wilkaniec, A., Gassowska-Dobrowolska, M., Strawski, M., Adamczyk, A. & Czapski, G. A. Inhibition of cyclin-dependent kinase 5 affects early neuroinflammatory signalling in murine model of amyloid beta toxicity. en. *Journal of Neuroinflammation* **15,** 1. ISSN: 1742-2094. https://jneuroinflammation.biomedcentral.com/articles/10.1186/s12974-017-1027-y (2024) (Dec. 2018).

86. Victoria, G. S., Arkhipenko, A., Zhu, S., Syan, S. & Zurzolo, C. Astrocyte-to-neuron intercellular prion transfer is mediated by cell-cell contact. en. *Scientific Reports* **6,** 20762. ISSN: 2045-2322. https://www.nature.com/articles/srep20762 (2024) (Feb. 2016).

87. Lau, S.-F., Cao, H., Fu, A. K. Y. & Ip, N. Y. Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in Alzheimer's disease. en. *Proceedings of the National Academy of Sciences* **117,** 25800–25809. ISSN: 0027-8424, 1091-6490. https://pnas.org/doi/full/10.1073/pnas.2008762117 (2024) (Oct. 2020).

88. Roses, A. *et al.* Understanding the genetics of *APOE* and *TOMM40* and role of mitochondrial structure and function in clinical pharmacology of Alzheimer's disease. en. *Alzheimer's & Dementia* **12,** 687–694. ISSN: 1552-5260, 1552-5279. https://alz-journals.onlinelibrary.wiley.com/doi/10.1016/j.jalz.2016.03.015 (2024) (June 2016).

89. Gabitto, M. I. *et al.* Integrated multimodal cell atlas of Alzheimer's disease. en. *Nature Neuroscience.* ISSN: 1097-6256, 1546-1726. https://www.nature.com/articles/s41593-024-01774-5 (2024) (Oct. 2024).

90. Megill, C. *et al. cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices* en. Apr. 2021. http://biorxiv.org/lookup/doi/10.1101/2021.04.05.438318 (2024).

91. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. en. *Genome Biology* **15,** 550. ISSN: 1474-760X. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8 (2024) (Dec. 2014).

92. Chen, Y. & Colonna, M. Microglia in Alzheimer's disease at single-cell level. Are there common patterns in humans and mice? eng. *The Journal of Experimental Medicine* **218,** e20202717. ISSN: 1540-9538 (Sept. 2021).

93. Bhaskar, K. *et al.* Microglial derived tumor necrosis factor- drives Alzheimer's disease-related neuronal cell cycle events. en. *Neurobiology of Disease* **62,** 273–285. ISSN: 09699961. https://linkinghub.elsevier.com/retrieve/pii/S0969996113002829 (2024) (Feb. 2014).

94. Ajoolabady, A., Lindholm, D., Ren, J. & Pratico, D. ER stress and UPR in Alzheimer's disease: mechanisms, pathogenesis, treatments. en. *Cell Death & Disease* **13,** 706. ISSN: 2041-4889. https://www.nature.com/articles/s41419-022-05153-5 (2024) (Aug. 2022).

95. Rangaraju, S. *et al.* Identification and therapeutic modulation of a pro-inflammatory subset of disease-associated-microglia in Alzheimer's disease. en. *Molecular Neurodegeneration* **13,** 24. ISSN: 1750-1326. https://molecularneurodegeneration.biomedcentral.com/articles/10.1186/s13024-018-0254-8 (2024) (Dec. 2018).

96. Sun, N. *et al.* Human microglial state dynamics in Alzheimer's disease progression. en. *Cell* **186,** 4386–4403.e29. ISSN: 00928674. https://linkinghub.elsevier.com/retrieve/pii/S0092867423009716 (2024) (Sept. 2023).

97. Damodaran, B. B., Flamary, R., Seguy, V. & Courty, N. An Entropic Optimal Transport loss for learning deep neural networks under label noise in remote sensing images. en. *Computer Vision and Image Understanding* **191,** 102863. ISSN: 10773142. https://linkinghub.elsevier.com/retrieve/pii/S1077314219301559 (2024) (Feb. 2020).

98. Liaw, R. *et al. Tune: A Research Platform for Distributed Model Selection and Training* Version Number: 1. 2018. https://arxiv.org/abs/1807.05118 (2025).

99. Bergstra, J., Yamins, D. & Cox, D. D. Hyperopt: Distributed asynchronous hyper-parameter optimization. *Astrophysics Source Code Library.* ADS Bibcode: 2022ascl.soft05008B, ascl:2205.008. https://ui.adsabs.harvard.edu/abs/2022ascl.soft05008B (2025) (May 2022).

100. Yu, G. Thirteen years of clusterProfiler. en. *The Innovation* **5,** 100722. ISSN: 26666758. https://linkinghub.elsevier.com/retrieve/pii/S2666675824001607 (2025) (Nov. 2024).

101. Pang, Z. *et al.* MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. en. *Nucleic Acids Research* **52,** W398–W406. ISSN: 0305-1048, 1362-4962. https://academic.oup.com/nar/article/52/W1/W398/7642060 (2025) (July 2024).

102. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. eng. *Genome Research* **13,** 2498–2504. ISSN: 1088-9051 (Nov. 2003).

Acknowledgements

This work was supported by National Institutes of Health grants, R21 NS128761, RF1MH128695, R01AG067025, R21NS128761, RF1AG054047, and National Science Foundation Career Award 2144475, and a core grant to the Waisman Center from NICHD (P50 HD105353).

Author Information

Authors and Affiliations

Department of Population Health Sciences, University of Wisconsin-Madison, Madison, WI 53726, USA

Jerome J. Choi, Corinne D. Engelman, and Tianyuan Lu

Waisman Center, University of Wisconsin-Madison, Madison, WI 53705, USA

Jerome J. Choi, Noah Cohen Kalafut, and Daifeng Wang

Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 5706, USA

Noah Cohen Kalafut and Daifeng Wang

Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53076, USA

Tim Gruenloh and Daifeng Wang

Contributions

Study conception, D.W.; Conceptualization, J.J.C. and D.W.; Methodology, J.J.C. and N.C.K.; Formal Analysis, J.J.C. and T.G.; Investigation, J.J.C., T.L., and D.W.; Writing–Original Draft, J.J.C.; Writing–Review and Edit, J.J.C., N.C.K., T.G., C.D.E., T.L., and D.W.; Supervision, T.L. and D.W.; Funding Acquisition, C.D.E. and D.W..

Corresponding Authors

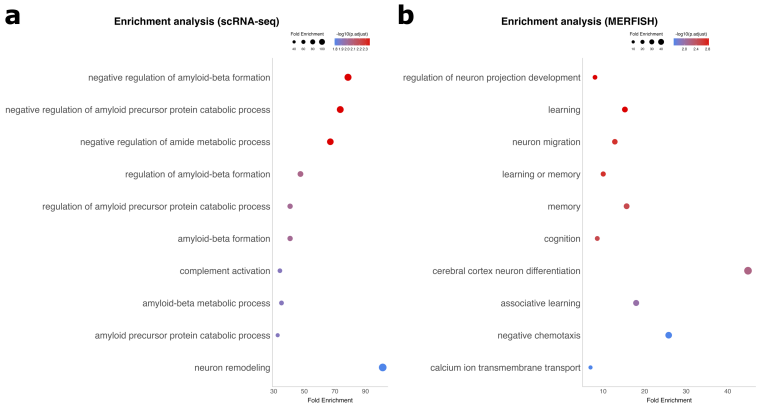
Correspondence to Daifeng Wang.

Ethics Declarations

Competing interests

The authors declare no competing interests.

Supplementary information



Supplementary Figure 1: Enrichment analyses prioritized for each view.
a The top 20 prioritized genes (absolute feature importance values) in microglia **b** The top 20 prioritized genes (absolute feature importance values)in astrocytes.

Supplementary Data 1

Table showing the prediction performance of COSIME models and benchmark models for binary outcome classification.

Supplementary Data 2

Table showing the prediction performance of COSIME models and benchmark models for continuous outcome classification.

Supplementary Data 3

Feature importance and interactions for the simulated binary high signal early fusion.

Supplementary Data 4

Feature importance and interactions for simulated binary low signal early fusion.

Supplementary Data 5

Feature importance and interactions for the simulated continuous high signal early fusion.

Supplementary Data 6

Feature importance and interactions for the simulated continuous low signal early fusion.

Supplementary Data 7

Feature importance and interactions for the ROSMAP binary classification early fusion.

Supplementary Data 8

Feature importance and interactions for the SEA-AD scRNA-seq and scATAC-seq (oligodendrocytes) early fusion.

Supplementary Data 9

Feature importance and interactions for the SEA-AD scRNA-seq (microglia) and MERFISH (astrocytes) early fusion.

Algorithm 1: COSIME Model Training

Training Loop
for $m = 1$ **to** M **do**

Input:

- Training: $X_1 \in \mathbb{R}^{n \times p_{X_1}}$, $X_2 \in \mathbb{R}^{n \times p_{X_2}}$
- Validation: $(X_1^{\text{val}}, X_2^{\text{val}}, y_1^{\text{val}}, y_2^{\text{val}})$
- Test: $(X_1^{\text{test}}, X_2^{\text{test}}, y_1^{\text{test}}, y_2^{\text{test}})$

Initialization:

- Set training iterations M, optimizer, and early stopping criteria
- Set tuning hyperparameters:
  $\text{batch\_size}, \text{learning\_rate}, \text{learning\_gamma}, \lambda_{\text{KLD\_A\_weight}}, \lambda_{\text{KLD\_B\_weight}}, \lambda_{\text{OT\_weight}}, \lambda_{\text{CL\_weight}}, \text{dropout}, \text{dim}, p_{\text{earlystop\_patience}}, \delta, \text{decay}$

Forward Pass:

- Compute KLD Losses, LOT Loss, and Prediction Loss: $\mathcal{L}_{\text{KLD}_A}(\theta)$, $\mathcal{L}_{\text{KLD}_B}(\theta)$, $\mathcal{L}_{\text{LOT}}(\theta)$, $\mathcal{L}_{\text{pred}}(\theta)$
- Compute Total Weighted Loss: $\mathcal{L}_{\text{Total\_weighted}}(\theta)$

Backpropagation:

- Backpropagate gradients for all parameters, including the transport plan for $\mathcal{L}_{\text{LOT}}$
- Perform optimization step: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{Total\_weighted}}(\theta)$

Early Stopping:
Compute validation prediction loss: $\mathcal{L}_{\text{pred}}(\theta)$ for validation
If $\mathcal{L}_{\text{pred}}(\theta) < \mathcal{L}_{\text{pred}}(\theta_{\text{best}})$ **(improvement in validation loss)**:
`Update` $\theta_{\text{best}} = \theta$
`Reset patience counter:` $p_{\text{counter}} = 0$
Else If $p_{\text{counter}} > p_{\text{earlystop\_patience}}$ **(early stopping)**:
`Break training loop`
Else:
`Increment patience counter:` $p_{\text{counter}} \leftarrow p_{\text{counter}} + 1$
End If
end for

Holdout Evaluation
Evaluate best model on the test set: compute losses and metrics;

Output
Trained best COSIME model and the evaluation metrics;

Algorithm 2: COSIME Feature Importance and Interaction Computation

Input

$X \in \mathbb{R}^{N \times F}$: Input feature matrix,
$y \in \mathbb{R}^{N}$: Target labels

Initialization

- Define the model f and input matrix X with N samples and F features.
- Set number of Monte Carlo iterations M.
- Define the batch size, or compute it based on available memory.
- Compute the number of batches: $B = \lceil \frac{N}{\text{batch size}} \rceil$.
- Compute the memory required for each batch:

$$\text{Batch memory (GB)} = \frac{\text{batch size} \times \text{element size of } X \times F}{1 \times 10^9}.$$

- If batch memory exceeds available memory, adjust the batch size dynamically.

Feature Importance Computation

1. For each feature i in X, compute the marginal contribution over M Monte Carlo iterations.
2. Store feature importance values in matrix $\mathcal{S} \in \mathbb{R}^{N \times F}$:
   $\mathcal{S}_{ij} = \phi_j(s_i)$, where $\phi_j(s_i)$ is the feature importance value for feature j for sample s_i.

Feature Interaction Computation

1. For each pair of features (i, j), compute the interaction effect over M Monte Carlo iterations.
2. Store interaction effects in matrix $\mathcal{I} \in \mathbb{R}^{F \times F}$:
   $\mathcal{I}_{ij} = \mathcal{I}_{ji}$, where $\mathcal{I}_{ij}$ is the interaction effect between features i and j.

Output

$\mathcal{S} \in \mathbb{R}^{N \times F}$: Feature importance matrix,
$\mathcal{I} \in \mathbb{R}^{F \times F}$: Feature interaction matrix.

CHAPTER 3: Multi-Omics Integration of Transcriptomics and Metabolomics with Machine Learning Uncovers Novel Risk Factors for Alzheimer's Disease

Multi-Omics Integration of Transcriptomics and Metabolomics with Machine Learning Uncovers Novel Risk Factors for Alzheimer's Disease

Jerome J. Choi[1], Corinne D. Engelman[1], Tianyuan Lu[1,2]

[1]Department of Population Health Sciences, University of Wisconsin-Madison
[2]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison

Abstract

INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disorder marked by cognitive decline, memory impairment, and functional deterioration. Its complex pathogenesis involves factors such as amyloid plaques, tau tangles, neuroinflammation, and synaptic dysfunction, but the precise mechanisms remain unclear, hindering effective treatment. Genetic, environmental, and lifestyle factors contribute to AD risk, yet their interactions are poorly understood. Recent advances in transcriptomics and metabolomics have shed light on the molecular underpinnings of AD, with gene expression alterations and metabolic disruptions implicated in disease progression. However, the interplay between these omics layers suggests a need for integrative approaches.

METHODS

This study utilizes the Cooperative multi-view integration and Scalable and Interpretable Model Explainer (COSIME) machine learning method to combine imputed whole blood transcriptomics

and measured plasma metabolomics data. A predictive model was developed through 5-fold cross-validation to identify genes and metabolites associated with AD-related phenotypes, including Preclinical Alzheimer's Cognitive Composite 3 (PACC3) and plasma phosphorylated tau 217 (ptau217), from the Wisconsin Registry for Alzheimer's Prevention (WRAP) cohort (N = 1,046 for both PACC3 and ptau217). The model performance was validated using an independent dataset from the Wisconsin Alzheimer's Disease Research Center (ADRC) cohort (N = 85 for PACC3 and N = 89 for ptau217). Feature importance and interactions were assessed to identify potential risk factors.

RESULTS

The normalized root mean square error (NRMSE) was 0.654 ± 0.004 for predicting PACC3 and 1.361 ± 0.046 for predicting ptau217 in the WRAP cohort. The NRMSE was 1.095 for predicting PACC3 and 1.795 for predicting ptau217 in the Wisconsin ADRC cohort. Biomarker importance and pairwise interaction values were computed. Several genes known to down-regulate cognitive function (e.g., *HOXA4* and *LRFN4*) and others associated with the up-regulation of ptau217 (e.g., *ACAD9* and *GPAT3*) were highly ranked based on our biomarker importance scores. Similarly, metabolites that down-regulate cognitive function (e.g., sphingomyelin (d17:2/16:0, d18:2/15:0) and N-carbamoylvaline) and those that up-regulate ptau217 (e.g., S-adenosylhomocysteine and propionylcarnitine (C3)) were also identified as highly ranked. Some metabolites, such as sphingomyelin (d17:2/16:0, d18:2/15:0) and tetradecadienoate (14:2), may interact synergistically, while sphingomyelin (d18:2/16:0, d18:1/16:1) and choline may interact antagonistically.

DISCUSSION

There is significant potential for using integrated transcriptomics and metabolomics data to identify novel biomarkers and predict AD phenotypes. Genes and metabolites associated with cognitive decline and tau pathology were identified, offering insights into additional molecular mechanisms underlying AD. The identified interactions between certain metabolites provide further clues to the complex metabolic dysregulation that may contribute to disease pathogenesis, underscoring the importance of integrative approaches for understanding AD.

1 BACKGROUND

Alzheimer's disease (AD) is a neurodegenerative disease characterized by progressive cognitive decline, memory impairment, and functional deterioration, representing a major public health challenge[1]. The pathogenesis of AD is multifactorial, involving the accumulation of amyloid plaques, neurofibrillary tangles, neuroinflammation, and synaptic dysfunction.[2–5] Despite significant advances in understanding the molecular underpinnings of the disease, the precise mechanisms remain unclear, and the complexity of AD presents a challenge for early diagnosis and effective therapeutic development. Genetic, environmental, and lifestyle factors all contribute to AD risk,[6,7] but the interactions between these factors are poorly understood. As a result, there is a pressing need for more comprehensive approaches to elucidate the molecular landscape of AD and identify novel biomarkers that could aid in early detection and intervention.

In recent years, transcriptomics and metabolomics have emerged as key omics approaches for investigating the molecular basis of AD.[8] Transcriptomics, which analyzes gene expression patterns, provides insights into the dysregulated cellular processes in AD, such as neuronal loss, synaptic dysfunction, and inflammatory responses.[9] Alterations in the expression of genes related

to amyloid processing, tau metabolism, and neuroinflammation have been linked to disease progression, highlighting the role of transcriptional dysregulation in AD pathophysiology.[10] On the other hand, metabolomics focuses on the small molecules involved in cellular metabolism, offering a unique window into the biochemical changes that occur in AD.[11] Studies have shown that metabolites associated with energy metabolism, oxidative stress, and neurotransmitter pathways are altered in AD, suggesting that disrupted metabolic networks contribute to disease onset and progression.[12] While both transcriptomics and metabolomics provide valuable insights into AD, the interplay between genes and metabolites suggests that a more integrative approach is required.[13] The interactions between genetic and metabolic alterations play a critical role in AD pathophysiology,[14] and understanding how these molecular networks influence each other is essential for unraveling the complexity of AD mechanisms.

Multi-view integration of omics data provides a comprehensive understanding of systems biology and the complex interactions between biomarkers, particularly in complex diseases such as AD.[15,16] By combining data from multiple omics layers, it is possible to capture a more holistic view of the molecular mechanisms underlying AD. Moreover, this integrative approach may allow for the identification of key biomarkers and pathways that may be missed when examining a single omics layer in isolation.[17] The vast and complex nature of multi-omics data requires advanced computational tools to analyze and interpret the intricate relationships between genes and metabolites. Machine learning techniques offer powerful methods for uncovering patterns within large, high-dimensional datasets.[18] By leveraging machine learning algorithms, we can integrate diverse omics data to predict disease phenotypes, identify novel biomarkers, and better understand the underlying pathophysiology of AD.[19] This study applies a state-of-art machine learning method, Cooperative multi-view integration and Scalable and Interpretable Model Explainer (COSIME)[20]

to integrate imputed transcriptomics and metabolomics, predict AD-related phenotypes, and identify potential AD-risk factors from the Wisconsin Registry for Alzheimer's Prevention (WRAP)[21] cohort. The cohort from the Wisconsin Alzheimer's Disease Research Center (ADRC) is used for independent validation.

2 METHODS

2.1 Study design

An overview of the study design is presented in **FIGURE 1**. Gene expression data were imputed using the genotypes of WRAP and Wisconsin ADRC participants via PrediXcan[22]. To reduce the dimensionality, untargeted transcriptomic and metabolomic analyses were performed, including age, sex, *APOE* score, and years of education (for Preclinical Alzheimer's Cognitive Composite 3 (PACC3) only) as covariates, to predict AD-related phenotypes in WRAP using linear regression. Subsequently, for each AD-related phenotype, nominally significant genes and metabolites (p-value < 0.05) were input into a COSIME model, which was trained to predict the outcome. After training the COSIME model, biomarker importance for individual genes and metabolites, as well as their interaction values, were computed. Lastly, the Wisconsin ADRC cohort was used to independently validate the prediction performance of the COSIME[20] models.

