Leveraging machine learning & interpretability methods for limited data: from systems biology to healthcare applications

Nicasia Beebe-Wang

A dissertation

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee

Su-In Lee, Chair

Sara Mostafavi

Christopher Althoff

Program Authorized to Offer Degree:

Paul G. Allen School of Computer Science and Engineering

University of Washington

Abstract

Leveraging machine learning & interpretability methods for limited data: from systems biology to healthcare applications

Nicasia Beebe-Wang

Chair of the Supervisory Committee:
Su-In Lee
Paul G. Allen School of Computer Science and Engineering

Although advances in machine learning (ML) have led to human-level performance in many domains, including some healthcare settings, high-performing models often rely on access to large datasets with densely observed features and labels. However, in practice, many biomedical applications involve some aspect of scarcity or sparsity that limit our ability to use standard ML models trained on big data. In particular, we focus on common limitations in computational biology and health settings including limited sample size, sparsely labeled data, and missing features or limited capacity to collect features at prediction time. We present projects, ranging in applications from computational molecular biology to healthcare applications, and for each, we describe our strategies for leveraging machine learning and explainability methods to enable prediction and interpretation of complex biomedical systems in the face of real-world data limitations.

In this work, we first describe a unified framework that we developed to uncover relationships between gene expression and Alzheimer's disease neuropathologies despite having a limited sample size and sparsely available labels by. By using multi-task deep learning in a multi-cohort setting and applying interpretability methods, we were able to identify nuanced sex-specific relationships among genes and AD. Next, we describe an automatic integrative method for learning interpretable communities of biological

pathways, which we developed in an effort to aid researchers in interpreting outcomes of computational biology analyses pipelines.

Finally, we turn to clinical risk prediction applications, in which medical practitioners often have limited time to collect features and assess a patient's risk for various health-related outcomes. We first describe a project in which we used feature attributions from a dementia prediction model to identify a globally relevant subset of features. We then demonstrated that a risk prediction model trained on these features achieved similar performance compared with a standard neuropsychological battery, but in in one-fifth of the time. However, this standard approach of retraining a model on a predetermined set of selected features may not be ideal in cases when different features are already available at prediction time, or if the individual is missing these key features. Thus, we finally propose a method for clinical risk prediction which simultaneously generates risk prediction intervals given sparse existing information and dynamically suggests useful features to collect next given a patient's current context.

Contents

Acknowledgments

First and foremost, I would like to thank Su-In Lee for her mentorship, guidance, inspiration, and encouragement over the last 6 years. Her endless support has helped me grow tremendously throughout my PhD, and was particularly impactful when I most doubted myself. I am also grateful to my wonderful committee members: Sara Mostafavi, for her amazing encouragement and mentorship over the last several years; Tim Althoff, for the wonderful guidance and conversations; and Sasha Aravkin for going above-and-beyond as my GSR.

I'm incredibly thankful for mentorship from other brilliant researchers both during and prior to my PhD. In particular, Safiye Celik has personally taught me so much and shared wisdom through many conversations continuing long past her time at UW. I've been fortunate to learn from many mentors throughout internships in the last few years, including Guang Li, Jon Irish, Sayna Ebrahimi, Sercan Arik, and Jinsung Yoon. I also want to thank Peter Koo for introducing me to the world of computational biology and for encouraging me through the grad school application process, and Scott Moeller for taking me under his wing (taking the time to personally teach statistics to a high schooler), and for opening a door for me to pursue a career in science.

I'm also grateful to have collaborated with other wonderful PhD students, including Alex Okeson, Ayse Dincer, Ethan Weinberger, Pascal Sturmfels, and Wei Qiu. It's been a highlight of my PhD to get to work with people that are both brilliant researchers and great friends. The Lee Lab and my UW CSE cohort-mates have been a great source of inspiration and camaraderie. It has been a joy to celebrate labmates' and classmates' accomplishments – and support each other through difficult times – and I'm grateful for the wonderful discussions/advice/commiseration sessions throughout the years. I'm also grateful to the Allen School Staff, particularly Elise and Joe, for their advice and assistance in navigating UW.

Finally, I am eternally grateful to my family and friends, without whom this would not have been possible. I would like to thank my parents for their endless love and support; Grammy, who got her PhD under much more challenging circumstances, but has always been unconditionally proud of me; Aunt Yun for welcoming me to her home in California; and the rest of my family and friends for their love and encouragement. I'm grateful beyond words to Geoff for being a better partner than I could ever have imagined. Lastly, have been blessed to have made wonderful friends throughout my time at UW, including Erin, Matt, Alex, Brian, Naveena, Ayse, Ian, Gabe, Pascal, my housemates, and many others. They have turned Seattle into a home, and brought so much joy, comfort, and laughter into my life over last six years.

Chapter 1. Introduction

Machine learning (ML) and technological advancements in the last several decades have resulted in large strides in our ability to design algorithms that rival human performance[1,2]. Often, these state of the art approaches involve training highly complex machine learning models using massive amounts of data (e.g., the BERT pre-trained language model[3]). In the biological and medical domains, we have also seen examples of highly successful models trained on large-scale datasets. For example, computer vision models have been developed to rival medical doctors in their ability to identify skin cancer from skin lesion images[4] and diabetic retinopathy from retinal fundus photographs[5] (both of which were trained on over one hundred thousand labeled images). However, in many cases, problems in biomedicine involve key limitations to data that preclude our ability to straightforwardly train a complex ML model on a large complete dataset. In this work, we present multiple projects designed to address various real-world data limitations by leveraging the use of ML and interpretability methods to better predict, relate, and ultimately understand biomedical phenomena.

Although many recent developments in machine learning methods assume the availability of large datasets with complete feature sets, biomedical applications often involve key data limitations. In the field of systems biology, we are often faced with the challenge of modeling highly complex phenomena (e.g., tens of thousands of genes) with limited sample sizes. Desired data is often prohibitively expensive to collect at large scale, and often the number of features may far exceed the number of samples. For example, in Chapter 2, we use brain gene expression data and Alzheimer's disease (AD) neuropathology data, which may only be obtained during autopsy from individuals who have agreed to donate their brains to science. While this form of data is highly valuable for computational drug discovery research, such studies to gather these samples require large amounts of funding and tremendous generosity on the part of participants, and thus only a handful of studies have each collected on the order of a few hundred samples each[6–10]. Moreover, because such studies have often been conducted independently of each other, there are often inconsistencies in what kinds of data were collected leading to feature and label-level sparsity if one wants to combine data across multiple datasets for increased statistical power (as described in Chapter 2), requiring solutions that can accommodate lower sample size and label sparsity.