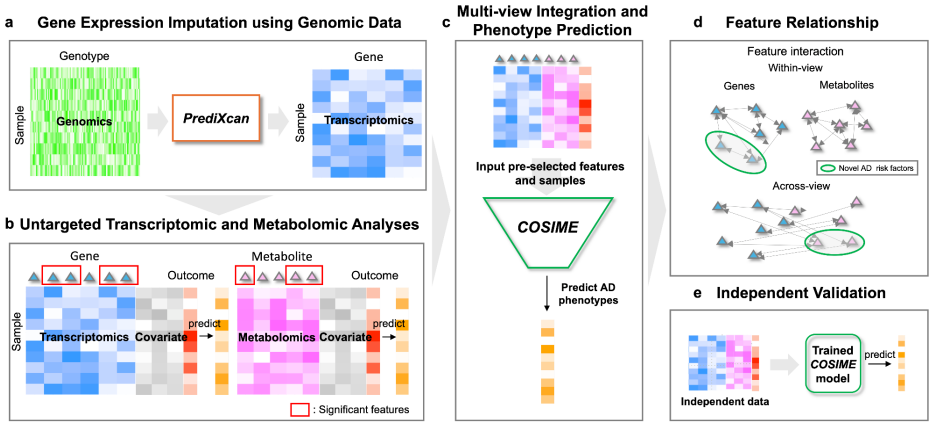


FIGURE 1

Workflow diagram. **a** Gene expression data for the WRAP participants were imputed using PrediXcan. **b** Untargeted transcriptomics and metabolomics adjusting for covariates were implemented to pre-select significant biomarkers. **c** Multi-omics data were integrated and AD phenotypes were predicted using COSIME. **d** Biomarker importance and interaction values for both within-view and across-view were computed by COSIME. **e** Independent validation for predicting AD phenotypes using the Wisconsin ADRC cohort as independent multi-view data was performed.

2.2 Study participants and cohort description

The cohort for the main analysis, WRAP, is an ongoing longitudinal study designed to identify midlife factors associated with the development of AD. Participant enrollment began in 2001, with the first follow-up visit occurring 2 to 4 years after the baseline visit, and all subsequent visits taking place at 2-year intervals thereafter. WRAP participants were dementia-free at enrollment (mean age 54 years). All study procedures were approved by the University of Wisconsin School of Medicine and Public Health Institutional Review Board and are in accordance with the Declaration of Helsinki.

Wisconsin ADRC participants were recruited from multiple sources such as memory clinic providers, events, media, partner organizations, and other studies. Adults aged 45 years or older with decisional capacity and English fluency were eligible for enrollment. Exclusion criteria included active major medical or psychiatric illnesses and lack of a study partner. This study included participants who have normal cognitive status based on their diagnoses.

2.3 Alzheimer's disease outcomes

PACC3 is a global cognitive composite score used to assess the cognitive performance of participants in the WRAP study. During each study visit, participants complete a detailed cognitive battery, which is described in more detail elsewhere[21]. Longitudinal global cognitive performance is evaluated using a three-test version of the modified PACC[23], which includes the Rey Auditory Verbal Learning Test (AVLT; Trials 1–5)[24], the Wechsler Memory Scale Logical Memory II[25], and the Wechsler Adult Intelligence Scale-Revised Digit Symbol Substitution.[26] The Logical Memory II test was introduced at Visit 2 for most participants (94%), so the PACC3 baseline is generally at Visit 2, with Visit 1 being the baseline for recent enrollees where it was administered at Visit 1. Each test score is converted into a Z-score, using the mean and standard deviation from baseline scores of cognitively unimpaired individuals.

The Wisconsin ADRC uses a slightly different set of neuropsychological tests compared to WRAP. To maximize the use of available data, we collaborated with neuropsychologists at the Wisconsin ADRC to create a PACC-3-Trail Making Test (TMT) score for the replication analysis. Specifically, we converted the Craft Story score into an estimated Logical Memory score using a

published crosswalk table and followed previous methodology by substituting the Digit Symbol score with the total completion time from the TMT-B test.[23,27]. Since the crosswalk table only provides a conversion for the Craft Story score to the Logical Memory score for the first five visits, we limited the replication analysis to data from these five visits. The final PACC-3-TMT score was calculated by standardizing and averaging the results from the AVLT, the estimated Logical Memory score, and the reversed TMT-B test scores.

Plasma ptau217 assay results are from University of Wisconsin ADRC Biofluid Lab and were analyzed using the ALZpath pTau217 v2 assay on the Quanterix HD-X. Samples were collected between Aug 2011 and June 2023. 5,082 samples were analyzed using kit lot# 999008 and 309 samples were analyzed using kit lot# 999024. A bridging study was performed using 35 samples to determine the relationship between lots. The Pearson correlation coefficient was 0.97 ($p < 2.2e-16$) and linear regression was used to standardize the data from lot# 999024 to lot# 999008 ($y = 0.65514x + 0.02667$).

2.4 Plasma metabolite collections and quality control

2.4.1 Collection

Participants underwent venipuncture, most after fasting ≥8 hours, and provided thirty mL of blood into 3×10 mL lavender top EDTA tubes (BD 366643; Franklin Lakes, New Jersey, USA). Samples were mixed gently by inverting 10–12 times and were centrifuged 15 min at 2000 g at room temperature within 1 h of collection. Plasma samples were aliquoted into 2 mL cryovials (Wheaton Cryoelite W985863; Millville, New Jersey, USA). Aliquoted plasma was frozen at −80°C within 90 min and stored until overnight shipment to Metabolon, Inc. (Durham, NC), which similarly

kept samples frozen at -80° C until analysis. Participants who fasted <8 hours prior to venipuncture were excluded (n = 298).

2.4.2 Quality control

Samples with missing values for more than 40% of metabolites were removed (0 plasma samples) prior to analysis. Non-xenobiotic metabolites with missing values for >75% (n=43), and xenobiotic metabolites with missing values for >99% (n=75) of all samples were removed.

2.4.3 Fasting status

Samples from individuals who fasted less than 8 hours prior to venipuncture, or who did not provide fasting status were removed.

2.5 Genotyping, quality control, and imputation

2.5.1 Genotyping array data

DNA was extracted from whole blood samples and genotyped at the University of Wisconsin Biotech Center in three batches (2017, 2021, 2024) using the Illumina Multi-Ethnic Genotyping Array (MEGA).

2.5.2 Processing and Quality Control

Initial quality control was performed separately for each batch using PLINK v1.9[29]. SNPs mapped to chromosome 0 were removed, followed by variants missing in >5% of samples and samples with >5% missing variants. Additional filters were applied to remove variants missing in >2% of remaining samples and samples with >2% missing variants. SNPs within the ACTG nucleotide

bases were retained, and sex discrepancies were checked using X chromosome homozygosity, removing 10 samples with inconsistent reported sex. The data were remapped from hg37 to hg38 using the UCSC liftover tool[29]. The HRC checker tool was used to process the data before imputation, removing duplicates, mismatches, and monomorphic SNPs. Data were sorted by chromosome and saved as VCF files using BCFtools[31]. Imputation was done via the TOPMed Imputation Server using Eagle v2.4 and the TOPMed reference panel r3[31–33]. After imputation, variants with a low INFO score (R^2 <0.8) were removed. Data were merged and filtered for genotyping rate >98% and MAF >0.001. The final dataset consisted of 14,980,298 SNPs shared across all datasets.

2.5.3 Relatedness and genetic ancestry

A pruned set of the genotyped SNPs, using PLINK v1.9 --indep-pairwise 50 10 0.1, was used to estimate relatedness and genetic similarities to 1000 Genomes superpopulations[35]. Relationship inference was performed using KING v2.3.1[36] to identify related samples. The pruned SNPs, in a subset of unrelated samples, were used in a supervised learning model in ADMIXTURE v1.3.0, using the default maximum likdelihood estimation, to determin proportions of genetic similarities between the Wisconsin samples and the 1000 Genomes defined superpopulations (AFR, AMR, EAS, EUR, SAS). Then the subset of related samples was projected onto the resulting population structure output from ADMIXTURE[37].

2.5.4 *APOE* score

The *APOE* score is a weighted risk score for the effect of the six *APOE* genotypes on AD neuropathology. It is calculated using the natural log (ln) of the odds ratios (OR) for the six

different *APOE* genotypes, based on a study of AD neuropathology[39]. The *APOE* genotypes considered are ε2ε2, ε2ε3, ε3ε3, ε2ε4, ε3ε4, and ε4ε4, with ε3ε3 serving as the reference group.[40]

2.6 Gene expression imputation

The recently published pre-trained database, which used GTEx whole blood data, was obtained from the PredictDB Data Repository[41]. Genotypes from WRAP and Wisconsin ADRC participants were used to impute whole blood gene expression levels for the participants via PrediXcan. Specifically, genetic risk scores for predicting tissue-specific gene expression levels were pre-trained based on GTEx, which contains measured individual-level genetics and transcriptomics data. PrediXcan then calculates genetically predicted whole blood gene expression levels using genotypes of WRAP and Wisconsin ADRC participants, although transcriptomics data for these participants were not measured.

2.7 Statistical analyses

2.6.1 Sample selection

Participants were selected based on the availability of plasma metabolite measurements, genotypes, outcome data, and covariates such as age, sex, years of education, and *APOE* score. Among multiple visits, we selected the last observation for each participant, where the outcome was measured at or within six months after the plasma metabolites. Participants were then filtered based on their European genetic ancestry admixture (>0.5). For the Wisconsin ADRC cohort, participants diagnosed with dementia were excluded.

2.7.1. Pre-selecting potential AD-risk factors

A multivariate linear regression model was fitted for each biomarker (6,678 genes and 1,125 plasma metabolites) and each outcome in WRAP. In each model, a biomarker and covariates such as age, gender, *APOE* score, and years of education (for PACC3 only). were used as predictors to predict an outcome as follows:

$$PACC3 = age + gender + APOE\ score + years\ of\ education + gene$$

$$PACC3 = age + gender + APOE\ score + years\ of\ education + metabolite$$

$$ptau217 = age + gender + APOE\ score + gene$$

$$ptau217 = age + gender + APOE\ score + metabolite$$

Years of education were included only for PACC3, as PACC3 and years of education are correlated. For each outcome, genes and metabolites with a p-value < 0.05 were considered nominally significant biomarkers. These untargeted imputed transcriptomic and metabolomic analyses were performed to improve the computational efficiency of downstream machine learning modelling by focusing on potentially relevant associations.

2.7.2. Multi-omics integration and phenotype prediction

Significant biomarkers, pre-selected from untargeted imputed transcriptomics and metabolomics in WRAP, were input into the multi-omics models using COSIME[20], Cooperative Multi-view Integration and Scalable Interpretable Model Explainer. COSIME integrates multi-view data leveraging deep neural network encoders and Learnable Optimal Transport techniques and

implements a mechanism for assessing feature importance within each view, as well as quantifying feature interactions by estimating Shapley values and Shapley-Taylor indices.

Two models were trained for predicting PACC3 and ptau217, respectively. Hyperparameter optimization was performed using *Ray Tune* with Distributed Asynchronous Hyperparameter Optimization (*Hyperopt*). Early stopping was employed to prevent overfitting, based on performance improvements on the test data during the training phase. 5-fold cross-validation was used to evaluate the prediction performance of the best model for each fold. The best model was chosen based on the minimum prediction loss.

The prediction performance of the models was assessed using NRMSE to compare different scales of prediction. The NRMSE standardizes the error by normalizing it with respect to the interquartile range (IQR) of the observed data. The equation for NRMSE using IQR is given by:

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{IQR(y)}$$

Where y_i represents the actual values, $\hat{y}_i$ denotes the predicted values, and n is the number of participants. The term $IQR(y)$ refers to the interquartile range of actual values, which is calculated as $Q_3 - Q_1$, where Q_3 and $Q_1$are the third and first quartiles of the actual values, respectively.

2.7.3 Computing biomarker importance and interaction

Individual biomarker importance and interactions within each omics data type and across-omics data types in WRAP were computed using a trained model based on 75% of the samples and holdout data including the remaining 25% of the samples for each outcome. These calculations were conducted using COSIME.

3 RESULTS

3.1. Descriptive statistics for the participants

This study included 1,046 participants from the WRAP cohort for both outcomes: the PACC3 and ptau217, for the main analysis. Additionally, 85 participants for PACC3 and 89 participants for ptau217 were included from the Wisconsin ADRC cohort for independent validation. The demographic characteristics of individuals across cohorts and outcomes are described in **TABLE 1**.

TABLE 1. Demographic of participants in two cohorts.

Cohort	**WRAP**		**Wisconsin ADRC**	
Outcome	PACC3	ptau217	PACC3	ptau217
Total participants	1,046	1,046	85	89
Age (mean [SD])	67.04 (6.89)	67.20 (6.97)	69.14 (7.65)	68.99 (7.70)
Females	716 (68.5%)	716 (68.5%)	50 (58.8%)	54 (60.7%)

Males	330 (31.5%)	330 (31.5%)	35 (41.2%)	35 (39.3%)
APOE score (mean [SD])	0.63 (1.04)	0.63 (1.04)	0.61 (1.03)	0.64 (1.03)
Years of education (mean [SD])	15.80 (2.23)	-	16.88 (2.16)	-

3.2 Untargeted imputed transcriptomics and metabolomics

Out of 6,678 genes with imputed whole blood expression levels, 359 were significant for PACC3 and 348 for ptau217. Among 1,125 measured plasma metabolites, 150 were significant for PACC3 and 301 for ptau217 (**FIGURE 2**). These genes and metabolites were included in downstream machine learning modelling using COSIME.

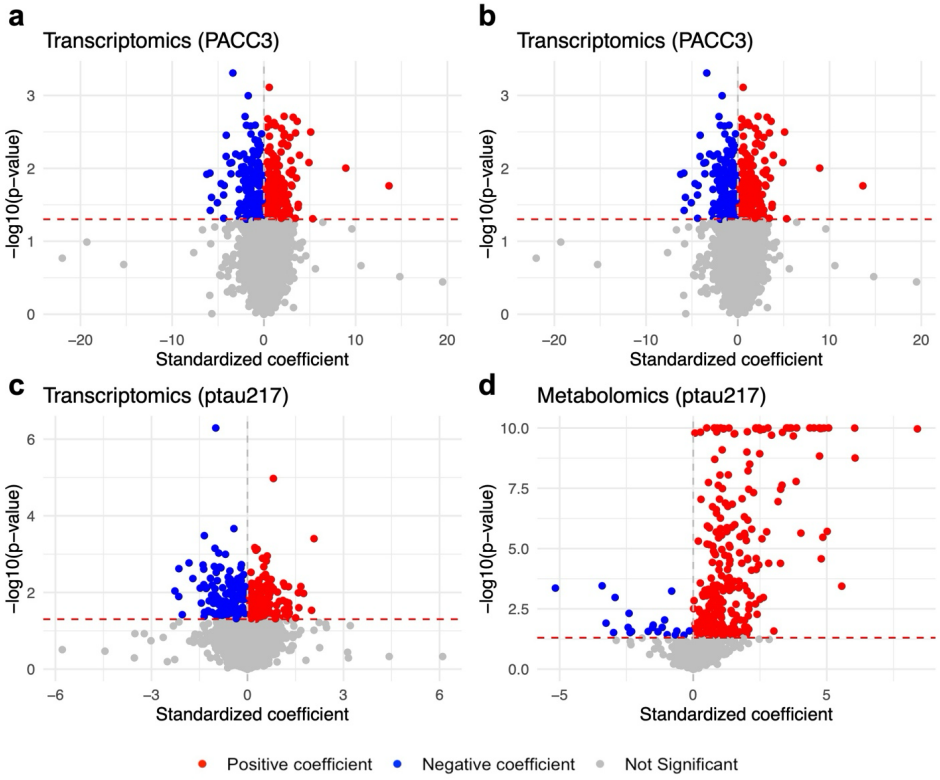


FIGURE 2

Volcano plots summarizing results of untargeted imputed transcriptomic and metabolomic analyses. P-values (y-axis) and magnitudes of associations (x-axis) for **a** untargeted imputed transcriptomics predicting PACC3, **b** untargeted metabolomics predicting PACC3, **c** untargeted imputed transcriptomics predicting ptau217, and **d** untargeted metabolomics predicting ptau217.

3.3 AD phenotype prediction

For the main analysis, the prediction performance of each model was evaluated five times (using 5-fold cross-validation) on holdout test data from the WRAP cohort. As an independent validation,

the trained COSIME models for the outcomes were used to assess their prediction performance on the Wisconsin ADRC cohort. As a result, the NRMSE was 0.654 ± 0.004 for predicting PACC3 and 1.361 ± 0.046 for predicting ptau217 in the WRAP cohort (**TABLE 2**). The NRMSE was 1.095 for predicting PACC3 and 1.795 for predicting ptau217 in the Wisconsin ADRC cohort (**TABLE 2**).

TABLE 2. Normalized Root Mean Squared Error (NRMSE) for each cohort and their outcomes.

Cohort	WRAP		Wisconsin ADRC	
Outcome	PACC3	ptau217	PACC3	ptau217
NRMSE	0.654 ± 0.004	1.361 ± 0.046	1.095	1.795

WRAP: mean ± 1.96 * standard deviation.

3.4 Biomarker importance and interactions

FIGURE 3a displays the top 20 genes ranked by their absolute biomarker importance values for predicting PACC3, while **FIGURE 3b** shows the top 20 metabolites with the highest biomarker importance values for predicting PACC3. Biomarkers with positive values for PACC3 may be down-regulated in AD or dementia-related pathways, whereas those with negative values may be up-regulated. **FIGURE 3c** presents the top 20 genes based on absolute biomarker importance values for predicting ptau217, and **FIGURE 3d** shows the top 20 metabolites for ptau217 prediction. For ptau217, biomarkers with positive values may be up-regulated in AD or dementia-related pathways, and those with negative values may be down-regulated. Several genes known to down-regulate cognitive function (e.g., *HOXA4* and *LRFN4*) and others associated with the up-regulation of ptau217 (e.g., *ACAD9* and *GPAT3*) were highly ranked based on our biomarker

importance scores. Several metabolites that down-regulate cognitive function (e.g., sphingomyelin (d17:2/16:0, d18:2/15:0) and N-carbamoylvaline) and those that up-regulate ptau217 (e.g., S-adenosylhomocysteine and propionylcarnitine (C3)) were also identified as highly ranked.

For PACC3 (**FIGURE 3a**), *HOXA4* and *LRFN4* are downregulated, while *ACAD9* and *GPTA3* were upregulated. In terms of metabolites (**FIGURE 3b**), sphingomyelin (d17:2/16:0, d18:2/15:0) and N-carbamoylvaline were downregulated, whereas S-adenosylhomocysteine and propionylcarnitine (C3) were upregulated. For ptau217 (**FIGURE 3c**), *LIPN* and *ARL11* were upregulated, while *CLEC4A* and *PDE1B* were downregulated. Lastly, for metabolites (**FIGURE 3d**), margoylcarnitine and N−acetyl−2−aminooctanoate were upregulated, while Undecenoylcarnitine (C11:1) and 2'-O-methyluridine were downregulated.