Beyond facing the challenge of modeling complex phenomena with relatively scarce samples, researchers often find that results from such studies are inherently difficult to interpret. For example, many computational biology studies aim to identify groups of genes that are associated with or differentially expressed with respect to phenotypes of interest[11,12]. Pathway enrichment analysis is commonly used as a crucial step to associate these gene-level findings with biological processes, providing both biological

context for genes and a systems perspective to the analysis. However, these efforts are complicated by the growing number of pathway databases and resources. Thus, discovery efforts may be hindered by an inability to understand and relate complex relationships among current biological pathways and their relationships with each other (Chapter 3).

So far we have discussed limitations in developing and understanding ML models for the sake of scientific discovery. Shifting towards healthcare applications, we've seen the emergence of AI solutions for medical tasks such as diagnostics and risk prediction[5,13]. In practice, when these models are deployed, they often assume that they can be applied to new samples for which we receive an equally rich set of features as was used during training. However, in many clinical settings, we may only have a limited amount of information about a person (i.e., observed features), or limited time or resources with which to gather more information about them (i.e., budget with which to collect features, which we explore in Chapters 4 and 5).

In the projects described below, we aim to address real-world biomedical data limitations by adapting existing ML methods towards these challenges. In this work, interpretability plays an important role in each project, either as a means to understand a model's learned representation of complex relationships (e.g., Chapter 2), the goal itself (e.g., Chapter 3), or as way of identifying and prioritizing relevant features for a prediction model (e.g., Chapters 4 and 5). We present four projects ranging from molecular biology to healthcare applications, which each address key data limitations in the biomedical domain, and briefly summarize here and in Table 1:

- **A unified framework to uncover relationships between gene expression and Alzheimer's disease neuropathologies using multi-task deep learning (Chapter 2).** Because brain gene expression data is difficult to collect leading to relatively smaller data sets, we combined data from multiple aging cohorts along with their sparsely measured AD-related phenotypes. We jointly modeled several related AD phenotypes with a multi-task neural network, and were thus able to learn a unified representation between gene expression and neuropathology across multiple datasets. By interrogating the learned representations of our model with an interpretability method, we were further able to uncover more nuanced gene-phenotype relationships than those previously identified with linear models.
- **An automatic integrative method for learning interpretable communities of biological pathways (Chapter 3).** Although knowledge of biological pathways is essential for interpreting results from computational biology studies, the growing number of conflicting and redundant pathway databases complicates efforts to efficiently perform pathway analysis. Our method reconciles pathways from different databases and reduces pathway redundancy by revealing

informative groups with distinct biological functions via a Louvain community detection algorithm applied to a network of pathways from multiple databases. Our approach, combined with an interpretable web tool we provide, may help computational biologists more efficiently contextualize and interpret their biological findings.

- **Efficient and Explainable Risk Assessments for Imminent Dementia (Chapter 4).** As the aging population grows, scalable approaches are needed to identify individuals at risk for dementia. Common prediction tools have limited performance, involve expensive neuroimaging, or require extensive and repeated cognitive testing. By experimenting with ML models trained on several years of clinical data, and by applying interpretability methods, we identified a set of metrics collected in a single visit that can accurately predict imminent dementia onset with comparable accuracy to a standard neuropsychological battery which would take five times longer to administer.
- **Explanation-guided dynamic feature selection for medical risk prediction (Chapter 5).** Building on our efficient dementia risk prediction project above, we propose a more general method for evaluating health risks for an individual in settings where we have limited knowledge of their full feature set (e.g., a sparsely populated electronic health record). Our approach uses a fixed prediction model, a local feature explainer, and ensembles of imputed samples to generate risk prediction intervals and context-dependent feature recommendations, which we demonstrate on a myocardial infarction prediction task.

Table 1. Outline of projects' goals and challenges, along with ML and interpretability-based solutions. For each project, we summarize the main challenges, introduce the machine learning and interpretability components of our approach, and provide a brief explanation for how those methods provide a solution to the challenges introduced.

	Project (Chapter)	Challenges	Machine learning approach	Interpretability component
			→ Solution	
Systems biology	**MD-AD (2)**	Limited samples; Heterogeneous data and sparse labels when combining datasets	Multi-task deep learning	Integrated gradients to identify relevant genes for the model
			→ unified model to relating gene expression and Alzheimer's disease neuropathology + refined our understanding of the molecular basis of AD via model interpretations	
	PAC (3)	Unintuitive and inconsistent names across databases; High redundancy between pathways within and across databases	Represent pathways as a graph → Louvain algorithm to learn communities of related pathways	Querying new gene sets → adds graph context to enrichment results
			Graph/community-based approach → provides context of pathway relationships for enrichment analyses	
Medical risk prediction	**Efficient dementia prediction (4)**	Access to care and physician time is limited; Few elderly individuals receive regular cognitive testing	Risk prediction model: Gradient boosted decision trees (XGBoost)	SHAP explainer → identifies globally important features and explains individual risk predictions
			XGBoost + explanation-based feature selection → efficient dementia onset prediction with lower burden on clinicians	
	Explanation-guided dynamic feature selection (5)	Missing features at test time; Limited resources to collect features	Model agnostic: post-hoc approach involving fixed ML prediction model and imputer	Feature explanations guide feature selection and explain risk estimates
			Given partially observed data, we use a fixed model + model explanations to generate flexible risk predictions and context-dependent feature recommendations	

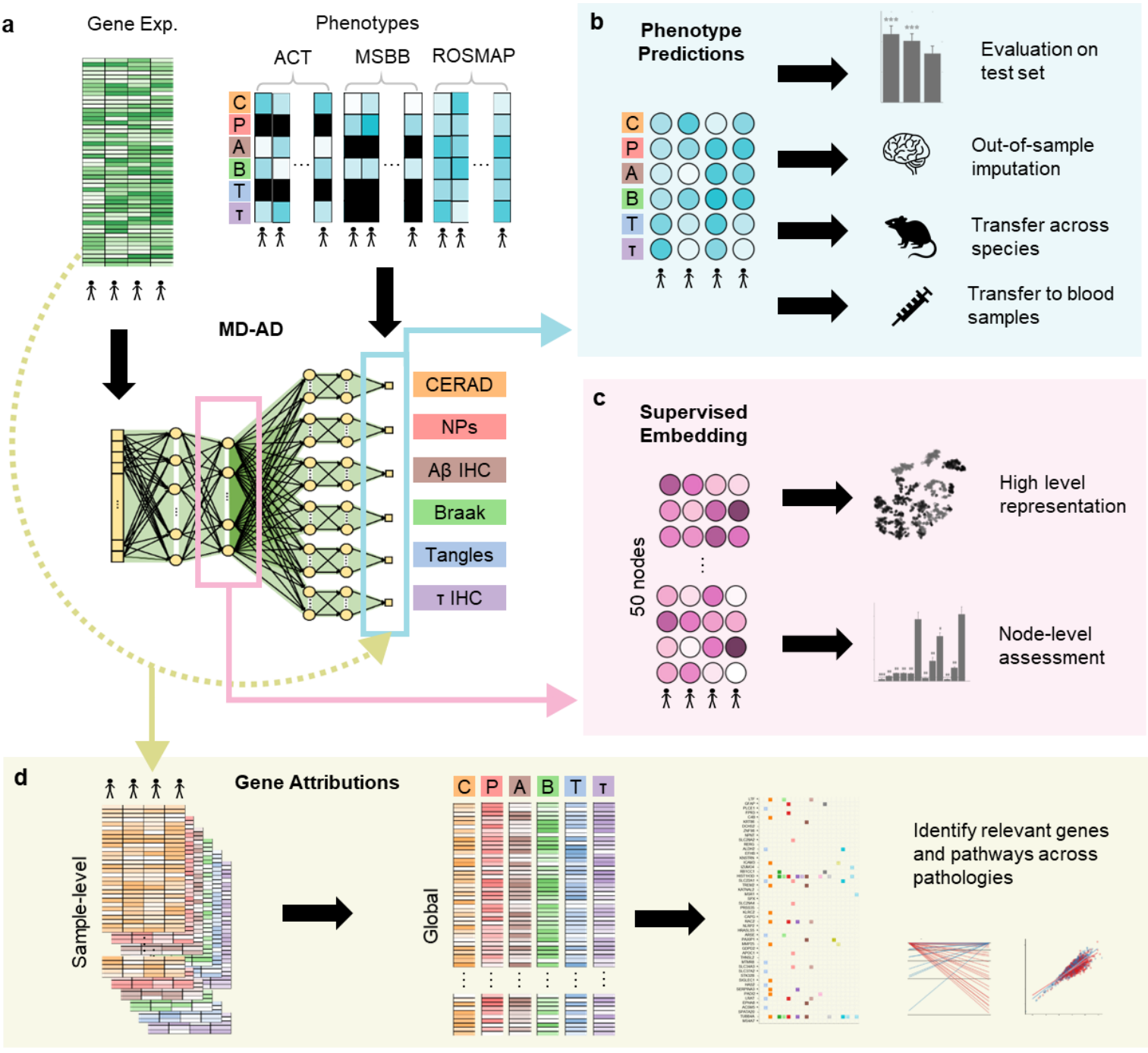
Chapter 2. Unified AI framework to uncover deep interrelationships between gene expression and Alzheimer's disease neuropathologies

Deep neural networks (DNNs) capture complex relationships among variables, however, because they require copious samples, their potential has yet to be fully tapped for understanding relationships between gene expression and human phenotypes. Here we introduce an analysis framework, namely MD-AD (Multi-task Deep learning for Alzheimer's Disease neuropathology), which leverages an unexpected synergy between DNNs and multi-cohort settings. In these settings, true joint analysis can be stymied using conventional statistical methods, which require "harmonized" phenotypes and tend to capture cohort-level variations, obscuring subtler true disease signals. Instead, MD-AD incorporates related phenotypes sparsely measured across cohorts, and learns interactions between genes and phenotypes not discovered using linear models, identifying subtler signals than cohort-level variations which can be uniquely recapitulated in animal models and across tissues. We show that MD-AD exploits sex-specific relationships between microglial immune response and neuropathology, providing a nuanced context for the association between inflammatory genes and Alzheimer's Disease.[*]

Introduction

Alzheimer's disease (AD), the sixth leading cause of death in the United States, is a degenerative brain condition with no known treatment to prevent, cure, or delay its progression. Primary challenges to treating and preventing AD include extensive heterogeneity in the clinicopathologic state of older individuals[14] and limited knowledge about genetic and molecular drivers and suppressors of AD-related (amyloid and tau) proteinopathies and AD dementia[15]. Recent efforts to identify molecular mechanisms underlying AD and its progression focus on two complimentary approaches. First, the assembly of large genome-wide association studies (GWAS) (N>100K subjects) enabled case/control analyses of genetic variants correlated with a clinical diagnosis of AD. Interestingly, some identified variants have implicated tau protein binding, amyloid precursor protein (APP) metabolism or immune pathways that play a role in their aggregation and/or uptake[16–18]. These results reinforce the need for detailed investigations of the drivers of neuropathological variation across individuals. Second, moderate-scale post-mortem transcriptomic studies

[*] This paper was joint work with Safiye Celik, Ethan Weinberger, Pascal Sturmfels, Philip L. De Jager, Sara Mostafavi & Su-In Lee. It has been published in *Nature Communications* [129].