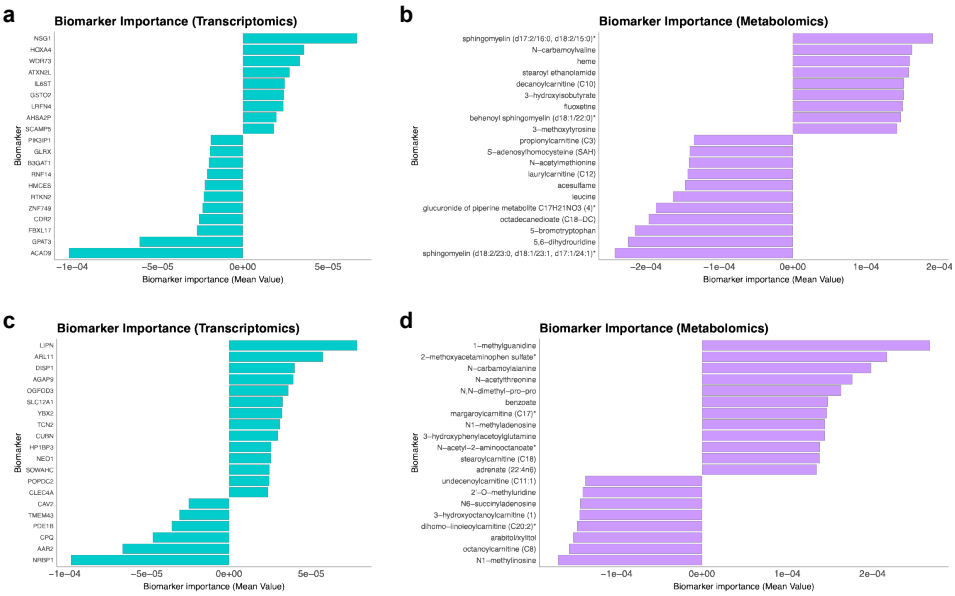


FIGURE 3

Biomarker importance values for PACC3 and ptau217. **a** Gene importance values for PACC3. **b** Metabolite importance values for PACC3. **c** Gene importance values for ptau217. **d** Metabolite importance values for ptau217.

The top 50 metabolite-metabolite pairs ranked by their absolute biomarker interaction values for predicting PACC3 are displayed in **FIGURE 4a** and the top 50 metabolite-metabolite pairs ranked by their absolute biomarker interaction values for predicting ptau217 are displayed in **FIGURE 4b**. Several pairs were notable, such as the potential antagonistic interaction between decadienedioic acid (C10:2−DC)** and tetradecadienoate (14:2) for PACC3 and the potential synergistic interaction between sphingomyelin (d18:2/16:0, d18:1/16:1) and choline for ptau217.

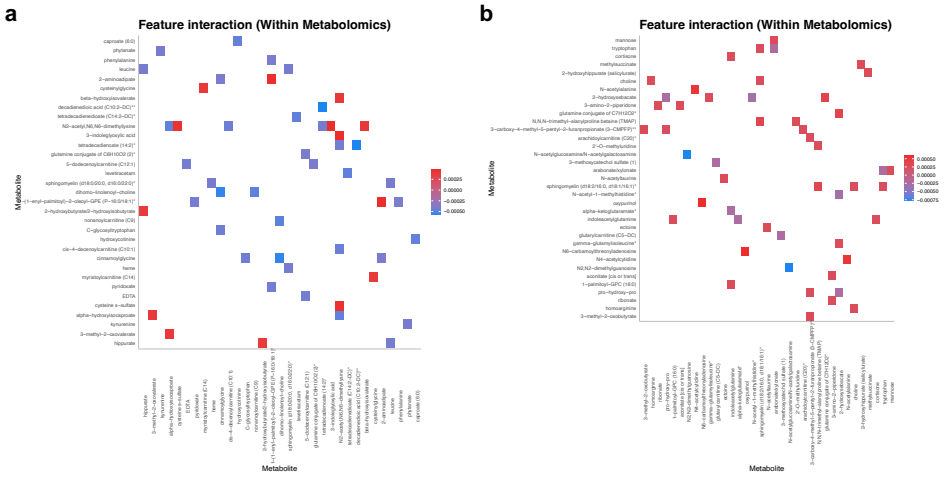


FIGURE 4

Metabolite pairwise interaction values for PACC3 and ptau217. **a** Metabolite pairwise interaction values for PACC3. **b** Metabolite pairwise interaction values for ptau217.

4 DISCUSSION

In this study, we aimed to predict two critical AD-related phenotypes—PACC3 and plasma ptau217—using multi-omics data, including imputed transcriptomics and metabolomics, combined with COSIME, a novel machine learning approach, for the WRAP cohort. We then assessed the individual biomarker importance and explored their pairwise interactions within the WRAP cohort. To validate the prediction performance of the trained models, we conducted independent validation by applying the same models to predict outcomes in the Wisconsin ADRC cohort. The analyses of biomarker importance and pairwise interactions provide valuable insights into potential AD risk factors and highlights key molecular features associated with disease progression.

Genes supporting neuronal growth and synaptic signaling are disrupted in AD. *HOXA4*, a key player in early brain development and neuronal differentiation, may be downregulated in AD, potentially impairing neuronal growth and connectivity[40]. Similarly, *LRFN4*, which supports synaptic communication, may also be reduced, leading to synaptic dysfunction—a core feature of AD[41]. *PDE1B*, important for cyclic nucleotide signaling and memory formation, could also be downregulated, disrupting synaptic plasticity and contributing further to cognitive deficits[42]. These downregulations are associated with lower PACC3 scores, while the absence of corresponding upregulation limits any compensatory effect linked to ptau217.

Disruptions in lipid metabolism impact membrane structure and function in AD. *GPAT3* supports lipid metabolism essential for maintaining cellular membranes and neuronal function[43]. In AD, dysregulation in lipid homeostasis may compromise these processes. *LIPN*, another lipid metabolism gene, might be upregulated in response to tau pathology, possibly acting as a

compensatory mechanism to preserve membrane integrity[44]. Sphingomyelins, such as sphingomyelin (d17:2/16:0, d18:2/15:0), are critical for membrane structure and signaling, and their downregulation in AD reflects impaired lipid processing[45]. Tetradecadienoate (14:2), which helps maintain membrane fluidity, may also be downregulated, further destabilizing membranes and exacerbating neurodegeneration[46]. The observed downregulations correspond with reduced PACC3 scores, while *LIPN* upregulation may indicate a ptau217-related compensatory stress response.

Mitochondrial dysfunction and impaired energy metabolism are consistent hallmarks of AD. *ACAD9* plays a central role in mitochondrial fatty acid metabolism, and its function is critical for sustaining neuronal energy needs[47]. Carnitines like propionylcarnitine (C3) and margaroylcarnitine (C17) facilitate fatty acid transport into mitochondria[48]. While C3 may be downregulated in AD, C17 could be upregulated in an attempt to support mitochondrial energy production. Similarly, undecenoylcarnitine (C11:1) may be reduced, indicating disrupted fatty acid metabolism and diminished neuronal energy capacity[49]. These mitochondrial expression patterns show that downregulations are associated with poorer PACC3 performance, whereas upregulated components such as C17 may be linked to ptau217 elevation.

Altered amino acid metabolism may reflect compensatory responses or metabolic breakdown. N-carbamoylvaline, involved in amino acid metabolism, could be downregulated in AD, reflecting general metabolic dysfunction[50]. In contrast, N-acetyl-2-aminooctanoate may be upregulated, potentially as an adaptive response to the altered metabolic environment in AD-affected neurons[50]. These changes suggest that amino acid metabolite downregulation is tied to reduced PACC3, while upregulation is related to increased ptau217 expression.

Epigenetic and RNA-processing disruptions contribute to AD pathology. S-adenosylhomocysteine (SAH), a key component in the methylation cycle, may be reduced in AD, suggesting impaired gene regulation and DNA repair capacity[51]. Similarly, 2'-O-methyluridine, a modified nucleoside involved in RNA metabolism, could be downregulated, pointing to stress in RNA processing and contributing to cognitive impairment[52]. These RNA and epigenetic disruptions are associated with decreased PACC3 scores, while the lack of compensatory upregulation suggests a minimal ptau217-related counter-response.

Immune system alterations may limit the brain's ability to respond to AD pathology. *CLEC4A*, involved in immune signaling and microglial function, may be downregulated in AD, which could impair the clearance of toxic proteins like tau[53]. On the other hand, *ARL11*, associated with cellular trafficking and signaling, might be upregulated in response to cellular stress from tau accumulation, reflecting a potential coping mechanism within the neuron[54]. These immune-related changes link *CLEC4A* downregulation to diminished PACC3 and *ARL11* upregulation to heightened ptau217 levels.

Metabolite interactions can amplify membrane instability and cognitive dysfunction. decadienedioic acid (C10:2−DC)**[55] and tetradecadienoate (14:2)[46] may interact antagonistically; their joint downregulation could destabilize both the structure and fluidity of membranes, accelerating neurodegeneration. In contrast, sphingomyelin (d18:2/16:0, d18:1/16:1)[56] and choline[57] might interact synergistically in the context of elevated ptau217, with disruptions in both potentially impairing acetylcholine synthesis and amplifying cognitive decline. Together, these metabolite changes show that downregulation contributes to lower PACC3 scores, while upregulated or synergistically altered molecules correlate with elevated ptau217.

While this study provides valuable insights into potential biomarkers and their interactions in AD, several limitations must be considered. First, the gene expression levels in this study were imputed based on genotypes rather than directly measured. The accuracy of the imputation depends on how precise the pre-trained models could capture true gene expression patterns, which vary strongly across genes and tissues. Therefore, although this approach enabled a more comprehensive analysis when measured transcriptomics data are unavailable, we acknowledge that it may introduce errors. Future studies could further improve accuracy by integrating more diverse reference datasets and refining the imputation models. Second, the use of the Wisconsin ADRC cohort as an independent validation set for the trained models may not provide a robust assessment of predictive performance due to its relatively small sample size. Utilizing larger independent cohorts with characteristics similar to WRAP and Wisconsin ADRC for validation would help improve the robustness of the models' predictive performance. To further deepen our understanding of AD, integrating a wider array of biomarkers—such as those from proteomics—would provide a more holistic view of its pathology. Furthermore, while this study focuses on specific biomarkers and their pairwise interactions, AD is a complex and multifactorial disease. Therefore, other factors, such as environmental influences and epigenetic changes, which were not captured in this analysis, could also play critical roles in disease progression. In the future study, investigating the complex interplay between genetic, environmental, and epigenetic factors will offer valuable insights into the multifactorial nature of AD progression.

ACKNOWLEDGMENTS

The authors especially thank the WRAP and Wisconsin ADRC participants and staff for their contributions to the studies. Without their efforts, this research would not be possible.

This study was supported by the National Institutes of Health (NIH) grants R01AG027161 (Wisconsin Registry for Alzheimer Prevention: Biomarkers of Preclinical AD), P30 AG062715 (the Wisconsin Alzheimer's Disease Research Center), and RF1AG054047 (Genomic and Metabolomic Data Integration in a Longitudinal Cohort at Risk for Alzheimer's Disease. Computational resources were supported by core grants from the Center for Demography and Ecology (P2CHD047873) and the Center for Demography of Health and Aging (P30AG017266).

CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests.

REFERENCES

1. DeTure, M. A. & Dickson, D. W. The neuropathological diagnosis of Alzheimer's disease. *Mol Neurodegeneration* **14**, 32 (2019).
2. O'Brien, R. J. & Wong, P. C. Amyloid precursor protein processing and Alzheimer's disease. *Annu Rev Neurosci* **34**, 185–204 (2011).
3. Sadigh-Eteghad, S. *et al.* Amyloid-beta: a crucial factor in Alzheimer's disease. *Med Princ Pract* **24**, 1–10 (2015).
4. Shankar, G. M. & Walsh, D. M. Alzheimer's disease: synaptic dysfunction and Abeta. *Mol Neurodegener* **4**, 48 (2009).

5. Cuddy, L. K. *et al.* Aβ-accelerated neurodegeneration caused by Alzheimer's-associated *ACE* variant R1279Q is rescued by angiotensin system inhibition in mice. *Sci. Transl. Med.* **12**, eaaz2541 (2020).
6. Killin, L. O. J., Starr, J. M., Shiue, I. J. & Russ, T. C. Environmental risk factors for dementia: a systematic review. *BMC Geriatr* **16**, 175 (2016).
7. Gatz, M. *et al.* Role of Genes and Environments for Explaining Alzheimer Disease. *Arch Gen Psychiatry* **63**, 168 (2006).
8. Horgusluoglu, E. *et al.* Integrative metabolomics-genomics approach reveals key metabolic pathways and regulators of Alzheimer's disease. *Alzheimers Dement* **18**, 1260–1278 (2022).
9. Lau, S.-F., Cao, H., Fu, A. K. Y. & Ip, N. Y. Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 25800–25809 (2020).
10. Rajmohan, R. & Reddy, P. H. Amyloid-Beta and Phosphorylated Tau Accumulations Cause Abnormalities at Synapses of Alzheimer's disease Neurons. *J Alzheimers Dis* **57**, 975–999 (2017).
11. Batra, R. *et al.* The landscape of metabolic brain alterations in Alzheimer's disease. *Alzheimer's & Dementia* **19**, 980–998 (2023).
12. Tönnies, E. & Trushina, E. Oxidative Stress, Synaptic Dysfunction, and Alzheimer's Disease. *JAD* **57**, 1105–1121 (2017).
13. Chu, S. H. *et al.* Integration of Metabolomic and Other Omics Data in Population-Based Study Designs: An Epidemiological Perspective. *Metabolites* **9**, 117 (2019).
14. Karch, C. M. & Goate, A. M. Alzheimer's disease risk genes and mechanisms of disease pathogenesis. *Biol Psychiatry* **77**, 43–51 (2015).
15. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol* **18**, 83 (2017).
16. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform Biol Insights* **14**, 1177932219899051 (2020).
17. Garg, M. *et al.* Disease prediction with multi-omics and biomarkers empowers case–control genetic discoveries in the UK Biobank. *Nat Genet* **56**, 1821–1831 (2024).

18. Hoffmann, J. *et al.* Machine learning in a data-limited regime: Augmenting experiments with synthetic data uncovers order in crumpled sheets. *Sci Adv* **5**, eaau6792 (2019).
19. Qiu, S. *et al.* Multimodal deep learning for Alzheimer's disease dementia assessment. *Nat Commun* **13**, 3404 (2022).
20. Choi, J. J. *et al.* COSIME: Cooperative multi-view integration and Scalable and Interpretable Model Explainer. Preprint at https://doi.org/10.1101/2025.01.11.632570 (2025).
21. Johnson, S. C. *et al.* The Wisconsin Registry for Alzheimer's Prevention: A review of findings and current directions. *Alzheimers Dement (Amst)* **10**, 130–142 (2018).
22. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091–1098 (2015).
23. Jonaitis, E. M. *et al.* Measuring longitudinal cognition: Individual tests versus composites. *Alz & Dem Diag Ass & Dis Mo* **11**, 74–84 (2019).
24. Vakil, E. & Blachstein, H. Rey auditory-verbal learning test: Structure analysis. *J. Clin. Psychol.* **49**, 883–890 (1993).
25. Atkinson, L. The Wechsler Memory Scale-Revised: Abnormality of selected index differences. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement* **24**, 537–539 (1992).
26. Wechsler, D. Wechsler Adult Intelligence Scale--Fourth Edition. https://doi.org/10.1037/t15169-000 (2012).
27. Monsell, S. E. *et al.* Results From the NACC Uniform Data Set Neuropsychological Battery Crosswalk Study. *Alzheimer Disease & Associated Disorders* **30**, 134–139 (2016).
28. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742-015-0047–8 (2015).
29. Nassar, L. R. *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Research* **51**, D1188–D1195 (2023).
30. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

31. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284–1287 (2016).
32. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782–784 (2015).
33. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
34. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).
35. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
36. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
37. Reiman, E. M. *et al.* Exceptionally low likelihood of Alzheimer's dementia in APOE2 homozygotes from a 5,000-person neuropathological study. *Nat Commun* **11**, 667 (2020).
38. Deming, Y. *et al.* Neuropathology-based *APOE* genetic risk score better quantifies Alzheimer's risk. *Alzheimer's & Dementia* **19**, 3406–3416 (2023).
39. Liang, Y., Nyasimi, F. & Im, H. K. Pervasive polygenicity of complex traits inflates false positive rates in transcriptome-wide association studies. Preprint at https://doi.org/10.1101/2023.10.17.562831 (2023).
40. Li, Q. S. *et al.* Association of peripheral blood DNA methylation level with Alzheimer's disease progression. *Clin Epigenet* **13**, 191 (2021).
41. de Wit, J. & Ghosh, A. Control of neural circuit formation by leucine-rich repeat proteins. *Trends Neurosci* **37**, 539–550 (2014).
42. Ahmad, N., Lesa, K. N., Sudarmanto, A., Fakhrudin, N. & Ikawati, Z. The role of Phosphodiesterase-1 and its natural product inhibitors in Alzheimer's disease: A review. *Front Pharmacol* **13**, 1070677 (2022).

43. Wang, Y. *et al.* Midlife Chronological and Endocrinological Transitions in Brain Metabolism: System Biology Basis for Increased Alzheimer's Risk in Female Brain. *Sci Rep* **10**, 8528 (2020).
44. Helbecque, N., Cottel, D. & Amouyel, P. Low-density lipoprotein receptor-related protein 8 gene polymorphisms and dementia. *Neurobiology of Aging* **30**, 266–271 (2009).
45. Baloni, P. *et al.* Multi-Omic analyses characterize the ceramide/sphingomyelin pathway as a therapeutic target in Alzheimer's disease. *Commun Biol* **5**, 1074 (2022).
46. Hammond, T. C. *et al.* Human Gray and White Matter Metabolomics to Differentiate APOE and Stage Dependent Changes in Alzheimer's Disease. *J Cell Immunol* **3**, 397–412 (2021).
47. He, M. *et al.* A New Genetic Disorder in Mitochondrial Fatty Acid β-Oxidation: ACAD9 Deficiency. *The American Journal of Human Genetics* **81**, 87–103 (2007).
48. Nho, K. *et al.* Serum metabolites associated with brain amyloid beta deposition, cognition and dementia progression. *Brain Communications* **3**, fcab139 (2021).
49. Hudson, S. A. & Tabet, N. Acetyl-l-carnitine for dementia. *Cochrane Database of Systematic Reviews* (2003) doi:10.1002/14651858.CD003158.
50. Griffin, J. W. D. & Bradshaw, P. C. Amino Acid Catabolism in Alzheimer's Disease Brain: Friend or Foe? *Oxid Med Cell Longev* **2017**, 5472792 (2017).
51. Panza, F. *et al.* Possible Role of S-Adenosylmethionine, S-Adenosylhomocysteine, and Polyunsaturated Fatty Acids in Predementia Syndromes and Alzheimer's Disease. *JAD* **16**, 467–470 (2009).
52. Zhang, X. *et al.* Small RNA modifications in Alzheimer's disease. *Neurobiol Dis* **145**, 105058 (2020).
53. Kaur, G. *et al.* DNA Methylation: A Promising Approach in Management of Alzheimer's Disease and Other Neurodegenerative Disorders. *Biology* **11**, 90 (2022).
54. Yin, R.-H., Yu, J.-T. & Tan, L. The Role of SORL1 in Alzheimer's Disease. *Mol Neurobiol* **51**, 909–918 (2015).
55. Schuck, P. F. *et al.* cis-4-decenoic acid provokes mitochondrial bioenergetic dysfunction in rat brain. *Life Sciences* **87**, 139–146 (2010).

552 56. Javaid, S. *et al.* Dynamics of Choline-Containing Phospholipids in Traumatic Brain Injury and
553 Associated Comorbidities. *Int J Mol Sci* **22**, 11313 (2021).
554 57. Yuan, J. *et al.* Is dietary choline intake related to dementia and Alzheimer's disease risks? Results from
555 the Framingham Heart Study. *Am J Clin Nutr* **116**, 1201–1207 (2022).
556

Conclusion

The integration of multi-omics data holds immense potential for unraveling the complex molecular mechanisms underlying AD. In this dissertation, I have successfully executed three ambitious aims that leverage cutting-edge techniques to address fundamental questions in AD research. The work presented here not only provides valuable insights into the molecular and genetic architecture of AD but also contributes to the advancement of precision medicine by identifying novel biomarkers and risk factors for AD.

In Aim 1, we identified cooperative TF pairs that regulate target genes in oligodendrocytes. By combining scRNA-seq and scATAC-seq data, we uncovered key TF interactions that play a pivotal role in the regulation of oligodendrocyte-specific genes. We identified several highly cooperative TF pairs, such as SOX10 and OLIG2, which are already known, and additionally discovered previously unreported cooperative pairs, such as SOX10 and NKX2.2. The use of deep learning models allowed for accurate predictions of target gene expression, revealing critical insights into how gene regulation is disrupted in AD. Additionally, the independent validation of these TF pairs through eQTLs further strengthened our findings and demonstrated the relevance of these regulatory interactions to neurodegenerative diseases like AD. This framework provides a deeper understanding of the GRNs involved in oligodendrocyte function and lays the foundation for future therapeutic strategies targeting transcriptional dysregulation in AD.

Aim 2 focused on developing COSIME (Cooperative Multi-view Integration and Scalable Interpretable Model Explainer), a novel machine learning approach designed to integrate multi-omics data and predict AD phenotypes. The integration of unsupervised neural networks with optimal transport methods proved to be an effective strategy for combining diverse data types,

such as genomics and transcriptomics, improving the accuracy of AD phenotype prediction. COSIME also provides a level of interpretability, allowing us to gain insights into the relative importance of features and their interactions within and across different omics modalities. Using COSIME, we identified that synergistic interactions between microglia and astrocyte genes associated with AD are more likely to be active at the edges of the middle temporal gyrus, as revealed through spatial transcriptomics data. We also constructed a gene-gene interaction network, which uncovered several astrocyte-specific MERFISH genes—such as *PLCB1* and *HPSE2*—as key hub genes linking multiple microglia genes. *PLCB1*, in particular, is involved in phosphatidylinositol signaling and plays a central role in inflammation and synaptic function, while *HPSE2* modulates heparan sulfate metabolism and may influence neuroinflammatory pathways in AD. This novel machine learning method has the potential to transform how we analyze complex, high-dimensional biological data, offering a robust framework for predicting disease outcomes and identifying novel AD risk factors.

In Aim 3, the application of COSIME to clinical multi-omics data from the Wisconsin Registry for Alzheimer's Prevention (WRAP) and the Wisconsin Alzheimer's Disease Research Center (ADRC) cohorts led to the identification and replication of key genomic and metabolomic risk factors for AD. By integrating transcriptomic and metabolomic data, we aimed to uncover how interactions between genes and metabolites influence AD phenotypes. For example, metabolite interactions such as the antagonistic relationship between sphingomyelin (d17:2/16:0, d18:2/15:0) and tetradecadienoate (14:2), which destabilize membrane structure and fluidity, could accelerate neurodegeneration. In contrast, sphingomyelin (d18:2/16:0, d18:1/16:1) and choline may interact synergistically in the presence of elevated ptau217, potentially impairing acetylcholine synthesis and amplifying cognitive decline. These metabolite changes are associated

with lower PACC3 scores when downregulated and higher ptau217 levels when upregulated. These findings have important implications for understanding the molecular interactions underlying AD, including gene-gene, metabolite-metabolite, and gene-metabolite relationships and could serve as the basis for the development of personalized treatments by enabling targeted therapies that address individual molecular profiles and specific interaction patterns unique to each patient. The ability to predict AD risk based on multi-omics data is a significant step toward precision medicine, with further work needed to validate these findings in larger, diverse cohorts and translate them into targeted therapeutic interventions tailored to the specific molecular signatures identified in high-risk individuals, as it enables the identification of individuals at high risk of developing AD before the onset of clinical symptoms.

The results of this dissertation contribute to a deeper understanding of the genetic and metabolic factors that drive AD and provide a new computational framework for integrating and analyzing multi-omics data. By identifying cooperative TFs, discovering novel biomarkers, and uncovering complex gene-metabolite interactions, this research paves the way for more targeted interventions in AD. For instance, the identification of interacting biomarkers—such as gene and metabolite biomarkers that co-regulate critical pathways in neurodegeneration—can allow for the development of combination therapies that target these pathways simultaneously, improving treatment efficacy by addressing multiple disease mechanisms at once. As we continue to improve the integration of multi-omics data, we move closer to the ultimate goal of precision medicine: providing personalized, effective treatments for individuals with AD.

Ultimately, this work demonstrates that a systems biology approach, integrating genomics, transcriptomics, and metabolomics, is crucial for advancing our understanding of AD pathology.

The application of novel computational tools like COSIME will allow for more accurate predictions, better identification of biomarkers, and the development of targeted therapies that address the complex nature of AD. This dissertation sets the stage for future research in neurodegenerative diseases, providing new directions for the diagnosis, prevention, and treatment of AD and other related cognitive disorders.

Building upon the findings and methodologies presented in this dissertation, several avenues for future research can be explored to further enhance our understanding of AD and improve the predictive capabilities of multi-omics integration models. One important direction is the exploration of high-order feature interactions, which could reveal more intricate relationships between genes, metabolites, and other molecular factors. Advanced clustering techniques could also be employed to identify subgroups within the AD population, potentially leading to the discovery of novel disease subtypes or risk profiles. Another promising development would be the incorporation of causal inference techniques into the COSIME model explainer. This would allow for a deeper understanding of the causal relationships between molecule factors in AD, thereby enabling more accurate predictions of disease progression. Finally, extending COSIME to handle longitudinal data represents a crucial next step, as it would allow for the analysis of temporal changes in multi-omics data, offering new insights into how AD evolves over time. These advancements would significantly improve our ability to predict disease onset and progression, leading to more personalized and effective therapeutic strategies.

Appendix (Publication) – CHAPTER 1: CoTF-reg reveals cooperative transcription factors in oligodendrocyte gene regulation using single-cell multi-omics

communications biology

Article

https://doi.org/10.1038/s42003-025-07570-6

CoTF-reg reveals cooperative transcription factors in oligodendrocyte gene regulation using single-cell multi-omics

Check for updates

Jerome J. Choi [1,2], John Svaren [1,3,6] & Daifeng Wang [1,4,5,6]

Oligodendrocytes are the myelinating cells within the central nervous system, but the mechanisms by which transcription factors (TFs) cooperate for gene regulation in oligodendrocytes remain unclear. We introduce coTF-reg, an analytical framework that integrates scRNA-seq and scATAC-seq data to identify cooperative TFs co-regulating the target gene (TG). First, we identify co-binding TF pairs in the same oligodendrocyte-specific regulatory regions. Next, we train a deep learning model to predict each TG expression using the co-binding TFs' expressions. Shapley interaction scores reveal high interactions between co-binding TF pairs, such as SOX10-TCF12. Validation using oligodendrocyte eQTLs and their eGenes that are regulated by these cooperative TFs show potential regulatory roles for genetic variants. Experimental validation using ChIP-seq data confirms some cooperative TF pairs, such as SOX10-OLIG2. Prediction performance of our models is evaluated through holdout data and additional datasets, and an ablation study is also conducted. The results demonstrate stable and consistent performance.

Oligodendrocytes play key functional roles in the central nervous system (CNS) function, including that they are responsible for myelination[1,2]. Myelination is a complex neurodevelopmental process that begins during brain development in the third trimester of pregnancy and increases steadily during childhood, but it can also be dynamically regulated in the context of learning and diseases affecting the mature CNS[3,4]. Also, Oligodendrocyte dysfunction and myelin abnormalities have been reported in CNS disorders[2,5,6]. Multidirectional interactions between neuronal and glial cells are required for CNS function[7], including interactions between oligodendrocytes and neurons through myelination[8]. Therefore, it is critical to better understand the functions and roles of oligodendrocytes and myelin.

Gene expression of oligodendrocyte development from oligodendrocyte progenitor cells (OPC) is governed by complex gene regulatory mechanisms involving transcription factors (TFs)[3,4]. TFs often work in a combinatorial fashion to regulate gene expression from regulatory elements[9,10]. For example, some TFs such as SOX10 and OLIG2 cooperate during the induction of genes for differentiation and myelin formation[11–14]. Enhancers can increase transcription levels from promoters and transcription start sites (TSS), and much of the regulatory code that drives cell type-specific gene expression resides in these distal regulatory elements. Especially, some active enhancers are associated with the gene expression that characterizes cell identity and functions[15]. Thus, it is important to identify active oligodendrocyte-specific enhancers as well as promoters and the co-binding TFs that are responsible for their activity.

Next-generation sequencing technologies, including single-cell RNA sequencing (scRNA-seq) and the assay for transposase-accessible chromatin sequencing (scATAC-seq), have provided important insights into cell-type-specific gene regulation. Recent functional genomic resources such as PsychENCODE2[16] and GTEx[17], and emerging tools for integrating multi-omics data enable creating cell-type-level gene regulatory networks (GRNs) linking TFs and their binding sites (TFBS), regulatory elements to target genes (TGs). Those networks can reveal the cell-type-specific regulatory roles of TFs via regulatory elements. Moreover, additional bioinformatic tools such as SCENIC+[18], Signac[19], and scGRNom[20] predict cell-type-specific gene regulatory networks to explain potential TF-TG relationships. However, most of these studies and tools focus on relationships between

[1]Waisman Center, University of Wisconsin-Madison, Madison, WI, USA. [2]Department of Population Health Sciences, University of Wisconsin-Madison, Madison, WI, USA. [3]Department of Comparative Biosciences, School of Veterinary Medicine, University of Wisconsin-Madison, Madison, WI, USA. [4]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA. [5]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA. [6]These authors jointly supervised this work: John Svaren, Daifeng Wang. e-mail: daifeng.wang@wisc.edu

https://doi.org/10.1038/s42003-025-07570-6

individual TFs and TGs instead of TF-TF interactions and their effects on TG expression. Consequently, due to the lack of tools, the mechanistic roles of cooperative TFs in establishing cell type-specific gene regulation remain uncharacterized.

To tackle these challenges, we introduce coTF-reg, an analytical framework that integrates scRNA-seq and scATAC-seq data to identify cooperative TFs co-regulating the TG. coTF-reg identifies cooperative co-binding TFs along with active regulatory elements for gene regulation as hallmarks of active oligodendrocyte-specific regulatory elements. First, it identifies co-binding TF pairs in these regulatory regions. Second, a deep learning model is trained to predict TG expression based on the expression profiles of co-binding TFs. Third, Shapley interaction scores are computed to evaluate the interactions between TF pairs. Our findings reveal high interactions between co-binding TF pairs, such as SOX10-TCF12. Validation using oligodendrocyte eQTLs and their eGenes that are regulated by these cooperative TFs showed potential regulatory roles for genetic variants. Experimental validation using ChIP-seq data confirmed some cooperative TF pairs, such as SOX10-OLIG2 and SOX10-NKX2.2. Prediction performance of our models was evaluated through holdout data and additional datasets, and an ablation study was also conducted. The results demonstrated stable and consistent performance. Overall, our results create an analytic framework in which co-binding TF pairs cooperatively activate the TG expression through oligodendrocyte-specific regulatory elements.

Results

Deep learning and single-cell multi-omics for identifying cooperative transcription factors in oligodendrocytes

In order to predict cooperative TFs involved in oligodendrocyte gene regulation, we designed coTF-reg, which integrates scRNA-seq and scATAC-seq data to identify the cooperative TFs that co-regulate the target gene (TG) expression in oligodendrocytes (Fig.1, Methods and Materials). Briefly, we first used scATAC-seq data with peak-to-gene links[21]. Second, among the regulatory regions for various cell types, we focused on those specific to oligodendrocytes. We then identified transcription factor binding sites (TFBSs) and co-binding TF pairs through motif co-occurrence and co-enrichment analyses. Third, we trained deep neural networks (DNNs) to predict the expression levels of the TGs and measure interaction effects between co-binding TFs on the expression levels of TGs using gene expression from scRNA-seq data[22] and computed Shapley interaction (SI)[23,24] scores for co-binding TF pairs and found cooperative TF pairs. Fourth, we built a gene regulatory network based on SI scores for co-binding TF pairs. Lastly, as an independent validation, to validate the cooperative TF pairs we found, we mapped oligodendrocyte eQTLs onto the regulatory regions where cooperative TF pairs exist, performed Liftover analysis and co-enrichment analysis using ChIP-seq data, and applied Boolean rules to characterize the cooperativity of regulatory factors. To evaluate the prediction performance of our models, we used other publicly available datasets and conducted an ablation study by generating random TF sets to predict TG expressions.

Identification of the co-binding transcription factors in oligodendrocyte-specific regulatory regions

First, we identified a set of 787 oligodendrocyte-differentially accessible and oligodendrocyte-specific regulatory regions by comparison of oligodendrocyte scATAC-seq data to other brain cell types. In this set, we identified 958 motifs for inferred TFBSs using the JASPAR database. Second, we used co-occurrence analysis and co-enrichment analysis to identify 8101 co-binding TF pairs out of 458,403 possible TF pairs ('Methods and Materials: Co-enrichment analysis' for more details). We removed TF pairs from the same families and applied a cutoff (<0.1) for false discovery rate (FDR) yielding 8101 co-binding TF pairs. There were 206 TFs that have co-binding TFs linked to 445 TGs (Supplementary Data) that are oligodendrocyte specific in 643 regulatory regions (Fig. 2a). We annotated the regulatory regions to categorize them into promoters (32.5%) and enhancers (67.5%) (Fig. 2b).

The density plots show the distributions of the number of co-binding TF pairs, the number of TGs, and the number of peaks for individual TFs that are co-bound to other TFs. Most of the TFs have 50 to 103 co-binding TFs (median = 78). The distribution of the number of TGs for TFs is right-

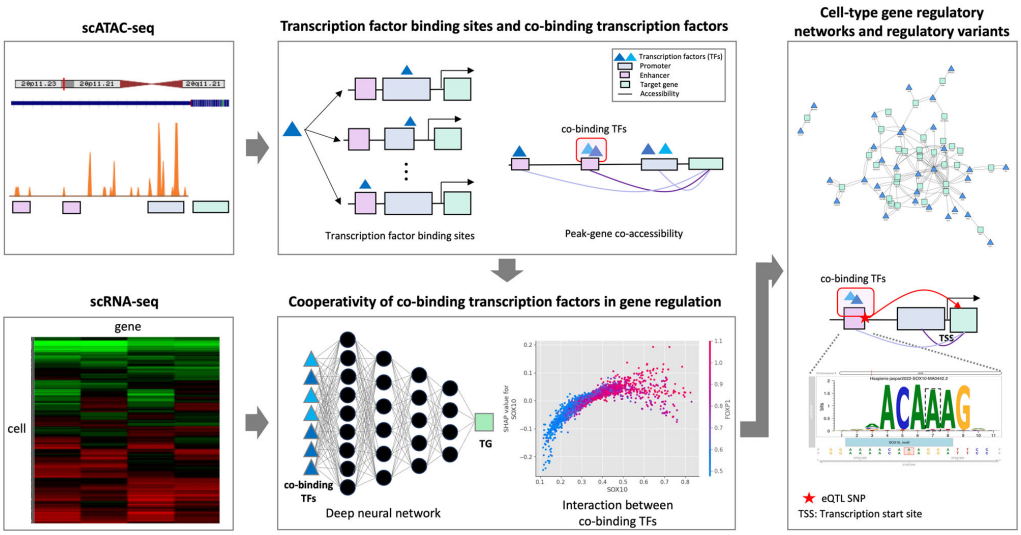


Fig. 1 | Deep learning and single-cell multi-omics for identifying cooperative transcription factors in oligodendrocytes. Inputs for the coTF-reg pipeline are scATAC-seq peak-gene links and scRNA-seq. It infers transcription factor binding sites (TFBSs) in regulatory regions and identifies co-binding TF pairs. Then, it measures cooperativity of co-binding TFs by predicting TF-TG relationships for the levels of expression using deep learning models and Shapley interaction scores. It outputs a gene regulatory network linking co-binding TF pairs with their TGs and regulatory variants on the regulatory regions where co-TFs have their binding sites.

https://doi.org/10.1038/s42003-025-07570-6

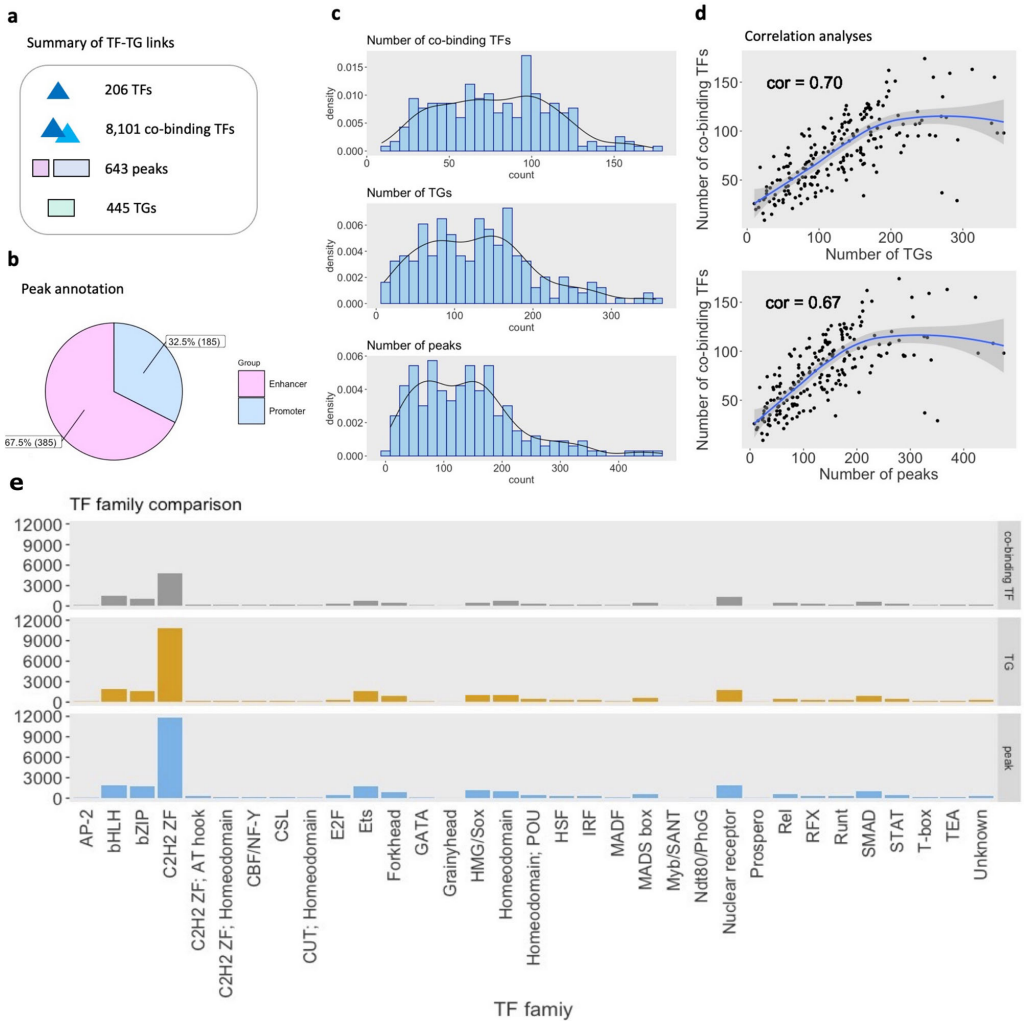


Fig. 2 | Distribution and correlation of the numbers of co-binding transcription factors, target genes, and peaks for individual transcription factors, peak annotation, and summary statistics for transcription factor-target gene links. **a** Summary statistics of transcription factor (TF)-target gene (TG) links. **b** Peak annotation. **c** Distributions of the numbers of co-binding TFs, TGs, and peaks for individual transcription factors. **d** Correlations between the numbers of co-binding transcription factors and target genes, as well as the numbers of co-binding TFs and peaks. **e** Distribution of the numbers of co-binding TF pairs, TGs, and peaks for transcription factors by families.

skewed, and many TFs have 76 to 172 TGs linked. The distribution of the number of peaks for TFs is also right skewed, and the most frequent intervals were between 75 and 180 peaks (Fig. 2c). The distributions of the numbers of TGs and peaks for co-binding TF pairs are approximately normal. On average, co-binding TF pairs have 60 TGs linked and 59 peaks (Supplementary Fig. 1). Additionally, other density plots show the distributions of the number of TGs and the number of peaks for co-binding TF pairs and bar plots display the numbers of co-binding TFs, TGs, and peaks for individual TFs by their family categories (Fig. 2d). Co-binding TFs have 4 to 115 TGs (median = 59) and 4 to 123 peaks (median = 56) and the most frequent motifs are associated with TF families with C2H2 zinc finger (ZF), bZip, and bHLH DNA-binding domains (Fig. 2e). C2H2 ZF proteins are a large family and C2H2 ZF TFs (e.g., ZNF24[25] and KLF9/13[26]) are known to play significant roles in the development and function of oligodendrocytes, which are the myelinating cells of the CNS. These TFs can regulate the expression of genes essential for oligodendrocyte differentiation, survival, and myelination processes[25,27,28].