e

	# Expression Samples	CERAD	NPs	Aβ IHC	Braak	Tangles	τ IHC
ACT	337	✓		✓	✓		✓
MSBB	879	✓	✓		✓		
ROSMAP	524	✓	✓	✓	✓	✓	✓

Figure 2.1. Overview of the MD-AD (Multi-task Deep learning for Alzheimer's Disease neuropathology) method and analyses. (**a**) Overview of the MD-AD framework: MD-AD is trained to predict six neuropathology phenotypes simultaneously from brain gene expression samples. During model training, samples do not need to have all available phenotypes; they influence only the layers for which they have labels (including shared layers). (**b**) Illustrates out-of-sample datasets we used to validate MD-AD's predictions. (**c**) Illustrates analyses used to validate the last shared layer of MD-AD. (**d**) By using model interpretability methods, we highlight genes relevant to MD-AD's predictions. Further analyses reveal non-linear effects among genes and their relationship with AD severity prediction. **(e)** Overview of data available from each cohort.

have investigated molecular correlates of a richer set of phenotypic and neuropathological outcomes[8,19–21]. Early work in this domain examined pairwise correlations among gene expression levels and AD related traits[22] or a diagnosis of AD[23]. More recent attempts have focused on learning statistical dependencies among gene expression using AD expression data collected from one cohort in order to infer gene regulatory networks[19] or co-expressed modules [8] associated with AD related phenotypes (see Methods for details). The relative scarcity of brain gene expression data collected from each cohort has posed a challenge to the use of complex models, such as deep neural networks.

The collection of postmortem brain RNA-sequencing datasets, assembled by the AMP-AD (Accelerating Medicines Partnership Alzheimer's Disease) consortium, provides a unique opportunity to combine multiple data sets in an integrative analysis. Previous work has applied existing co-expression methods to each dataset and used consensus methods to identify consistent gene expression modules across datasets[21]. To our knowledge, there has not yet been a unified approach to learn a single joint model that incorporates multiple AMP-AD datasets, which would enable the use of all samples to capture intricate interactions between gene expression levels and neuropathological phenotypes. A unified approach has been hindered by: (1) the need for "harmonized" phenotypes consistently measured across datasets, and (2) the limitation of current analysis methods that focus on linear relationships between variables (e.g., module analysis[21]) which capture only broad patterns in gene expression data. These often correspond to cohort-level variations which consequently obscure true disease signals[24]. To circumvent this issue, one approach has been to identify modules separately across brain regions and cohorts before performing using a consensus approach to cluster them[21].

Here, we develop MD-AD (Multi-task Deep learning for Alzheimer's Disease neuropathology), a unified framework for analyzing heterogeneous AD datasets to improve our understanding of an expression basis for AD neuropathology (Figure 2.1a-d). Unlike previous approaches, MD-AD learns a single neural network by jointly modeling multiple neuropathological measures of AD (Figure 2.1a), and hence it incorporates the largest collection of postmortem brain RNA-sequencing datasets assembled to date. The

combined AMP-AD dataset contains 1,758 samples distributed across 9 brain regions which are labeled with up to six neuropathological outcomes that are sparsely available across cohorts (Figure 2.1e). This unified framework has key advantages over separately trained models. First, MD-AD can accommodate sparsely labeled data, which is a natural characteristic of datasets aggregated through consortium efforts (Figure 2.1e). Even if different phenotypes only partially overlap in the measured samples, each sample contributes to the training of both phenotype-specific and shared layers (Figure 2.1a). Predicting multiple outcome variables at once biases shared network layers to capture relevant features of all those outcome variables (here, neuropathological phenotypes) at the same time[25]. This is of critical importance in our application: each neuropathological phenotype represents a different noisy measurement of the same underlying true biological process, and, as we demonstrate, joint training with these phenotypes allows MD-AD to average out the noise to extract the true hidden signal. Additionally, the increased sample size from combining cohorts (in our case, doubling the number of samples available from any individual study) facilitates using deep learning models, which are expressive and able to capture complex non-linear interactions among features. By composing layers of functions, deep neural networks collapse correlation patterns present in input data at intermediate layers in a way that is useful for prediction[26]. In particular, multi-layer perceptrons (MLPs) have been used to effectively perform disease classification and prediction from gene expression data[27–29]. However, training separate MLPs for each neuropathological phenotype (Supplementary Figure 2.1a) has limited scope: it can utilize only the samples measured for a specific phenotype, and it cannot share information across related phenotypes. We demonstrate that MD-AD's joint training approach improves prediction accuracy, enabling its predictions to generalize across species and tissue types (Figure 2.1b).

An obvious drawback of deep neural networks is their black-box nature, making it difficult to biologically interpret gene-phenotype associations that have been learned by a model. We present two ways to address this challenge. First, MD-AD adopts a well-known feature attribution method[30], which quantifies how much each input variable (here, gene expression level) contributes to a prediction (here, a neuropathological phenotype) to identify genes and pathways relevant to each neuropathological phenotype (Figure 2.1d). Second, because MD-AD is a deep learning model, we can interpret its intermediate layers as biologically relevant high-level feature representation of gene expression levels and its predictions as the amalgamation of AD-specific molecular markers. The last shared layer of MD-AD can be viewed as a supervised embedding influenced by each neuropathological phenotype used during training. Thus, by interpreting this layer's embedding, we gain understanding of model components and high-level dependencies between expression and neuropathology (Figure 2.1c). We identify globally important genes

not previously implicated in linear methods and then perform sex-specific analyses to explore implicitly captured non-linear effects among genes and their differing relationship with AD severity predictions.

In sum, the MD-AD framework makes the following contributions: (1) It is able to effectively impute accurate AD neuropathological phenotype predictions from broad compendia of heterogeneous brain gene expression data; (2) it produces learned representations that are more robust than separately learned models, improving generalizability to other datasets, species, and even tissue types; (3) it provides an improved understanding of inter-relationships among molecular drivers of AD neuropathology that is missed by linear methods; and (4) from a biological standpoint, MD-AD highlights a sex-specific relationship between microglial immune activation and neuropathology.