We computed Pearson correlation coefficient (r) to measure correlations between the number of co-binding TFs and the number of TGs and the number of co-binding TFs and the number of peaks for individual TFs (Fig. 2d). The number of co-binding TFs and the number of TGs for individual TFs are strongly positively correlated ($r = 0.70$). It suggests that TFs that are co-bound to other TFs tend to have more TGs linked to them. The number of co-binding TFs and the number of peaks for individual TFs are also strongly positively correlated ($r = 0.67$). It shows that co-binding TFs may exist in many different peaks.

https://doi.org/10.1038/s42003-025-07570-6

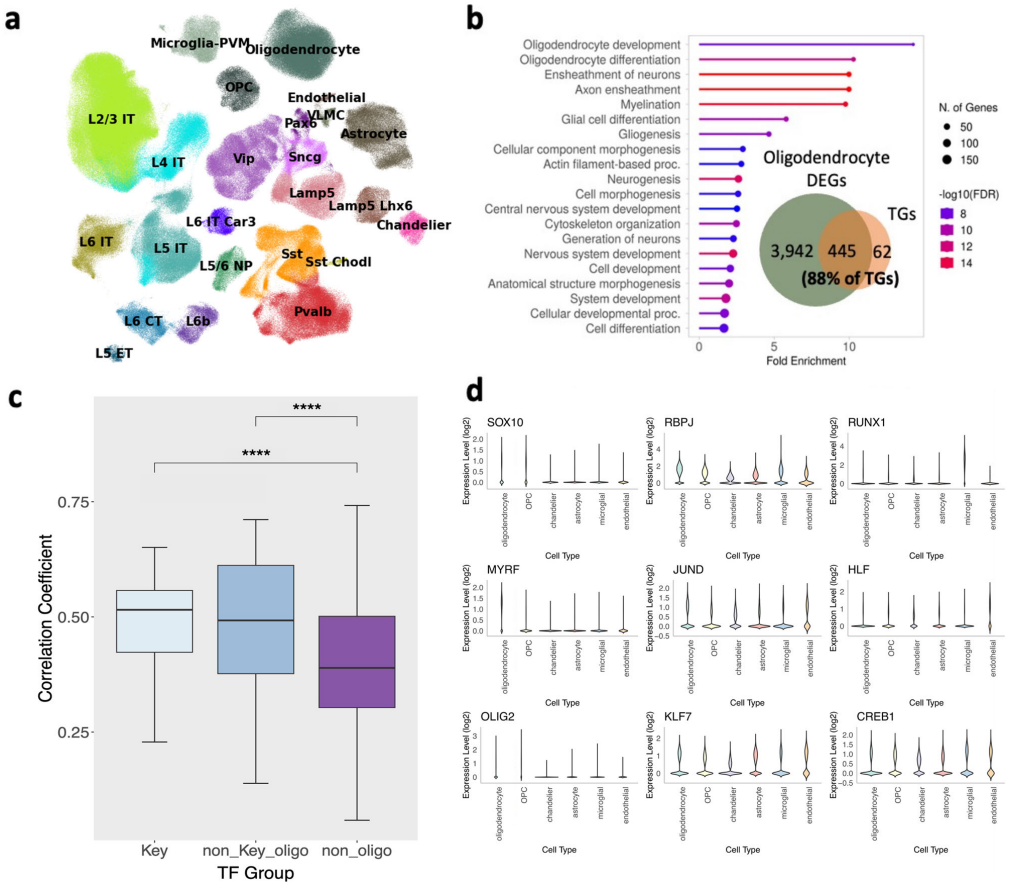


Fig. 3 | Oligodendrocyte gene expression relationships between transcription factors and target genes. a UMAP for eighteen cell type in middle temporal gyrus region, **b** Enrichments analysis for target genes that are oligodendrocyte-specific, **c** Pairwise two-sided t-tests for the correlation (between the expression of the TFs and their TGs) comparison for three TF groups: 10 oligodendrocyte key transcription factors, 83 oligodendrocyte-specific non-key transcription factors, and 103 non-oligodendrocyte-specific transcription factors, and **d** Boxplots for the expression levels of the three categories in (**c**) (each column is an example for each category).

In the following sections, we incorporate RNA-seq data to explore gene expression relationships between TFs and TGs, train deep learning models to predict TG expression using co-binding TFs, and compute TF interaction scores using the trained models.

Oligodendrocyte gene expression relationships between transcription factors and target genes

A single cell study identified the unique gene expression profile of oligodendrocytes compared to other brain cell types[22], as shown by the two dimensional Uniform Manifold Approximation and Projection (UMAP) space after computing latent representations of the neighborhood graph (Fig. 3a). The UMAP embeddings reveal that oligodendrocytes exhibit a distinct expression profile compared to other cell types. This separation suggests that oligodendrocytes have unique transcriptional programs that differentiate them from neighboring cell types. The distinct clustering of oligodendrocytes in the UMAP space indicates specialized functional roles and may reflect their involvement in myelination and maintenance of neural integrity. In order to focus on oligodendrocyte-specific mechanisms of gene regulation, we conducted differential expression testing using 17,946 genes and 20,191 metacells and identified 4387 differentially expressed genes (DEGs) for oligodendrocytes. We found 445 TGs out of 507 TGs of oligodendrocyte-specific regulatory elements (88%) were DEGs for oligodendrocytes. Subsequently, we conducted enrichment analysis for these 445 TGs revealing their involvement in crucial biological processes for oligodendrocytes, such as oligodendrocyte development, oligodendrocyte differentiation, and myelination (Fig. 3b).

We categorized TFs into oligodendrocyte key TFs, oligodendrocyte-specific non-key TFs, and non-oligodendrocyte-specific TFs using oligodendrocyte expression level and the list of key TFs (see 'Methods and Materials: Key TFs' for more details). 'Oligodendrocyte-specific key TFs' are oligodendrocyte differentially expressed TFs and key TFs, 'oligodendrocyte-specific non-key TFs' are oligodendrocyte differentially expressed TFs but not key TFs, and 'non-oligodendrocyte-specific TFs' are neither oligodendrocyte differentially expressed TFs nor key TFs. The key oligodendrocyte TFs were defined based on mouse loss-of-function studies that have shown that specific TF's are critical for oligodendrocyte differentiation. The key TF's include SOX10[29], SOX2[30,31], SOX8[32], MYRF[33], OLIG1[34], OLIG2[35], TCF7L2[36,37], ZNF24[25], NKX2.2[38], and NKX6.2[39].

https://doi.org/10.1038/s42003-025-07570-6

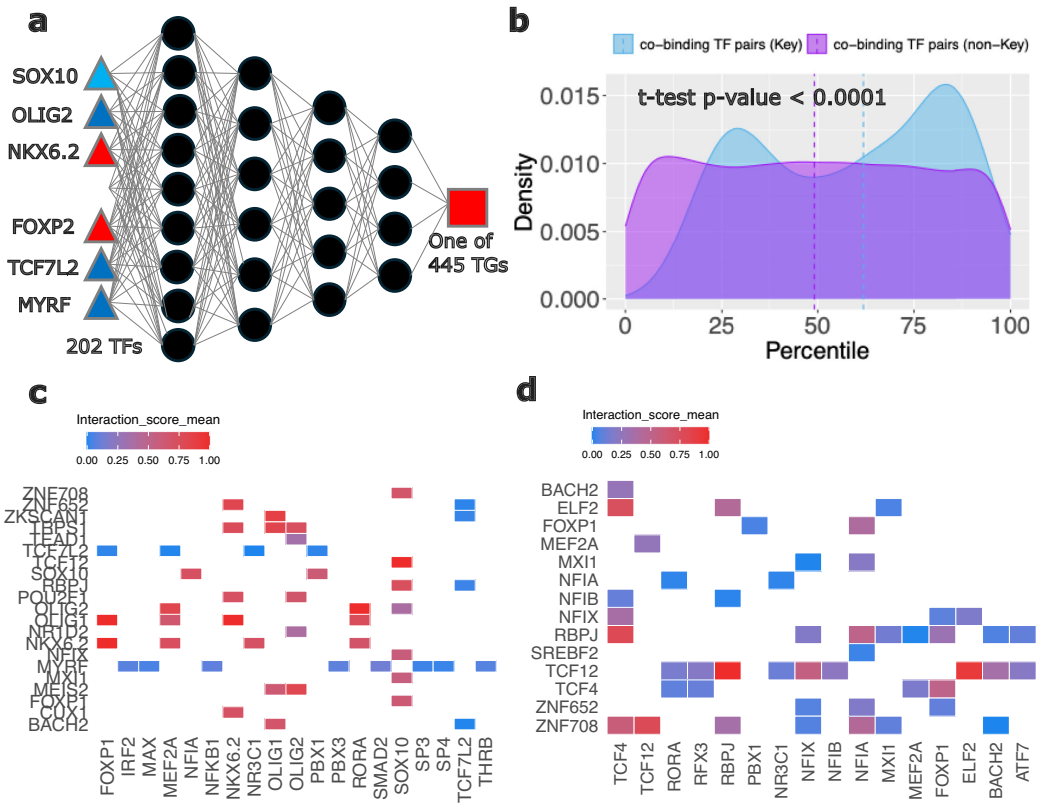


Fig. 4 | Cooperative transcription factor pairs by Shapley interaction scores. **a** Deep learning architecture, **b** Percentile distribution of interaction scores for 577 key-transcription factor pairs and 7029 non-key transcription factor pairs, **c** Top forty-eight SI interaction scores for key transcription factor pairs, and **d** Top forty-eight SI interaction scores for non-key transcription factor pairs.

Each of 206 TF, who have co-binding TFs, regulates a different set of TGs, and we computed correlations between the expression of the TFs and their TGs in the three categories. Pairwise two-sided t-tests show that correlations between TFs and TGs in oligodendrocyte key TF pairs and those in non-oligodendrocyte-specific TF pairs are significantly different ($p < 0.001$). It also indicates that correlations between TFs and TGs in oligodendrocyte-specific non-key TF pairs and those in non-oligodendrocyte-specific TF pairs are significantly different ($p < 0.001$) (Fig. 3c). The results for differential expression testing show that the six TFs in the two categories, oligodendrocyte key TF pairs and oligodendrocyte-specific non-key TF pairs are all significant and up-regulated (Fig. S8).

We color-coded the UMAP embeddings based on the expression level of the TFs (Fig. 3a) and selected three TFs as examples for each category. Oligodendrocyte-specific key TFs such as SOX10, MYRF, and OLIG2 are specifically highly expressed in oligodendrocytes. Oligodendrocyte-enriched non-key TFs, including RBPJ, JUND, and KLF7, are expressed in multiple cell types but are more highly expressed in oligodendrocytes. Non-oligodendrocyte-specific TFs, such as RUNX1, HLF, and CREB1, are not specifically expressed in oligodendrocytes (Fig. 3d).

Deep learning and Shapley interaction scores to measure cooperativity of co-binding transcription factors

To understand the complex relationships between TFs for predicting TGs, we built deep learning models. We trained a deep learning model for each of the 445 TG. Each model used the expression levels for the 206 TFs that have co-binding TFs to predict a TG expression level. We used seven hidden layers in each DNN (Fig. 4a). We excluded co-binding TF-TG pairs that exhibited high variability in their SI scores (coefficient of variance > 0.5). Using a trained model and a hold-out test dataset, we computed SI scores for TFs in each DNN. Additionally, we determined the percentile SI score for all co-binding TF pairs. Then, a two-sided t-test to compare the mean values for the percentile SI scores of key co-binding TF pairs and non-key co-binding TF pairs revealed a significant difference between the two groups ($p < 0.0001$) (Fig. 4b).

To emphasize the several important key co-binding TF pairs, we selected the top forty-eight interacting pairs for each key co-binding TF pair, such as SOX10, MYRF, OLIG1, OLIG2, NKX6.2, and TCF7L2, and generated a heatmap for their SI scores scaled from 0 to 1 (Fig. 4c). Similarly, we chose the top forty-eight interacting co-binding TF pairs for non-key TFs and created another heatmap for their SI scores scaled from 0 to 1 (Fig. 4d). We noticed that the SI scores for key-TF co-binding pairs have higher values than those for non-key co-binding TF pairs.

We also validated our model prediction performance for one TG, myelin basic protein (*MBP*), using additional data (Supplementary Fig. 2)[40]. We regressed the scaled actual values on the scaled predicted values. For our primary dataset, we obtained an R-squared of 0.81 and a r of 0.90 (Supplementary Fig. 2a). Furthermore, when analyzing another dataset, we observed an R-squared of 0.69 and a r of 0.83, affirming the predictive capability of our model architecture (Supplementary Fig. 2b).

https://doi.org/10.1038/s42003-025-07570-6

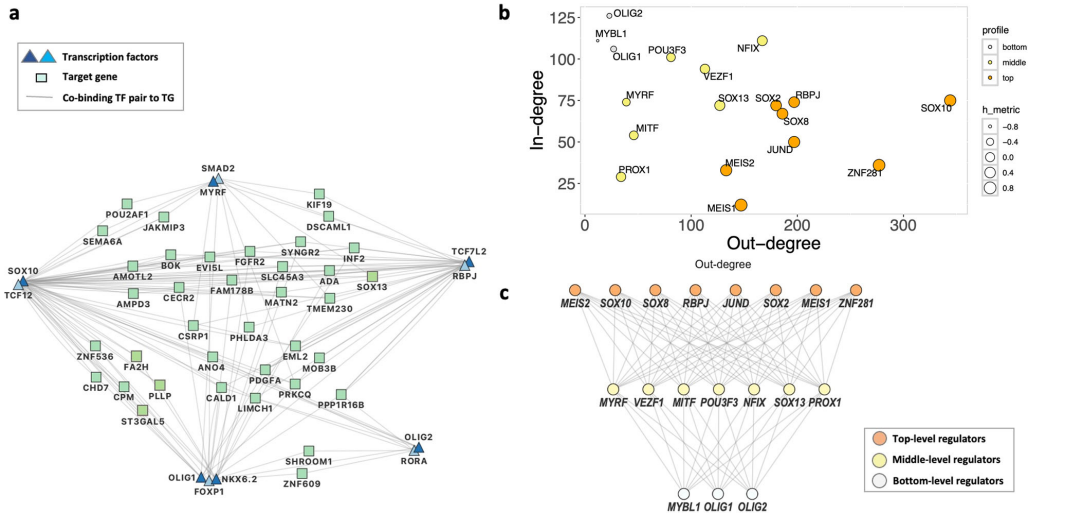


Fig. 5 | Gene regulatory network and transcription factor hierarchy. a Gene regulatory network for six key cooperative transcription factor pairs with the highest interaction scores, **b** A plot of in-degree (I) vs out-degree (O) for the transcription factors that have I and O in the gene regulatory network, and **b** Transcription factor hierarchy. Each node depicts a transcription factor. In (**b**, **c**), the edges colored in orange are the top-level (master) regulators, the edges colored in yellow are the middle-level regulators, and the edges colored in white are the bottom-level regulators.

Oligodendrocyte gene regulatory network analysis for cooperative TF pairs and transcription factor hierarchy

We chose one pair of co-binding TF with the highest interaction scores from six key co-binding TF pairs, including SOX10, MYRF, OLIG1, OLIG2, NKX6.2, and TCF7L2. We built a gene regulatory network (GRN) for these cooperative TF pairs and their TGs that are co-regulated by them (Fig. 5a). We found that a TG, *CALD1*, is co-regulated by three key cooperative TF pairs, SOX10-TCF12, RORA-OLIG2, and FOXP1-NKX6.2 and another TG, *PPP1R16B*, is co-regulated by three key cooperative TF pairs, RORA-OLIG2, FOXP1-NKX6.2, and FOXP1-OLIG1. There are other TGs, such as *AMOTL2*, *BOK, CALD1, FA2H*, and *CPM*, in the GRN that are co-regulated by two pairs of cooperative TFs.

We computed in-degree and out-degree for eighteen TFs that can also be TGs at the same time since TF feed forward and feedback loops are common (Fig. 5b). Then, we conducted a TF hierarchy analysis and found eight top-level regulators (Fig. 5c), called 'Master regulators', including SOX10, SOX2, and SOX8, which are key TFs that are known to play critical roles in oligodendrocyte differentiation[30–32]. The other five master regulators, MEIS1, MEIS2, RBPJ, JUND, and ZNF281, are categorized as oligodendrocyte-specific non-key TFs in Fig. 3c. All TFs that are middle-level regulators and bottom-level regulators, except for MYRF, are categorized as oligodendrocyte-specific non-key TFs. MYRF is one of the key TFs which is specifically activated in myelinating oligodendrocytes. PROX1 has been identified as being important for oligodendrocyte differentiation[41,42]. Most of these eighteen TFs are expressed in both oligodendrocytes and OPCs (Supplementary Fig. 4). It provides evidence that oligodendrocyte differentiation is pre-set in OPCs[43].

Independent validation for cooperative TFs

eQTL mapping. As an independent assessment of the regulatory regions we mapped oligodendrocyte eQTLs[44] onto oligodendrocyte-specific regulatory regions to explain the causal relationships between the expression levels of the co-binding TF pairs we identified and their target genes (TGs). Using chromosome and position of eQTL SNPs (eSNPs) from oligodendrocyte eQTLs, eSNPs integrated with a total of 643 oligodendrocyte-specific regulatory regions (Fig. 2a). This integration facilitates the identification of potential regulatory connections between the eSNPs and the co-binding TFs in these regions, enhancing our understanding of how genetic variations influence the expression levels of the identified co-binding TF pairs and their corresponding TGs. Notably, it provides evidence of causation if the eQTL genes and TGs are identical where co-binding TF pairs occur, indicating that these co-binding TF pairs are co-regulating TG expressions.

First, among 4.8 million oligodendrocytes eQTLs, we filtered 2 million significant (FDR < 0.05) eQTLs. Second, we mapped these significant eSNPs onto oligodendrocyte-specific regulatory regions (Fig. 6a). In total, 383 eSNPs and 159 eGenes were mapped onto 188 regulatory regions. Among these, 373 eSNPs and 153 eGenes (and TGs) were found in 179 regulatory regions associated with key TF pairs. Enrichment analysis for TGs indicates their strong involvement in biological processes such as oligodendrocyte development, myelination, and oligodendrocyte differentiation. (Fig. 6b)

Validation of cooperative TF pairs. The model generated from human epigenome and expression data predicted a number of enriched TF pairs within oligodendrocyte-specific TF regulatory elements. In order to test if the coordination occurs as predicted, we utilized rat oligodendrocyte ChIP-seq data that were available for selected transcription factors. One predicted pair was OLIG2/SOX10, which had previously been shown to be extensively colocalized in analyses of rat oligodendrocytes[45]. To visualize the preferential binding of SOX10 on a global scale, a read density plot for SOX10 ChIP-seq reads[11] was generated centered on the previously defined OLIG2 peaks[45] in oligodendrocytes (Fig. 6c). In line with previous analysis, the average read density of SOX10 is highly enriched over OLIG2 bound sites. A newly found pair predicted by the model was that of NKX2.2 and SOX10, and we generated a similar plot of SOX10 ChIP-seq reads over a defined set of NKX2.2 ChIP-seq peaks in oligodendrocytes[46], and we found a similarly high enrichment of SOX10 binding on ~40% of NKX2.2 binding sites (Fig. 6d). An example of the colocalization is shown for the *MBP* gene, which *MBP* is a crucial TG in oligodendrocytes as a key component of the myelin sheath[47,48]. Expression of *MBP* is essential for the differentiation and maturation of oligodendrocytes[49,50], and *MBP* maintains the structure and integrity of

https://doi.org/10.1038/s42003-025-07570-6

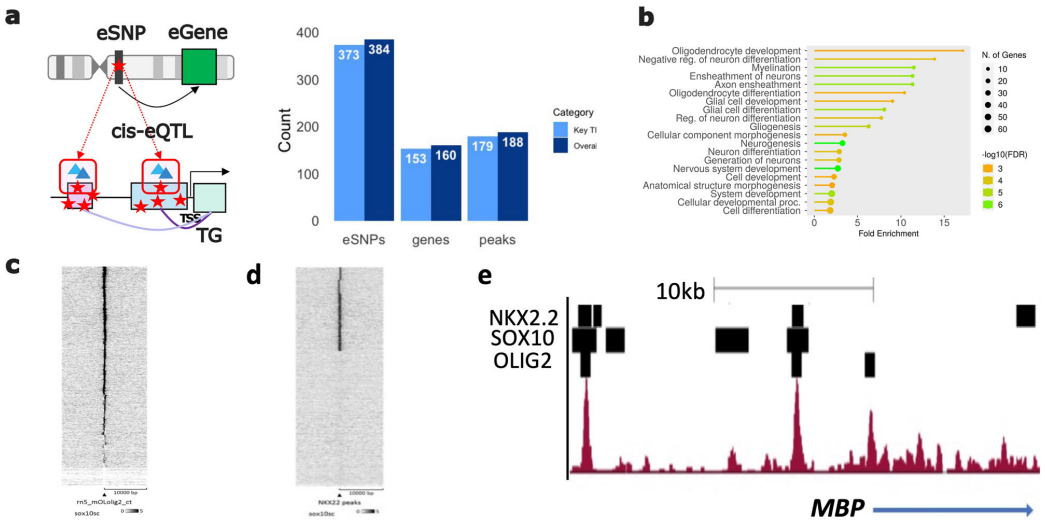


Fig. 6 | Independent validation using eQTLs and ChiP-seq data. a eQTL mapping onto oligodendrocytes regulatory regions, **b** ChIP-seq peak for cooperative TF pairs that co-regulate *MBP*. The ChIPseq profile is for SOX10, and solid blocks indicate called peaks for the specified transcription factors. **c** Gene ontology enrichment analysis of target genes associated with oligodendrocyte eQTL's, and **d**,**e** Heatmaps show distribution of SOX10 ChIP-seq reads centered on the previously defined OLIG2 and NKX2.2 sites.

the myelin sheath[51]. As shown in Fig. 6e, there are at least 2 sites upstream of *MBP* where there is colocalization of SOX10 with NKX2.2 and OLIG2.

Boolean cooperativity of TF pairs. We applied a logic circuit to characterize Boolean cooperativity of TFs using Loregic[52]. A total of 206 TFs that form 8101 co-binding TF pairs were input. 6660 (82.2%) out of the 8101 co-binding TF pairs have consistent triplets—matching the same logic gate across all targets, demonstrating strong cooperation between the activities of the two TFs on the TGs. More than half of the TF1-TF2-TG pairs are categorized as "AND" indicating a positive correlation between TG expression and the expression of both TF1 and TF2 (Fig. S6a). We also achieved permutation scores to remove logic gates chosen by random. Still, 6092 TF pairs have consistent triplets and 64% of triplets are categorized as "AND" (Fig. S6b).

Independent validation for the prediction performance of the models

Model prediction validation and ablation study. Using Multi-omics scRNA-seq data[21] from the same cells as the scATAC-seq data in the main analysis, we trained deep learning models and computed SI scores. Forty-eight SI scores for key-TF pairs in Fig. 4c were selected. The correlation between the SI scores computed from the main data and the Multi-omics data are shown in Fig. S5a. We also ran a two-sided t-test to compare the mean values for the percentile SI scores of key co-binding TF pairs and non-key co-binding TF pairs as we did for the main data (Fig. 4b). There was a significant difference between the two groups ($p < 0.0001$) (Fig. S5b).

Model performance was evaluated using the holdout data. Additionally, we included three more publicly available scRNA-seq datasets: Multi-omics, ROSMAP[40], and Cross-disorder[53], and validated the prediction performance of our model for each TG. Here, TGs were predicted using the trained models and the entire datasets. The holdout data, Multi-omics, and ROSMAP show consistently low normalized root mean squared error (NRMSE), while more than 75% of predictions in Cross-disorder also have low NRMSE. NRMSE can be compared across genes (Fig. S7a).

We also conducted an ablation study to compare the prediction performance of our models. Another dataset for 206 random TFs that are neither co-binding nor cooperative was created and their prediction performance was compared to that of 206 co-binding TFs (Fig. S7b). The model prediction performance is much better overall when 206 TFs used for predicting TGs are either co-binding, cooperative, or both.

Discussion

With resources provided by advances in single-cell sequencing, some studies[54–57] have elucidated the roles of several TFs, enabling the construction of cell type-specific gene regulatory networks to explain potential TF-TG relationships using bioinformatic tools. However, most of these studies and tools primarily focus on relationships between independent TFs and TGs.

This study introduces an analytical framework, coTF-reg, which identifies co-binding TFs and their TGs in oligodendrocyte-specific regulatory regions. Deep learning models predicted TG expression levels using the expression levels of co-binding TF pairs, and we computed TF SI scores to define highly interacting co-binding TF pairs as 'cooperative' TFs that co-regulate TG expression levels. We found that the key co-binding TF pairs tend to highly interact with each other compared to non-key co-binding TF pairs for predicting TG expression levels. Independent validation, such as mapping eQTLs onto the regulatory regions, provides evidence for causal relationships between co-binding TF pairs and TGs. Additionally, converting these regions to the rat genome assembly coordinates and measuring the density of ChIP-seq signals for key cooperative TFs show that many of these TF pairs are enriched in the regulatory regions, indicating their collaborative role in co-regulating TG expression levels. We defined specific key TFs and examined co-binding TF pairs containing them, along with their interactions in predicting TG expression levels. We then compared these results with those of non-key TF pairs. Overall, co-binding TF pairs with known regulators of oligodendrocyte development exhibit higher SI scores, suggesting that they not only regulate TG expression individually but also cooperatively. We identified several highly cooperative TF pairs, such as SOX10 and OLIG2[12,58], which are already known. Additionally, we

https://doi.org/10.1038/s42003-025-07570-6

discovered previously unreported cooperative pairs, such as SOX10 and NKX2.2.

Our study demonstrates several strengths. First, we concentrate on interactions between co-binding TF pairs and their impact on TG expression using deep learning approaches. Deep learning can elucidate complex TF relationships and their effects on TG expression levels. Second, the coTF-reg pipeline can be used by general users with any scATAC-seq and scRNA-seq data. The code for coTF-reg is openly available on GitHub, allowing users to input their scATAC-seq and scRNA-seq data for specific purposes. Third, we provide a comprehensive analytical framework that incorporates analyses utilizing co-bindings by motif and expression levels. We define 'cooperative' TF pairs as TF pairs significantly co-enriched across regulatory regions, exhibiting high SI scores in terms of expression when predicting TG expression. The term cooperativity has often been applied to co-bindings of TFs to nearby sites that facilitates stabilized binding due to protein-protein interactions, but in our model, we use TF pairs that can bind to sites in the same regulatory regions, since TF's can coordinately activate enhancers without direct interactions.

Nevertheless, there are some limitations to our study. To begin with, it's important to note that more than two TFs can co-regulate TG expression[59,60]. However, our current tool is limited to analyzing interactions between two co-binding TFs. In future research, developing or applying more sophisticated methods capable of handling clusters of TFs that co-regulate the same TG expression will be informative. Moreover, our method for identifying binding sites relies on the position frequency matrices in the motif database. While both SOX10 and MYRF are key TFs for oligodendrocytes, we encountered difficulty in obtaining sufficient binding sites for MYRF. Consequently, we had to supplement with a different motif for MYRF based on our prior knowledge. More generally, the definition of TF motifs relies on disparate methods, and limitations of motif generation and analysis have been noted previously. Nonetheless, our analysis provided TF-TF coordination that we could validate using data from previous studies. We predict that future analysis can be used to determine if the predicted TF pairing plays a role in oligodendrocyte differentiation, since reliance on single factor studies is not able to recapitulate the important combinatorial functions of TF's in generating cell type-specific gene expression patterns. Lastly, there can be alternative methods for establishing cooperative relationships between TFs, such as Boolean rules[61–65]. Logic-based models are also powerful tools for understanding the complex interactions among regulatory TFs in gene regulation. Developing new tools that incorporate Boolean rules and machine learning approaches will help us effectively infer more intricate TF relationships, paving the way for future research aimed at unraveling the complexities of gene regulation.

Methods

coTF-reg pipeline workflow

First, published scATAC-seq data with peak-to-gene links[21] is inputted into the coTF-reg pipeline. Second, transcription factor binding sites (TFBSs) and co-binding TF pairs in the oligodendrocyte regulatory regions are identified through motif co-occurrence and co-enrichment analyses. Third, deep neural networks (DNNs) to predict the expression levels of the TGs are trained and the interaction effects between co-binding TFs on the expression levels of TGs using gene expression from scRNA-seq data[22] are measured by computing Shapley interaction (SI)[23,24] scores. Fourth, a gene regulatory network is built based on SI scores for co-binding TF pairs. Fifth, a TF hierarchy analysis is used to define TFs as regulators in three categories. Lastly, as an independent validation, to validate the cooperative TF pairs: 1. The oligodendrocyte eQTLs are mapped onto the regulatory regions where cooperative TF pairs exist, 2. Liftover analysis and co-enrichment analysis using ChIP-seq data are conducted, 3. Boolean rules are applied to characterize the cooperativity of regulatory factors. To evaluate the prediction performance of our models: 1. Other publicly available datasets are used as validation data to predict TG expressions, 2. Ablation study is implemented by generating random TF sets to predict TG expressions.

Step 1: Infer transcription factor binding sites. We inferred transcription factor binding sites (TFBSs) in 787 scATAC-seq peak regions that have linkages with TGs.

a) The R package *GenomicRanges* was used to format the ATAC-seq peaks into genomic ranges.
b) Position frequency matrices (PFMs) for the 949 motifs in *JASPAR2022* database[66] were set in R, along with nine additional PFMs for the important modified motifs based on our prior knowledge.
c) TFBSs in the scATAC-seq peak regions were inferred using a R package, *motifmatchr*[67].

Step 2: Identify co-binding transcription factor pairs. We identified co-binding TF pairs using the inferred TFBSs in Step 1.

a) All possible TF-TF pairs with binding sites in the scATAC-seq peak regions were considered.
b) TF pairs from the same families were excluded.
c) Co-enrichment analysis: Co-occurrence analysis was conducted to find TF pairs that have overlapping regions. We then conducted hypergeometric tests to find significantly enriched TF pairs in the same regions. We used multiple testing corrections via FDR and applied FDR < 0.1 cutoff. We define the TF pairs that are co-enriched (FDR < 0.1) as 'co-binding' TF pairs.
d) Gene regulatory networks (GRNs) were constructed for TG-co-TF pair-peak links and matched TGs and co-TF pairs to the scRNA-seq data.
e) Lowly expressed TGs and TFs were removed from the GRNs by applying a cutoff, median expression level > 1; more than half of the cells are expressed, from the GRNs.
f) Differential expression testing was implemented using Seurat[68] and selected TGs that are oligodendrocyte specific in the GRNs.
g) Peaks were annotated as promoters or enhancers using *annotatr*[69].

Step 3: Measure cooperativity of co-binding transcription factors. Gene expression levels of the co-binding TF pairs from scRNA-seq data were incorporated into deep learning models to predict the expression levels of the TGs and measure interaction effects between co-binding TFs on the expression levels of TGs using Shapley interaction (SI) scores.

a) Metacells for the cells in scRNA-seq data were projected using a Python package, *metacells*[70].
b) Expression levels of TFs that have co-binding TFs and TGs were used to construct deep learning models for each TG using *PyTorch*[71] in Python.
c) SI scores for TF pairs were computed in each deep learning model.
d) Interaction matrices for the SI scores were generated in deep learning models and the mean interaction scores for co-binding TF pairs were calculated.
e) Coefficients of variation (CV)[72] of the interaction scores for each co-binding TF pair were computed and the pairs with CV values higher than 0.5 were removed.

Step 4: Gene regulatory network and TF hierarchy analysis. A gene regulatory network was built for six key cooperative TFs.

a) One cooperative TF pair for each of the six key TFs was selected based on the top interaction scores.
b) A gene regulatory network was bulit linking cooperative TF pairs to TGs.
c) TGs co-regulated by cooperative TF pairs were selected.
d) A network plot was generated using Cytoscape[73].

https://doi.org/10.1038/s42003-025-07570-6

Step 5: TF hierarchy analysis. TFs that can be TGs were chosen, and we implemented hierarchy analysis[74] for those TFs.

a) In-degree (I) and out-degree (O) for the TFs were calculated.
b) Hierarchy height metrics for the TFs were computed.
c) TFs were classified as top-regulator, middle-regulator, or bottom-regulator.

Step 6: Independent validation. We implemented eQTL mapping, ChIP-seq enrichment analysis, and Boolean cooperativity analysis for validating cooperative TF pairs and model prediction validation and an ablation study for validating the prediction performance of the models.

Validation of cooperative TF pairs

eQTL mapping. We mapped the significant (FDR < 0.05) oligodendrocyte eQTLs onto the scATAC-seq peak regions.

a) Publicly available oligodendrocyte eQTL data[44] were downloaded and the significant (FDR < 0.05) eQTLs were extracted.
b) The significant eQTLs were mapped to the scATAC-seq peak regions in the GRNs.
c) The results were verified by comparing the number of eQTLs mapped onto the peak regions for key-TF pairs and non-key-TF pairs.

ChIP-seq enrichment analysis

We performed the LiftOver analysis to convert genome coordinates for rat to human hg38 assembly using UCSC Genome Browser[75].

a) Genome coordinates for human hg38 assembly were converted to the rn5 rat genome coordinates for human (hg19) assembly.
b) Overlapping genome coordinates between conserved (from hg38 to rn5) assembly and the regulatory regions in the GRN were identified.
c) Cooperative TF pairs in the overlapping regions identified, along with the TGs they co-regulate.

Using the results from the LiftOver analysis, we tried to find signals in co-enriched binding sites for cooperative key TF pairs in rat oligodendrocyte ChIP-seq data. Heatmaps were created via EAseq[76]. ChIP-seq tracks were visualized using UCSC genome browser. Previous ChIP-seq datasets for SOX10, OLIG2, and NKX2.2 are available at GEO accession numbers: GSE64703, GSE42447 and GSM1906296.

Boolean cooperativity of TF pairs

We applied a logic circuit to characterize Boolean cooperativity of TFs using Loregic[52]. Loregic is a computational tool, integrating gene expression and regulatory network, to characterize the cooperativity of regulatory factors. It uses 16 possible two-input-one-output logic gates (e.g. AND) to describe triplets of two factors regulating a common target. The GRN was inputted including co-binding TFs-TG links. Then, we binarized the gene expression levels to Boolean values 1 and 0 to represent high and low gene expression, respectively, using BoolNet[77]. BoolNet assigned Boolean values to expression data on the basis of modular co-expression patterns by K-means clustering across inputted samples and therefore accounts for differences in the dynamic ranges of expression among genes in the input data. The triplet gene expression data was extracted and matched to all possible logic gates. We selected consistent logic gates. We also ran 100 permutation tests to find significant logic gates.

Validation of the prediction performance

Model prediction validation. To verify the performance of deep learning model architectures, we trained a deep learning model for predicting a TG, *MBP* using another data[40]. The trained model was used to predict the expression level of *MBP* and compared the results with the model for *MBP* using the main data.

Using Multi-omics scRNA-seq data[21] from the same cells as the scATAC-seq data in the main analysis, we trained deep learning models and computed SI scores, following the same processes we did in coTF-reg pipeline for identifying cooperative TFs in oligodendrocyte gene regulation ('Step 2 Measure cooperativity of co-binding TFs') for the main scRNA-seq data.

Model performance was evaluated using the SEA-AD[22] holdout data. We also include three more publicly available scRNA-seq datasets: Multi-omics[21], ROSMAP[40], and Cross-disorder[53], and validate the prediction performance of our model for each TG. Here, TGs were predicted using the trained models and the entire datasets. Normalized root mean squared error (NRMSE) is used to compare the performance across different datasets.

Ablation study

It is important to assess whether the 206 co-binding TFs effectively predict their TGs. Another dataset with 206 random TFs that are neither co-binding nor cooperative was generated to evaluate the prediction performance of our models. We used our trained models to predict holdout data for random TFs and compared their prediction performance to that of 206 co-binding TFs.

Single-cell ATAC-seq data

Chromatin accessibility data[21] was used for the main analyses. Brain samples were selected and eight thousand nuclei from each sample were subjected to the Chromium Next GEM Single-Cell Multiome ATAC-seq. We filtered oligodendrocyte-specific peak-gene links for our analyses. 930 peaks and 606 genes were initially chosen.

Single-cell RNA-seq data

SEA-AD (Main analysis). The data for the whole taxonomy collected from dorsolateral prefrontal cortex (1,395,601 cells) were downloaded through the Open Data Registry on AWS as AnnData objects (h5ad format)[22]. The cells for disease were excluded and only the controls were retained. Then, we projected metacells for the whole taxonomy and found 2004 metacells and 17,946 genes for oligodendrocytes.

Multi-omics. The normalized and quality controlled data was gained from the CELLxGENE (RRID:SCR_021059) portal. Brain samples were selected and eight thousand nuclei from each sample were subjected to the Gene Expression protocol (10x Genomics). We filtered 5459 cells for oligodendrocyte.

ROSMAP. The processed count matrix for oligodendrocyte was downloaded from a supplementary website for 'Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology'[40]. We projected metacells for the controls only and found 7072 metacells and 16,707 genes.

Cross-disorder. Post quality control filtered data was obtained from the CELLxGENE portal. We projected metacells for oligodendrocyte controls and found 1004 metacells and 21,248 genes for oligodendrocytes.