Results

MD-AD provides a unified framework to learn a single model of multiple neuropathological phenotypes across multiple cohort datasets

The MD-AD model takes as input brain gene expression profiles and simultaneously predicts several AD-related neuropathological phenotypes (Figure 2.1a). In particular, the model is trained on expression data from the ROSMAP [6–8], ACT [9] and MSBB [10] cohort studies, which together have 1,758 gene expression profiles for 925 distinct individuals (with no participant overlap between cohorts). These data are normalized for study batch (Methods, Supplementary Figure 2.1d) [31]. As shown in Figure 2.1a, the MD-AD model simultaneously predicts six AD-related neuropathological phenotypes: three related to amyloid plaques and three to tau tangles. The former include: (1) Aβ IHC: amyloid-β protein density via immunohistochemistry, (2) NPs: neuritic amyloid plaque counts from stained slides, and (3) CERAD score: a semi-quantitative measure of neuritic plaque severity [32]. The latter include: (4) τ IHC: abnormally phosphorylated τ protein density via immunohistochemistry, (5) tangles: neurofibrillary tangle counts from silver stained slides, and (6) Braak stage: a semi-quantitative measure of neurofibrillary tangle pathology [33]. Thus, MD-AD generates six highly related predictions simultaneously and covers each of the two main hallmarks of AD neuropathology (plaques and tangles) at three levels of granularity. The three studies measure partially overlapping subsets of the six neuropathological phenotypes described above (Figure 2.1e, Figure 2.1a, Supplementary Figure 2.1a and Supplementary Tables 2.1-2.2), so across our combined dataset some variables are sparsely labeled, although Braak and CERAD are each measured in all studies (Figure 2.1e). During training, the MD-AD model continually updates model parameters via backpropagation, but only for labeled phenotypes from a given sample. Thus, for each phenotype for a given sample, MD-AD updates parameters from associated separate layers along with all shared layers.

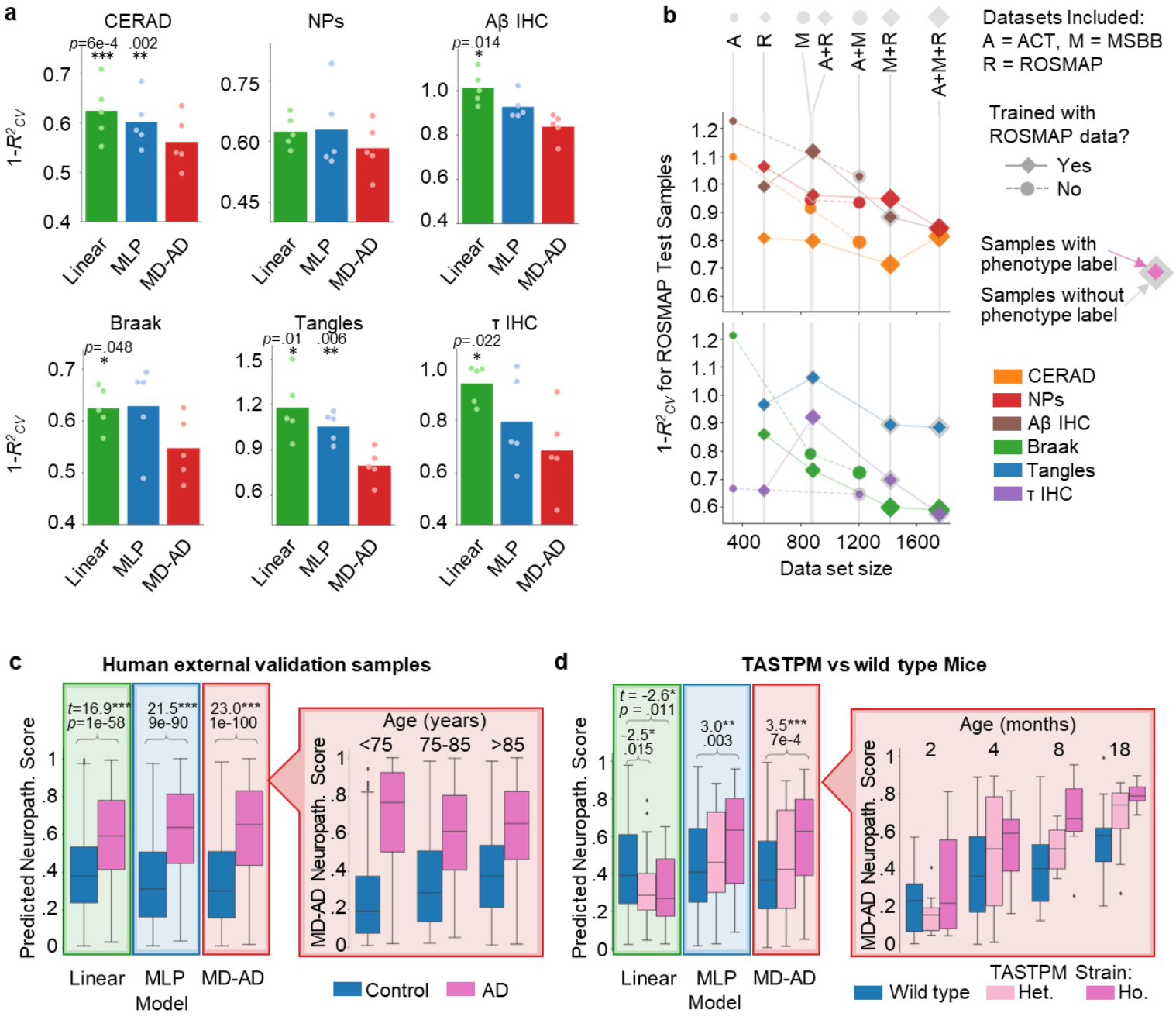


Figure 2.2. MD-AD prediction performance for within-sample test data and out-of-sample (external validation) data. **(a)** Average test set 1-R^2_{CV} (prediction error metric calculated by dividing the mean squared error by the label's variance) for phenotype predictions across 5 test splits (values from each run as dots). MLP: Multiple Layer Perceptron. Linear: linear model using L2 regularization (2-sided paired *t*-test comparing alternative methods with MD-AD, *p*-values: *<.05, **<.01, ***<.001; n=5 runs of each method). **(b)** Average 1-R^2_{CV} for ROSMAP test set samples when training on subsets of the available data sets in the training set. **(c)** For samples from three external validation data sets, we obtain neuropathology scores for each sample from each model. *Left*: Box plots displaying the distribution of predicted neuropathology scores from each method. T-test highlight between-group differences for each method (2-sided independent *t*-test, ***p<.001) *Right*: Box plots displaying the distribution of MD-AD's