Uniform manifold approximation and projection for dimension reduction

We gained scRNA-seq data for the whole taxonomy collected from dorsolateral prefrontal cortex through the Open Data Registry on AWS as AnnData objects (h5ad format)[22]. There were 1,395,601 cells across 18 subcell types. A total of 18,431 hg38 protein-coding genes, obtained via BioMart[78], were selected from 36,517 genes. We normalized the data to a depth of 10,000 and log1 transformed it using Scanpy[79] in Python. Then, the highly variable genes (HVGs) were identified using dispersion-based methods[80] to normalize dispersion, obtained by scaling with the mean and standard deviation of the dispersions for genes falling into a given bin for mean expression of genes. The cutoffs for the mean dispersions for genes were a minimum of 0.0125 and a maximum of 3, and for the minimum dispersion was 0.5. We identified 3032 HVGs and scaled each gene to unit

https://doi.org/10.1038/s42003-025-07570-6

variance to clip values exceeding standard deviation of 10. To reduce the dimensionality of the data, we ran principal component analysi and used top 30 PCs to compute the neighborhood graph of the cells. Finally, we embedded the neighborhood graph with 20 neighbors in two dimensions using Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)[81].

Differential expression testing

We inputted metacells for all cell types to identify oligodendrocyte-specific genes using Seurat v4 in R. We used the Poisson likelihood ratio test in FindMarkers function assuming that gene expression follows the negative binomial distribution. Oligodendrocytes and oligodendrocyte precursor cells (OPCs), and astrocytes that are known as major cell types among glia in the CNS were grouped and the other fifteen cell types were compared. We used a cutoff, FDR (<0.05) to select differentially expressed genes in the oligodendrocyte group.

Position Frequency Matrices

Position frequency matrices (PFMs) for the 949 motifs in JASPAR2022 were used to infer TF binding sites. We added PFMs for MYRF, SP7, and OLIG2 that are one of the key TFs from another study[82], Mus musculus in JASPAR2022[66], and HOCOMOCO v12[83], respectively. We also included shorter motifs for other key TFs, such as SOX10, MYRF, ZNF24, NKX2.2, and SP7, considering their importance in oligodendrocytes (Supplementary Fig. 3).

Co-enrichment analysis

We used a hypergeometric test to assess whether a number of overlaps in the binding sites for two TFs follows a hypergeometric distribution. Specifically, given that a random variable X represents the possible outcomes of a hypergeometric process, the probability of getting k or more overlapping binding sites between two TFs inside a particular chosen set, as a hypergeometric random process, is

$$\Pr(X \geq k | n; N; m) = \sum_{x=k}^{\min(n,m)} \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}} \quad (1)$$

where N is the total number of transcription binding sites for all TFs, m is the number of binding sites for TF1, n is the number of binding sites for TF2, and x is the number of overlapping binding (co-occurrence) sites between TF1 and TF2. We applied an FDR adjusted p-value as a cutoff (<0.1) for all possible TF pairs and chose co-binding TF pairs.

Key transcription factors

We defined ten key TFs that are oligodendrocyte marker genes based on mouse loss-of-function studies that have shown that specific TF's are critical for oligodendrocyte differentiation. This includes SOX10[29], SOX2[30,31], SOX8[32], MYRF[33], OLIG1[34], OLIG2[35], TCF7L2[36,37], ZNF24[25], NKX2.2[38], and NKX6.2[39] were chosen as key TFs. Ten 'Oligodendrocyte-specific key TFs' are oligodendrocyte differentially expressed TFs and key TFs, eighty-three 'oligodendrocyte-specific non-key TFs' are oligodendrocyte differentially expressed TFs but not key TFs, and a hundred-thirteen 'non-oligodendrocyte-specific TFs' are neither oligodendrocyte differentially expressed TFs nor key TFs. Especially, ten 'Oligodendrocyte-specific key TFs' play crucial roles in the development and differentiation of oligodendrocytes. They regulate various stages of oligodendrocyte maturation and promote the expression of myelin genes; essentially, they are key players in the process of myelination within the CNS.

Deep learning models

We inputted expression levels of TFs that have co-binding pairs into the deep neural network (DNN) models to predict TG expression levels. 2004 metacells (samples), 206 TFs (features), and a TG expression level (label) were used in the DNN models. A DNN for each TG was built to predict a TG expression level. The mean squared error (MSE) between predicted TG expression and actual TG expression was used as the loss function in DNN models. We cross-validated the training dataset (80% of the input samples) with 5-fold cross-validation and validated the best trained model on the 20% of hold-out validation dataset for the best use of data and to achieve reliable model performance. We used an early stopping function with patience 10 and determined the number of epochs and we set the batch size to 32. Adam with a learning rate 0.001 was used for training the models. The structure of our neural network model can be written as

$$Z_i = f(W_i \cdot X + b_i) \quad (2)$$

where X denotes the input data and f represents the activation function, specifically the LeakyReLU function. TF expression levels serve as the input data, while Zi represents the output of the ith hidden layer. The final output of the model is the predicted TG expression level, and W_i and b_i are the weight matrix and bias vector for the i^{th} layer, respectively.

To evaluate the performance of our neural network model, we utilize the Mean Squared Error (MSE) loss function. The MSE quantifies the average squared difference between the predicted outputs of the model, Z and the true labels in our regression task. Mathematically, we can express the MSE as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Z_i - Y_{true_i})^2, \quad (3)$$

where Z_i represents the predicted output for the i^{th} sample, and Y_{true_i} denotes the true label corresponding to the i^{th} sample.

Shapley interaction scores

We denote the set of all TFs by F, a feature i∈F, and a feature set S⊆F. We define the interaction effect between TF i and j, with feature set S, of a neural network f at a data point X_k to be

$$\delta_{ij}^{f} = f(X_k; S\cup\{i,j\})) - f(X_k; S\cup\{i\})) - f(X_k; S\cup\{j\})) + f(X_k; S), \quad (4)$$

where $f(X_k; S)$ is the prediction at X_k when only TFs in S are used, which often requires retraining the NN multiple times. A common approximation is to replace the absent features (i.e., F\S) by the corresponding values in a baseline $C_{F\setminus S}$, such that

$$f(X_k; S) \approx f\left(X_{K,S}; C_{F\setminus S}\right) \quad (5)$$

The baseline is set as the empirical mean of each feature. The Shapley interaction score $SI_{ij}^{f}(X_k)$ is the expectation of $\delta_{ij}^{f}(X_k; S)$,

$$SI_{ij}^{f}(X_k) = E_{p(S)}\left[\delta_{ij}^{f}(X_k; S)\right], \quad (6)$$

over a uniformly random chosen feature set S from F. We use Monte-Carlo procedure[84] to approximate $SI_{ij}^{f}(X_k)$ by a small number of samples of S. To aggregate the local interaction effect at different data points into a global interaction effect, we take the expectation $\left|SI_{ij}^{f}(X_k)\right|$ of w.r.t. the empirical data distribution $p(X)$, such that

$$SI_{ij}^{f} = E_{p(X)}\left[\left|SI_{ij}^{f}(X)\right|\right] \quad (7)$$

For our deep ensemble of deep learning models, we utilize a posterior distribution of functions $q(f)$ induced by the ensemble distribution of the weights $q(w)$, as outlined in Eq. (2). This ensemble approach involves

https://doi.org/10.1038/s42003-025-07570-6 **Article**

training multiple instances of the model, each initialized with different random weights to promote diverse learning paths.

The weights w are drawn from a Gaussian prior, reflecting our initial uncertainty about their values. After training, we apply Bayesian inference techniques to update our beliefs about these weights and compute the posterior distribution $q(w)$. This posterior captures the uncertainty in the model parameters, providing a more comprehensive understanding of the model's behavior.

The function $q(f)$ represents the expected output of the model across this ensemble of weights. To compute the interaction score, we take the expectation of the interaction score SI_{ij} with respect to $q(f)$. This is estimated by averaging N_f samples drawn from the ensemble:

$$SI_{ij} = E_{q(f)}\left[SI_{ij}^{f}\right] \approx \frac{1}{N_f}\sum_{k=1}^{N_f} SI_{ij}^{f_k}. \quad (8)$$

We compute Shapley interaction scores[23,24] for the co-binding TF pairs, TF i and TF j using the trained DNN models and validation datasets. We calculate mean values for co-binding TF pairs using interaction matrices. We rank them by percentile and scaled them to 0 and 1 for easier interpretation.

Coefficient of variance

The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. The CV represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another. The CV is defined as the ratio of standard deviation to the mean as follows:

$$CV = \frac{\sigma}{\mu} \quad (9)$$

Hierarchy analysis

We computed connectivity statistics, out-degree (O) and in-degree (I), for individual TFs to get a 'hierarchy height' metric (h), a normalized value of the difference between O and I for each TF. The h is calculated as

$$h = \frac{O - I}{O + I} \quad (10)$$

We defined TFs as top-regulator ($h > 0.33$), middle-regulator ($-0.33 < h < 0.33$), and bottom-regulator ($h < -0.33$) by their h values.

Statistics and reproducibility

Data manipulation and analyses were performed using Python 3.10.14 and R 4.3.1. All relevant information including the sample sizes in the groups for statistical tests are included in the figure legends. The plots in this study are generated by Scanpy[79] (v1.10.3), and seaborn (v0.13.2) in Python and ggplot2 (v3.5.1) in R.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data supporting the results are included in Supplementary Data 1~5 and are publicly available on GitHub (https://github.com/daifengwanglab/coTF-reg). For the main analyses, scATAC-seq data were obtained from Supplementary Materials of the Multi-omics study[21] and scRNA-seq data was sourced from SEA-AD: Seattle Alzheimer's Disease Brain Cell Atlas (https://cellxgene.cziscience.com/collections/1ca90a2d-2943-483d-b678-b809bf464c30). For the independent validation, scRNA-seq data were acquired from the following websites: (1) Multiome (https://cellxgene.cziscience.com/collections/ceb895f4-ff9f-403a-b7c3-187a9657ac2c); (2) ROSMAP (https://compbio.mit.edu/ad_aging_brain/#loading-the-raw-data); (3) Cross-disorder (https://cellxgene.cziscience.com/collections/c53573b2-eff4-4c5e-9ad0-b24d422dfd9b).

Code availability

The code for the analyses and figures is available at https://github.com/daifengwanglab/coTF-reg.

Received: 19 June 2024; Accepted: 17 January 2025;
Published online: 05 February 2025

References

1. Bercury, K. K. & Macklin, W. B. Dynamics and mechanisms of CNS myelination. *Dev. Cell* **32**, 447–458 (2015).
2. Nasrabady, S. E., Rizvi, B., Goldman, J. E. & Brickman, A. M. White matter changes in Alzheimer's disease: a focus on myelin and oligodendrocytes. *Acta Neuropathol. Commun.* **6**, 22 (2018).
3. Singh, D. K., Ling, E. & Kaur, C. Hypoxia and myelination deficits in the developing brain. *Int. J. Dev. Neurosci.* **70**, 3–11 (2018).
4. Emery, B. & Lu, Q. R. Transcriptional and epigenetic regulation of oligodendrocyte development and myelination in the central nervous system. *Cold Spring Harb. Perspect. Biol.* **7**, a020461 (2015).
5. Valdés-Tovar, M. et al. Insights into myelin dysfunction in schizophrenia and bipolar disorder. *WJP* **12**, 264–285 (2022).
6. Maitre, M. et al. Myelin in Alzheimer's disease: culprit or bystander? *Acta Neuropathol. Commun.* **11**, 56 (2023).
7. Quan, L., Uyeda, A. & Muramatsu, R. Central nervous system regeneration: the roles of glial cells in the potential molecular mechanism underlying remyelination. *Inflamm. Regen.* **42**, 7 (2022).
8. Simons, M. & Nave, K.-A. Oligodendrocytes: myelination and axonal support. *Cold Spring Harb. Perspect. Biol.* **8**, a020479 (2016).
9. Elbaz, B. & Popko, B. Molecular control of oligodendrocyte development. *Trends Neurosci.* **42**, 263–277 (2019).
10. Ibarra, I. L. et al. Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. *Nat. Commun.* **11**, 124 (2020).
11. Lopez-Anido, C. et al. Differential Sox10 genomic occupancy in myelinating glia. *Glia* **63**, 1897–1914 (2015).
12. Sock, E. & Wegner, M. Using the lineage determinants Olig2 and Sox10 to explore transcriptional regulation of oligodendrocyte development. *Dev. Neurobiol.* **81**, 892–901 (2021).
13. Bujalka, H. et al. MYRF is a membrane-associated transcription factor that autoproteolytically cleaves to directly activate myelin genes. *PLoS Biol.* **11**, e1001625 (2013).
14. Weider, M. et al. Nfat/calcineurin signaling promotes oligodendrocyte differentiation and myelination by transcription factor network tuning. *Nat. Commun.* **9**, 899 (2018).
15. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* **16**, 144–154 (2015).
16. Emani, P. S. et al. Single-cell genomics and regulatory networks for 388 human brains. *Science* **384**, eadi5199 (2024).
17. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet* **45**, 580–585 (2013).
18. Bravo González-Blas, C. et al. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. Methods* **20**, 1355–1367 (2023).
19. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
20. Jin, T. et al. scGRNom: a computational pipeline of integrative multi-omics analyses for predicting cell-type disease genes and regulatory networks. *Genome Med.* **13**, 95 (2021).

21. Zhu, K. et al. Multi-omic profiling of the developing human cerebral cortex at the single-cell level. *Sci. Adv.* **9**, eadg3754 (2023).
22. Gabitto, M. I., Travaglini, K. J., Rachleff, V. M. et al. Integrated multimodal cell atlas of Alzheimer's disease. *Nat. Neurosci.* **27**, 2366–2383 (2024).
23. Cui, T. et al. Gene–gene interaction detection with deep learning. *Commun. Biol.* **5**, 1238 (2022).
24. Dhamdhere, K., Agarwal, A. & Sundararajan, M. The Shapley Taylor Interaction Index. https://doi.org/10.48550/ARXIV.1902.05622. (2019).
25. Elbaz, B. et al. Phosphorylation sState of ZFP24 controls oligodendrocyte differentiation. *Cell Rep.* **23**, 2254–2263 (2018).
26. Bernhardt, C. et al. KLF9 and KLF13 transcription factors boost myelin gene expression in oligodendrocytes as partners of SOX10 and MYRF. *Nucleic Acids Res.* **50**, 11509–11528 (2022).
27. Howng, S. Y. B. et al. ZFP191 is required by oligodendrocytes for CNS myelination. *Genes Dev.* **24**, 301–311 (2010).
28. Al-Naama, N., Mackeh, R. & Kino, T. C2H2-Type Zinc finger proteins in brain development, neurodevelopmental, and other neuropsychiatric disorders: systematic literature-based analysis. *Front. Neurol.* **11**, 32 (2020).
29. Aprato, J. et al. Myrf guides target gene selection of transcription factor Sox10 during oligodendroglial development. *Nucleic Acids Res.* **48**, 1254–1270 (2020).
30. Zhao, C. et al. Sox2 sustains recruitment of oligodendrocyte progenitor cells following CNS demyelination and primes them for differentiation during remyelination. *J. Neurosci.* **35**, 11482–11499 (2015).
31. Zhang, S. et al. Sox2 is essential for oligodendroglial proliferation and differentiation during postnatal brain myelination and CNS remyelination. *J. Neurosci.* **38**, 1802–1820 (2018).
32. Turnescu, T. et al. Sox8 and Sox10 jointly maintain myelin gene expression in oligodendrocytes. *Glia* **66**, 279–294 (2018).
33. Garnai, S. J. et al. Variants in myelin regulatory factor (MYRF) cause autosomal dominant and syndromic nanophthalmos in humans and retinal degeneration in mice. *PLoS Genet.* **15**, e1008130 (2019).
34. Chen, Y. et al. The oligodendrocyte-specific G protein–coupled receptor GPR17 is a cell-intrinsic timer of myelination. *Nat. Neurosci.* **12**, 1398–1406 (2009).
35. Wang, J. et al. Olig2 ablation in immature oligodendrocytes does not enhance CNS myelination and remyelination. *J. Neurosci.* **42**, 8542–8555 (2022).
36. Zhao, C. et al. Dual regulatory switch through interactions of Tcf7l2/Tcf4 with stage-specific partners propels oligodendroglial maturation. *Nat. Commun.* **7**, 10883 (2016).
37. Ye, F. et al. HDAC1 and HDAC2 regulate oligodendrocyte differentiation by disrupting the beta-catenin-TCF interaction. *Nat. Neurosci.* **12**, 829–838 (2009).
38. Qi, Y. et al. Control of oligodendrocyte differentiation by the Nkx2.2 homeodomain transcription factor. *Development* **128**, 2723–2733 (2001).
39. Southwood, C. et al. CNS myelin paranodes require Nkx6-2 homeoprotein transcriptional activity for normal structure. *J. Neurosci.* **24**, 11215–11225 (2004).
40. Mathys, H. et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. *Cell* **186**, 4365–4385.e27 (2023).
41. Bunk, E. C. et al. Prox1 is required for oligodendrocyte cell identity in adult neural stem cells of the subventricular zone. *Stem Cells* **34**, 2115–2129 (2016).
42. Kato, K. et al. Prox1 inhibits proliferation and is required for differentiation of the oligodendrocyte cell lineage in the mouse. *PLoS ONE***10**, e0145334 (2015).
43. Suzuki, N. et al. Differentiation of oligodendrocyte precursor cells from Sox10-venus mice to oligodendrocytes and astrocytes. *Sci. Rep.* **7**, 14133 (2017).
44. Bryois, J. et al. Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat. Neurosci.* **25**, 1104–1112 (2022).
45. Yu, Y. et al. Olig2 targets chromatin remodelers to enhancers to initiate oligodendrocyte differentiation. *Cell* **152**, 248–261 (2013).
46. Aguado, L. C. et al. microRNA function is limited to cytokine control in the acute response to virus infection. *Cell Host Microbe* **18**, 714–722 (2015).
47. Galiano, M. R. et al. Myelin basic protein functions as a microtubule stabilizing protein in differentiated oligodendrocytes. *J. Neurosci. Res.* **84**, 534–541 (2006).
48. Aber, E. R. et al. Oligodendroglial macroautophagy is essential for myelin sheath turnover to prevent neurodegeneration and death. *Cell Rep.* **41**, 111480 (2022).
49. Ehrlich, M. et al. Rapid and efficient generation of oligodendrocytes from human induced pluripotent stem cells using transcription factors. *Proc. Natl Acad. Sci. USA* **114**, E2243–E2252 (2017).
50. Smirnova, E. V. et al. Comprehensive Atlas of the myelin basic protein interaction landscape. *Biomolecules* **11**, 1628 (2021).
51. Snaidero, N. et al. Antagonistic Functions of MBP and CNP establish cytosolic channels in CNS Myelin. *Cell Rep.* **18**, 314–323 (2017).
52. Wang, D. et al. Loregic: a method to characterize the cooperative logic of regulatory factors. *PLoS Comput. Biol.* **11**, e1004132 (2015).
53. Rexach, J. E. et al. Cross-disorder and disease-specific pathways in dementia revealed by single-cell genomics. *Cell* **187**, 5753–5774.e28 (2024).
54. Yashar, W. M. et al. Predicting transcription factor activity using prior biological information. *iScience* **27**, 109124 (2024).
55. Duren, Z. et al. Sc-compReg enables the comparison of gene regulatory networks between conditions using single-cell data. *Nat. Commun.* **12**, 4763 (2021).
56. Ferrari, C., Manosalva Pérez, N. & Vandepoele, K. MINI-EX: integrative inference of single-cell gene regulatory networks in plants. *Mol. Plant* **15**, 1807–1824 (2022).
57. Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W. H. Modeling gene regulation from paired expression and chromatin accessibility data. *Proc. Natl Acad. Sci. USA* **114**, E4914–E4923 (2017).
58. Liu, Z. et al. Induction of oligodendrocyte differentiation by Olig2 and Sox10: Evidence for reciprocal interactions and dosage-dependent mechanisms. *Dev. Biol.* **302**, 683–693 (2007).
59. Kim, J. et al. The co-regulation mechanism of transcription factors in the human gene regulatory network. *Nucleic Acids Res.* **40**, 8849–8861 (2012).
60. Fuda, N. J., Ardehali, M. B. & Lis, J. T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461**, 186–192 (2009).
61. Buchler, N. E., Gerland, U. & Hwa, T. On schemes of combinatorial transcription logic. *Proc. Natl Acad. Sci. USA***100**, 5136–5141 (2003).
62. Silva-Rocha, R. & De Lorenzo, V. Mining logic gates in prokaryotic transcriptional regulation networks. *FEBS Lett.* **582**, 1237–1244 (2008).
63. Krumsiek, J., Marr, C., Schroeder, T. & Theis, F. J. Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. *PLoS ONE* **6**, e22649 (2011).
64. Malekpour, S. A., Shahdoust, M., Aghdam, R. & Sadeghi, M. wpLogicNet: logic gate and structure inference in gene regulatory networks. *Bioinformatics* **39**, btad072 (2023).
65. Malekpour, S. A., Haghverdi, L. & Sadeghi, M. Single-cell multi-omics analysis identifies context-specific gene regulatory gates and mechanisms. *Brief. Bioinform.* **25**, bbae180 (2024).
66. Castro-Mondragon, J. A. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 (2022).
67. Alicia, S. motifmatchr: Fast Motif Matching in R. R package version 1.28.0. https://doi.org/10.18129/B9.bioc.motifmatchr (2024).

68. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
69. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).
70. Ben-Kiki, O., Bercovich, A., Lifshitz, A. & Tanay, A. Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis. *Genome Biol.* **23**, 100 (2022).
71. Paszke, A. et al. PyTorch: An imperative style, high-performance deep learning library. https://doi.org/10.48550/ARXIV.1912.01703. (2019).
72. Koopmans, L. H., Owen, D. B. & Rosenblatt, J. I. Confidence intervals for the coefficient of variation for the normal and log normal distributions. *Biometrika* **51**, 25–32 (1964).
73. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
74. Gerstein, M. B. et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
75. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
76. Lerdrup, M., Johansen, J. V., Agrawal-Singh, S. & Hansen, K. An interactive environment for agile analysis and visualization of ChIP-sequencing data. *Nat. Struct. Mol. Biol.* **23**, 349–357 (2016).
77. Müssel, C., Hopfensitz, M. & Kestler, H. A. BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics* **26**, 1378–1380 (2010).
78. Smedley, D. et al. BioMart – biological queries made easy. *BMC Genom.***10**, 22 (2009).
79. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
80. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
81. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://doi.org/10.48550/ARXIV.1802.03426. (2018).
82. Kim, D. et al. Homo-trimerization is essential for the transcription factor function of Myrf for oligodendrocyte differentiation. *Nucleic Acids Res.* **45**, 5112–5125 (2017).
83. Vorontsov, I. E. et al. HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res.* **52**, D154–D163 (2024).
84. Štrumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **41**, 647–665 (2014).

Acknowledgements

This work was supported by National Institutes of Health grants, R21 NS128761, RF1MH128695, R01AG067025, R21NS128761, and National Science Foundation Career Award 2144475, and a core grant to the Waisman Center from NICHD (P50 HD105353).

Author contributions

Conceptualization, J.S. and D.W.; Methodology, J.C., J.S., and D.W.; Formal Analysis, J.C.; Investigation, J.C., J.S., and D.W.; Writing – Original Draft, J.C.; Writing – Review & Editing; J.C., J.S., and D.W.; Supervision, J.S., and D.W.; Funding Acquisition, J.S. and D.W.

Competing interests

The authors declare no competing interests.

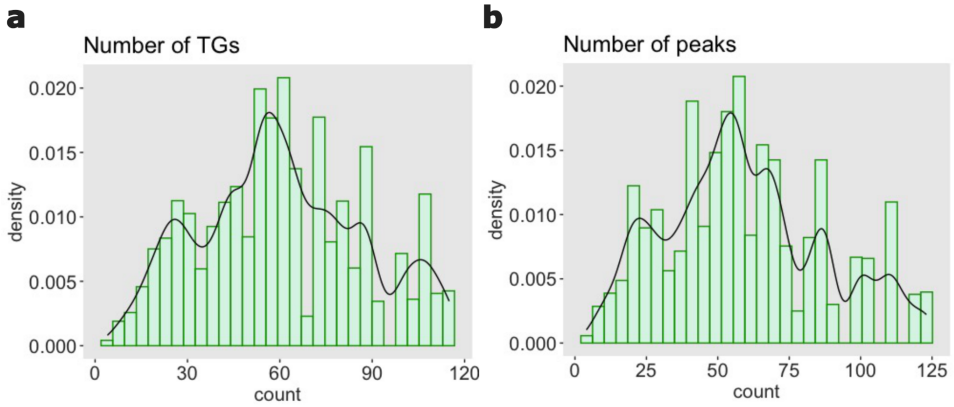
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-07570-6.

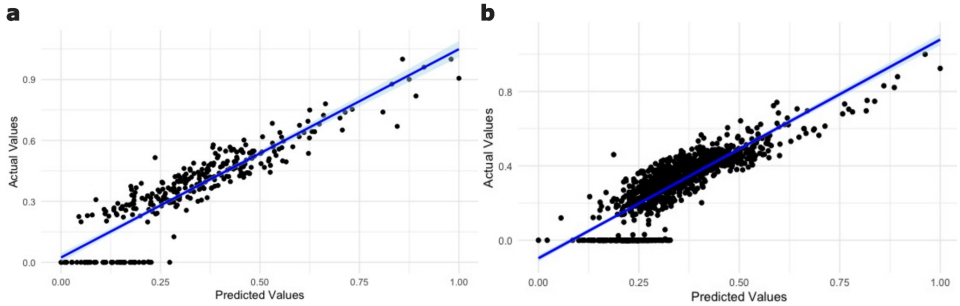
Correspondence and requests for materials should be addressed to Daifeng Wang.

Peer review information *Communications Biology* thanks Seyed Malekpour and Ping-Han Hsieh for their contribution to the peer review of this work. Primary Handling Editors:Chien-Yu Chen and Mengtan Xing.

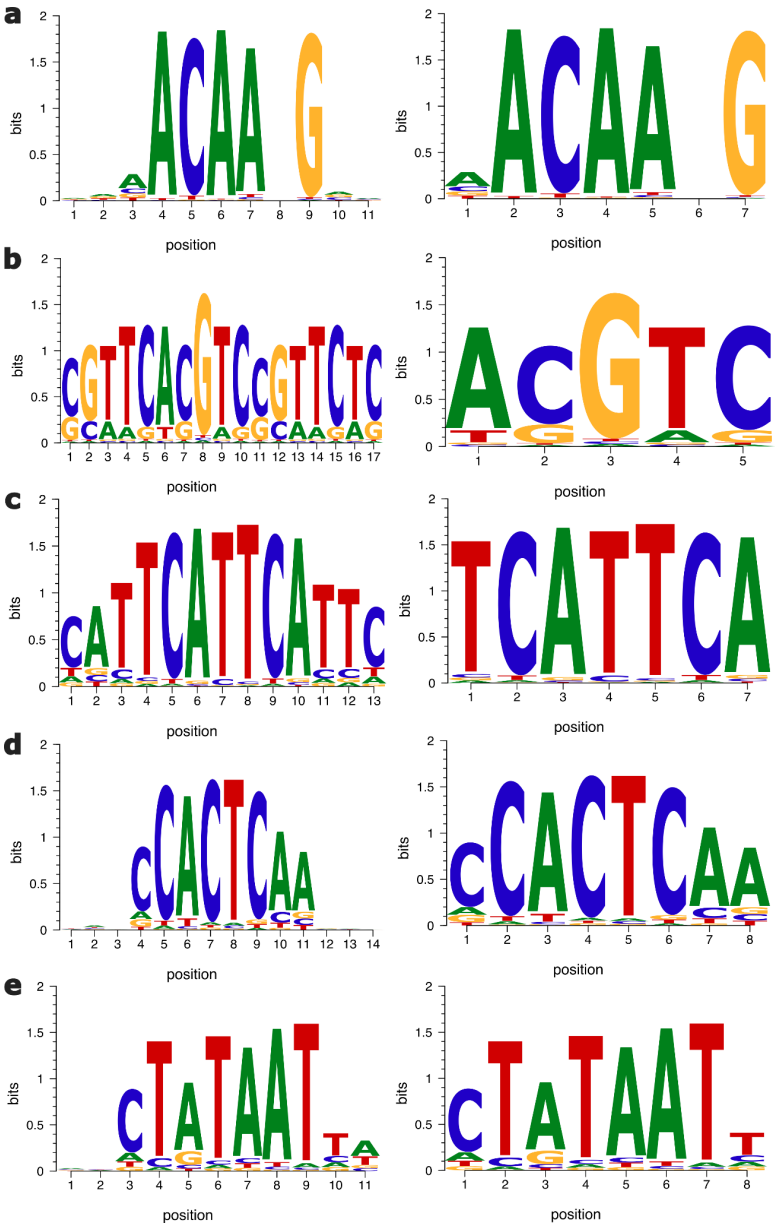
Supplementary information



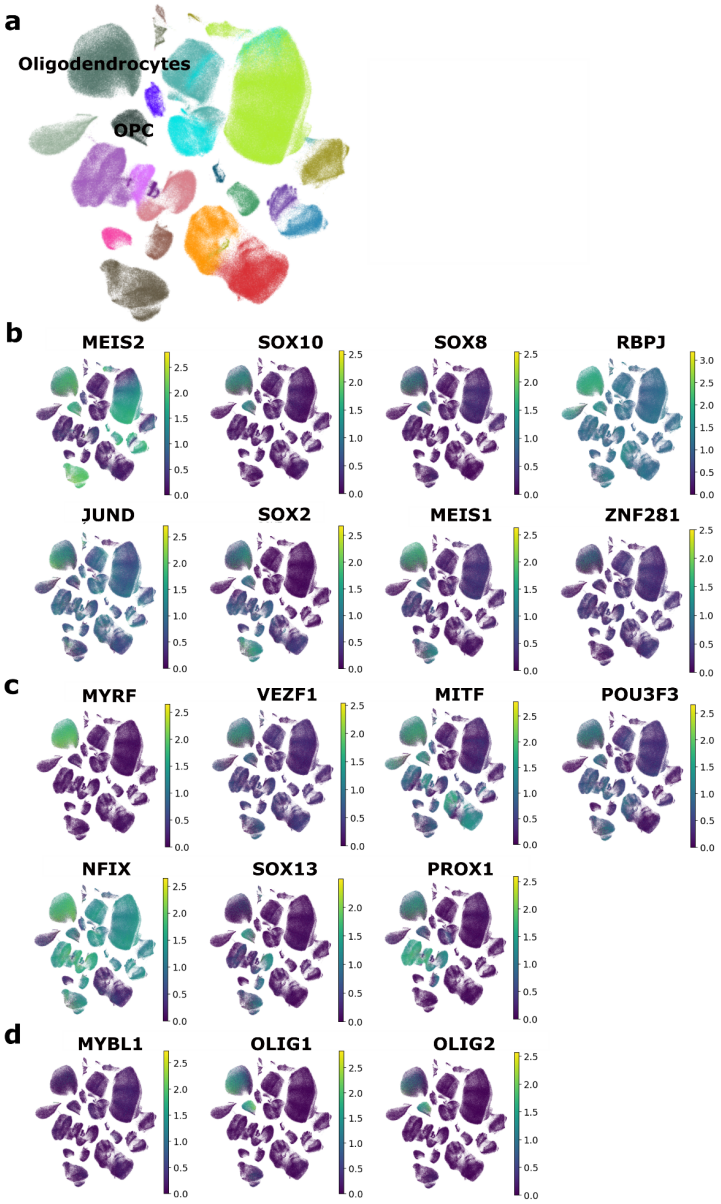
Supplementary Figure 1: Distributions of the numbers of target genes and peaks for co-binding TF pairs. a Distribution of the number of target genes (TGs) for co-binding transcription factor pairs. **b** Distribution of the number of peaks for co-binding transcription factor pairs.



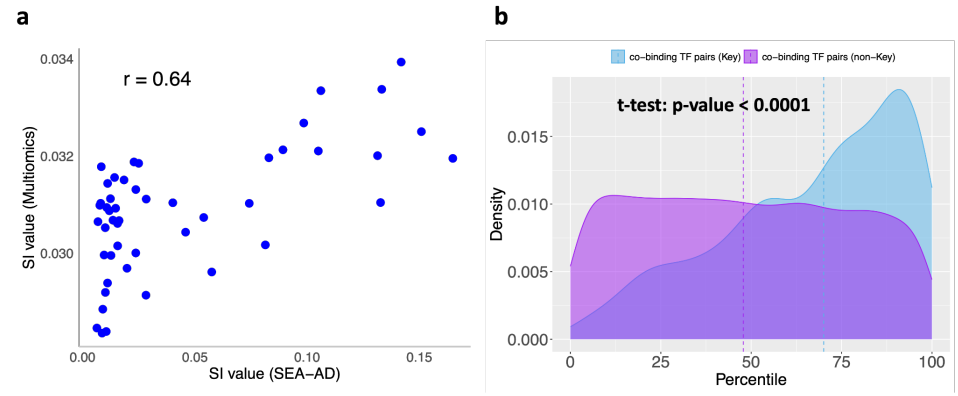
Supplementary Figure 2: Validation for the deep learning prediction. a Comparison between predicted values and actual labels in a deep learning model for predicting *MBP* using co-binding transcription factors in *Integrated Multimodal Cell Atlas of Alzheimer's Disease* data. **b** Comparison *between* predicted values and actual labels in a deep learning model for predicting *MBP* using co-binding transcription factors in *Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology* data.



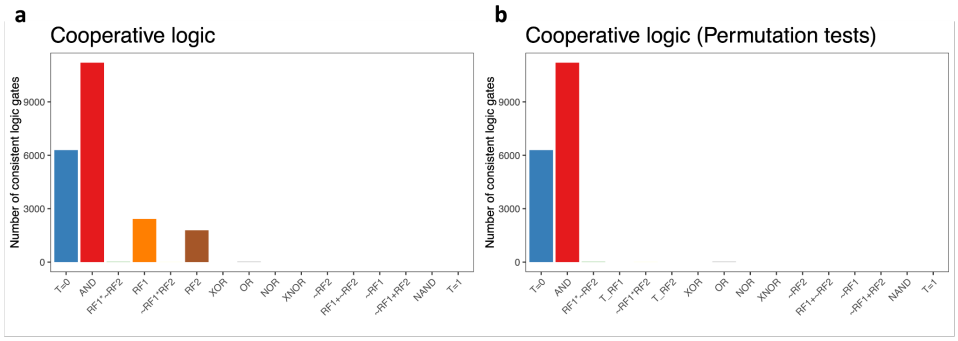
Supplementary Figure 3: Logo plots for the original and shorter motifs (The first column shows the long motifs and the second column shows the short motifs) a SOX10. **b** MYRF. **c** ZNF24. **d** NKX2.2. e SP7.



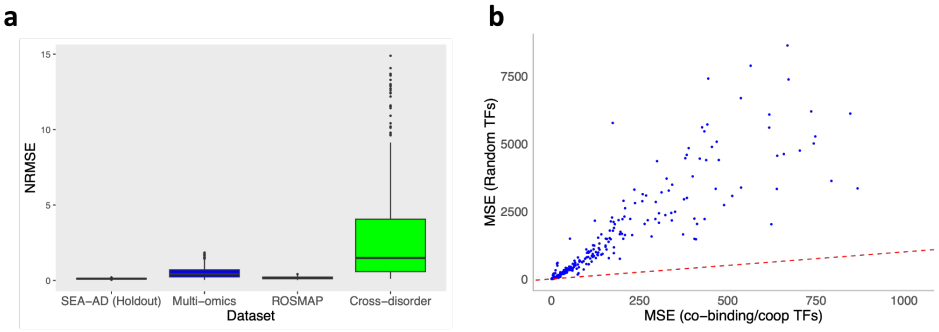
Supplementary Figure 4: UMAPs for all cell types and regulators. a UMAP for all cell types **b** top-level regulators ('Master regulators'). **c** middle-level regulators. **d** bottom-level regulators.



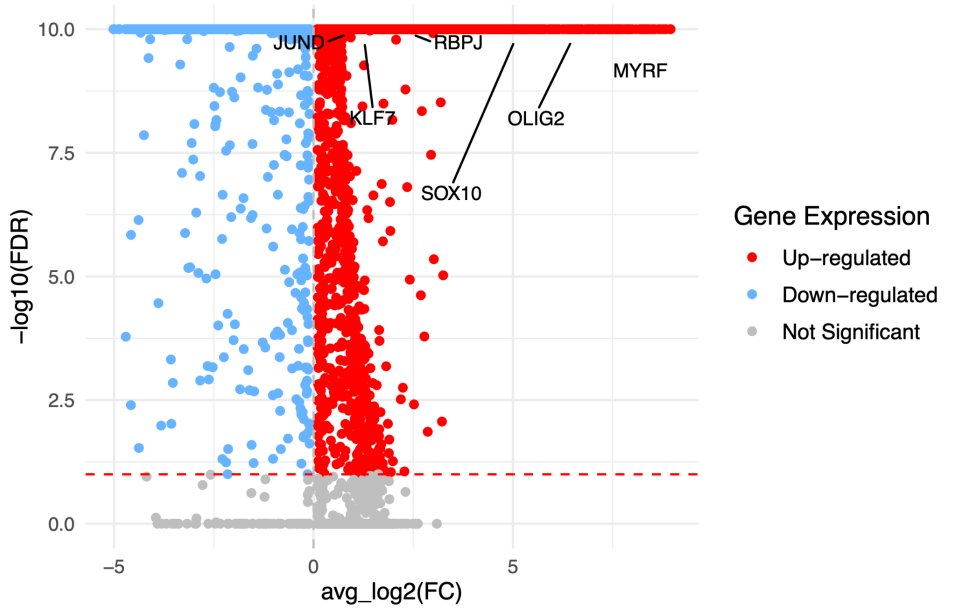
Supplementary Figure 5: TF interaction based on Multi-omics RNA-seq data. a Comparison for Top SI values for key-TF computed by two different scRNA-seq data, **b** T-test for the interaction score distributions of key-transcription factor pairs and non-key transcription factor pairs.



Supplementary Figure 6: Application of Boolean logic to find cooperative TFs. a Distribution of logic gates of gate-consistent triples. **b** Distribution of logic gates of gate-consistent triples after removing spurious logic gates by permutation tests.



Supplementary Figure 7: Holdout evaluation performance and model's prediction performance. a The x-axis represents four groups and the y-axis show normalized root mean squared error (NRMSE) **b** Mean squared error comparison for 206 co-binding/cooperative TFs and 206 random TFs that are neither co-binding nor cooperative.



Supplementary Figure 8: Differential gene expression testing for the 17,946 genes in SEA-AD data. The X-axis represents average log2 fold change values and the y-axis shows -log10 FDR values. Above the red dashed line, which is a cutoff for the FDR, the most up-regulated and significant genes are towards the

right (red), the most down-regulated and significant genes are towards the left (blue), and the statistically non-significant genes are under the red dashed line (grey).

Supplementary Data 1
Supplementary Data 2
Supplementary Data 3
Supplementary Data 4
Supplementary Data 5