predicted neuropathology scores split by age group and diagnosis (see **Supplementary Figure 4b** for sample sizes broken down by age and diagnosis and significance of pair-wise differences). Test statistics were calculated based on 565 AD cases and 482 controls. All box plots in this figure indicate median (center line), upper and lower quartiles (box limits), 1.5x interquartile range from quartiles (whiskers), and outliers (points). **(d)** *Left*: Box plots displaying the distribution of predicted neuropathology scores from each method for wild type and TASTPM (both heterozygous and homozygous) mice. T-tests highlight between-group differences for each method (2-sided *t*-test, *p*-values: *<.05, **<.01, ***<.001). *Right*: Box plots displaying the distribution of MD-AD's predicted neuropathology scores for mice split by age and strain (See **Supplementary Figure 4c** for sample sizes and significance of pair-wise differences). Test statistics were calculated based on 72 wild type, 32 heterozygous TASTPM and 30 homozygous TASTPM mice (same box plot elements as described in part b).

This lets us train a unified model from all available samples despite having many missing labels. Although in our application neuropathological phenotypes overlapped across datasets, MD-AD could accommodate non-overlapping phenotypes from different cohorts (as long as they are believed to be closely related and share a common underlying gene expression basis). Details of the MD-AD framework, modeling assumptions, and hyperparameter tuning are provided in Methods.

MD-AD accurately predicts neuropathology from gene expression, and its predictions are generalizable to external datasets.

In the first pass at model evaluation, we trained MD-AD using standard five-fold cross-validation (CV), and assessed the average $1\text{-}R^2_{CV}$ error (mean squared error divided by the phenotype's variance in the test set) on the held-out test samples (Figure 2.2a). Our hyperparameter tuning and evaluation procedures are described in detail in the Methods and Supplementary Figure 2.1e. We compared MD-AD to two simpler baseline models: a regularized linear model (ridge regression) and a single output deep neural network (MLP). These alternative results helped us assess two significant components of the MD-AD model: (1) its non-linear modeling of the relationship between gene expression and neuropathological phenotypes, and (2) its joint modeling of multiple related neuropathological phenotypes. In general, MLP models outperformed linear models, highlighting the advantage of deep learning over a linear approach. Furthermore, compared to the MLP models, MD-AD reduced the prediction error by 7% for CERAD score, 13% for Braak stage, 7% for NPs, 25% for tangles, 10% for Aβ immunohistochemistry (IHC), and 14% for τ IHC (Figure 2.2a). Interestingly, MD-AD showed its largest performance gain for the tangles variable, which also had the most missing labels (Figure 2.1e), highlighting a specific advantage of joint learning for sparsely labeled data. We additionally experimented with some alternative approaches (e.g., different training/test splits, covariate-corrected data) and found that performance results were robust to these changes (Methods).

Because our model was trained and evaluated on ACT, MSBB, and ROSMAP datasets, we assessed whether residual (uncorrected) batch effects affected performance. To do so, we performed additional validation experiments by leaving out specific datasets during training and then evaluating their performance for MD-AD trained on the other datasets (Figure 2.2b, Supplementary Figure 2.2a). We evaluated the prediction error for ROSMAP alone since it was the only dataset with all six phenotype labels; further, by evaluating a single dataset's performance, we can identify the influence of adding "external" data. We make several observations from this analysis. First, as one may expect, larger training samples always helped to reduce prediction error on test samples from the unseen study (ROSMAP), and especially so when datasets from multiple cohorts were included in the training (i.e., ACT and MSBB) (circular markers in Figure 2.2b). Second, when considering the effects of augmenting ROSMAP data with other datasets during training (diamond markers in Figure 2.2b), we observed that errors initially increased when adding a new dataset but tended to decline as more datasets were included in training. This may result from small differences in labeling conventions across studies, or batch effects in gene expression data. However, we find that the benefits of additional heterogeneous samples ultimately outweigh potential batch effects in prediction performance. Third, we observed that adding new samples improved performance for a neuropathological phenotype even when the phenotype in question was not measured in the new samples (see gray footprints around markers in Figure 2.2b). The same analysis repeated with the other two cohorts as test sets revealed similar findings (Methods, Supplementary Figure 2.3). This suggests that the shared representation learned by MD-AD (which is improved by access to additional sparsely labeled samples) captures the underlying biological signal common across noisy neuropathological phenotype measurements.

Next, as the ultimate test of MD-AD out-of-sample predictions, we assessed performance on three independent studies never seen by the model: Mount Sinai Brain Bank Microarray (MSBB-M; N=1,047; 565 AD cases and 482 controls), Harvard Brain Tissue Resource Center (HBTRC; N=338; 246 AD cases and 92 controls)[19], and Mayo Clinic Brain Bank (N=157; 81 AD cases and 76 controls)[34]. Because these datasets provide a sparse set of neuropathological labels, we evaluated whether MD-AD predictions were consistent with the (binary) neuropathological diagnosis of AD by calculating "MD-AD neuropathology scores" for each sample (by averaging ranked predictions across the six neuropathological phenotypes).

As shown in Figure 2.2c, we observed a highly significant difference in predicted neuropathology scores between AD cases and controls (two-sided t-test: $t = 22.98$, $p < 0.001$), and these differences were more pronounced for MD-AD compared to the other baseline models (results split by dataset are shown in Supplementary Figure 2.4a). More convincingly, when split by age group (Figure 2.2c right panel), we consistently observed a significant increase in predicted neuropathology for AD vs control samples, but the

difference was largest in individuals under 75 (between-groups *p*-values are shown in Supplementary Figure 2.4b. The same analysis comparing APOE ε4 carriers to non-carriers revealed a similar pattern, shown in Supplementary Figure 2.5). This is consistent with the observation that aging individuals who are cognitively non-impaired often have substantial neuropathology[9]. Together, these results indicate that MD-AD can identify generalizable gene expression patterns that are predictive of AD-related neuropathology across varied age ranges, and thus it is unlikely that these patterns merely capture normal aging.

Complex transcriptomic predictors of neuropathology are conserved across species.

We next evaluated how well MD-AD's learned expression patterns predictive of neuropathology recapitulated neuropathology in mouse models. We applied MD-AD trained on human datasets to make predictions based on 30 brain (hippocampal and cortical) gene expression samples from TASTPM mice that harbored double transgenic mutation in *APP* and *PSEN1* and compared the predictions to those for 76 samples from wild type mice[35,36]. We focused on TASTPM mice since they were found to robustly exhibit early signs of amyloid aggregation and plaque formation. As above, to simplify MD-AD predictions, we then predicted all six neuropathological phenotypes via MD-AD and generated an aggregate "neuropathology score" per mouse sample (as described in Methods).

As shown in Figure 2.2d, MD-AD predicted significantly higher neuropathology scores for the homozygous cross TASTPM than wild type mice (two-sided t-test: t=3.45, p <.001). The MLP baseline method also produced significant differences between homozygous and wild type mice, but less effectively (t=3.01, p<.01). Furthermore, there was a stronger trend for higher predictions in the heterozygous TASTPM cross samples (N=32) than wild type mice for MD-AD (t=1.38, p=.17) compared to MLP baselines (p=.38). The linear baseline model failed to make accurate predictions. None of the models produced significantly different neuropathology scores between other strains (i.e., TPM, TAS10, Tau) and wild type mice, consistent with lower neuropathological burden in these models (Supplementary Figure 2.4e). Notably, when we stratified the samples by age, we found that MD-AD tended to predict higher neuropathology in older mice (regardless of strain), but in particular it made higher neuropathology predictions for homozygous than heterozygous crosses followed by wild type mice (many of these groups differed significantly from one another, as shown in Supplementary Figure 2.4c). Overall, these results indicate that MD-AD learns a generalizable expression pattern associated with neuropathology that is conserved across species.

Deep transcriptomic signatures of neuropathology are predictive of AD dementia

Hidden layers of a deep neural network capture the embedding of input examples in the derived feature space, yielding a "hidden" representation that is predictive of the outcome(s) of interest. In this case, the last shared layer of MD-AD (Figure 2.1a, c) captures a latent (lower) dimensional representation of gene expression that is predictive of multiple types of neuropathology related to AD. To derive the biological basis of MD-AD predictions, we first visualized this embedding space in 2D using the t-SNE algorithm (Figure 2.3a)[37] (to improve stability, we used a consensus approach over many re-trainings of the MD-AD model, Supplementary Figure 2.6a). We observed that the representation in this space was impressively coherent with respect to all six neuropathological variables: individuals with similar overall neuropathology severities had similar MD-AD consensus representations for their gene expression profiles, and this observation was true for external test samples not used for model training (Figure 2.3d-e, Supplementary Figure 2.4d). This was remarkable because representations derived by unsupervised dimensionality reduction (e.g., K-means or PCA) failed to capture the components of gene expression relevant to neuropathology, and mainly captured effects relatable to batch or brain region differences, while those derived by standard single output MLP tended to overfit to each neuropathology variable and were incoherent *across* neuropathological measurements (Figure 2.3c and Supplementary Figure 2.7).

Next, we evaluated whether the MD-AD embedding can go beyond neuropathology to also capture the molecular manifestation of AD dementia. In particular, we considered three "higher-level" clinical phenotypes: AD dementia (a clinical diagnosis of AD), assessment of cognitive function, and assessment of AD duration. We then correlated the latent representation captured by the hidden nodes in the last shared layer with each of these three higher-level phenotypes. As shown in Figure 2.3b, we found that MD-AD consistently produced nodes that were significantly correlated with high-level AD phenotypes; using paired *t*-tests, these correlations often outperformed nodes from our MLPs and always outperformed unsupervised methods and module-based approaches ($p<.05$ after FDR correction over nodes). This indicates that MD-AD creates embeddings that most consistently capture the relationship between gene expression and general AD severity. Together, these results show that by jointly predicting several neuropathological phenotypes, the MD-AD framework produces a low dimensional representation of gene expression data that robustly captures a generalizable signature of AD beyond individual neuropathological phenotypes alone. Detailed annotations for MD-AD embedding nodes are provided in Supplementary Table 2.3 and Supplementary Figure 2.6b-d.

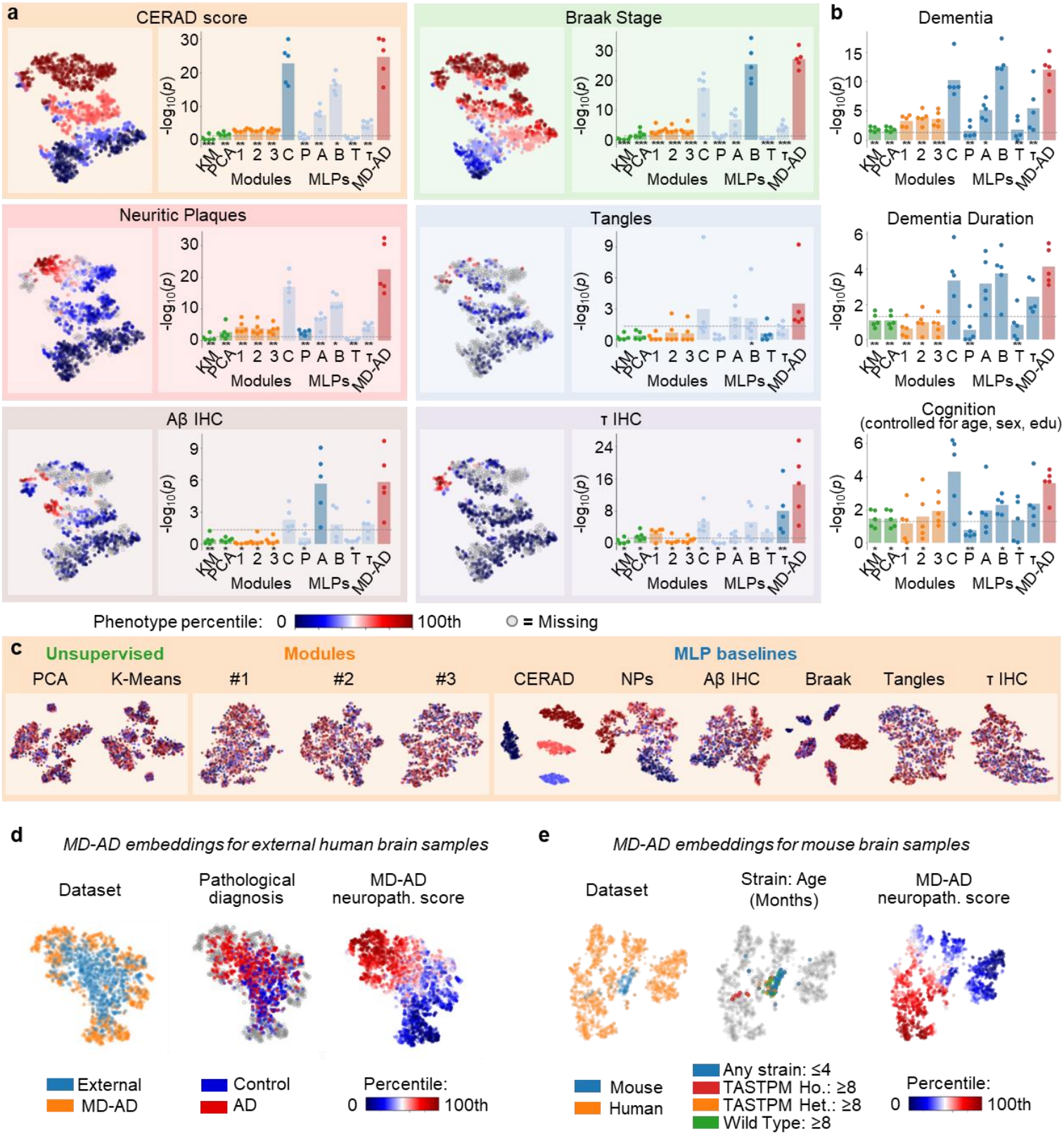

a
CERAD score
Braak Stage
Neuritic Plaques
Tangles
Aβ IHC
τ IHC
-log10(p)
KM PCA 1 2 3 C P A B T τ MD-AD
Modules
MLPs
b
Dementia
Dementia Duration
Cognition
(controlled for age, sex, edu)
Phenotype percentile: 0 100th
= Missing
c
Unsupervised
Modules
MLP baselines
PCA
K-Means
#1
#2
#3
CERAD
NPs
Aβ IHC
Braak
Tangles
τ IHC
d
MD-AD embeddings for external human brain samples
Dataset
Pathological diagnosis
MD-AD neuropath. score
External
MD-AD
Control
AD
Percentile:
0 100th
e
MD-AD embeddings for mouse brain samples
Dataset
Strain: Age (Months)
MD-AD neuropath. score
Mouse
Human
Any strain: ≤4
TASTPM Ho.: ≥8
TASTPM Het.: ≥8
Wild Type: ≥8
Percentile:
0 100th

Figure 2.3. Comparing MD-AD's supervised embedding to other embedding methods. **(a)** For each colored box, *Left*: 2-dimensional t-SNE embedding of MD-AD's last shared layer colored by neuropathological phenotype indicated in the title of the box, *Right*: -$\log_{10}$(*p*-value) of correlations (averaged across 5 folds) between "best" node from each embedding method and the neuropathological phenotype across 5 test folds. The "best" node was identified as the most significantly correlated in the training set, but bar height indicates the Pearson correlation -$\log_{10}$(*p*-value) of the node with the phenotype in their corresponding test sets after FDR correction across nodes (averaged results over five runs; individual points show –log10(p-value) from each run). Bar graph columns (left to right): two unsupervised embeddings (green; K-Means and PCA), three module-based embeddings (orange; Modules #1[19] and Modules #2[8], and Modules #3[21]), six singly-trained MLPs (blue), and MD-AD (red). Results from each method were compared with MD-AD using two-sided paired *t*-tests (*p*-values indicated below each bar: *<.05, **<.01, ***<.001). **(b)** Highest correlation -$\log_{10}$(*p*-values) (averaged across 5 folds) found between each embedding method and high-level AD variables: dementia (diagnosis prior to death), dementia duration (approximate time between dementia diagnosis and death; available for ACT and ROSMAP), and last available cognition score (controlling for age, sex and education; available for ROSMAP only). All *p*-values listed are shown after FDR correction over the nodes within each method. Bar height indicates the mean over 5 folds, and points show individual values from each run. **(c)** 2-dimensional t-SNE embedding of alternative embedding methods (described in **a**), colored by CERAD scores associated with each sample. **(d)** 2-dimensional t-SNE embeddings of MD-AD embeddings for training and external data sets. Each point represents a sample colored by dataset (Left), AD status for external samples (Middle) and MD-AD's predicted neuropathology score (Right). (**e**) 2-dimensional t-SNE embeddings of MD-AD embeddings for external human and mouse samples.

MD-AD reveals an interrelationship between sex and immune genes predictive of AD neuropathology

We next sought to interpret MD-AD's learned parameters to identify the set of genes (and their relationships) that underlie its impressive predictive performance. Integrated Gradients (IG)[30], one of the most widely used interpretability methods developed for deep neural networks, estimates the importance of input features on a model's predicted output for a particular input sample (See Methods for details). Here, we applied the IG algorithm on the fully trained model in an ensemble fashion to ensure robustness (Methods, Supplementary Figure 2.8), producing an "importance score" for each gene (Supplementary Table 2.4). For a global view, we first performed functional enrichment analysis (GSEA[38,39]) using these importance scores (aggregated across samples), and found that relevant genes for the MD-AD model were enriched for several pathways, including metabolism of RNA and proteins, immune system, cell-to-cell communication, and signal transduction (Figure 2.4b). Figure 2.4a shows the top 50 genes and their pathway annotations where the particular relevance of immune function is even more prominent.

We next assessed to what extent the learned gene importance varied between a linear model and a non-linear model like MD-AD. With a simple linear correlation-based gene ranking (Methods), we found that the top 50 genes were less likely to be annotated to REACTOME pathways (Supplementary Figure 3.9a). When we directly compared the top 1% of genes from MD-AD versus a correlation-based approach in Figure 2.4c, we observed that many genes belonging to metabolism, immune system, and signal

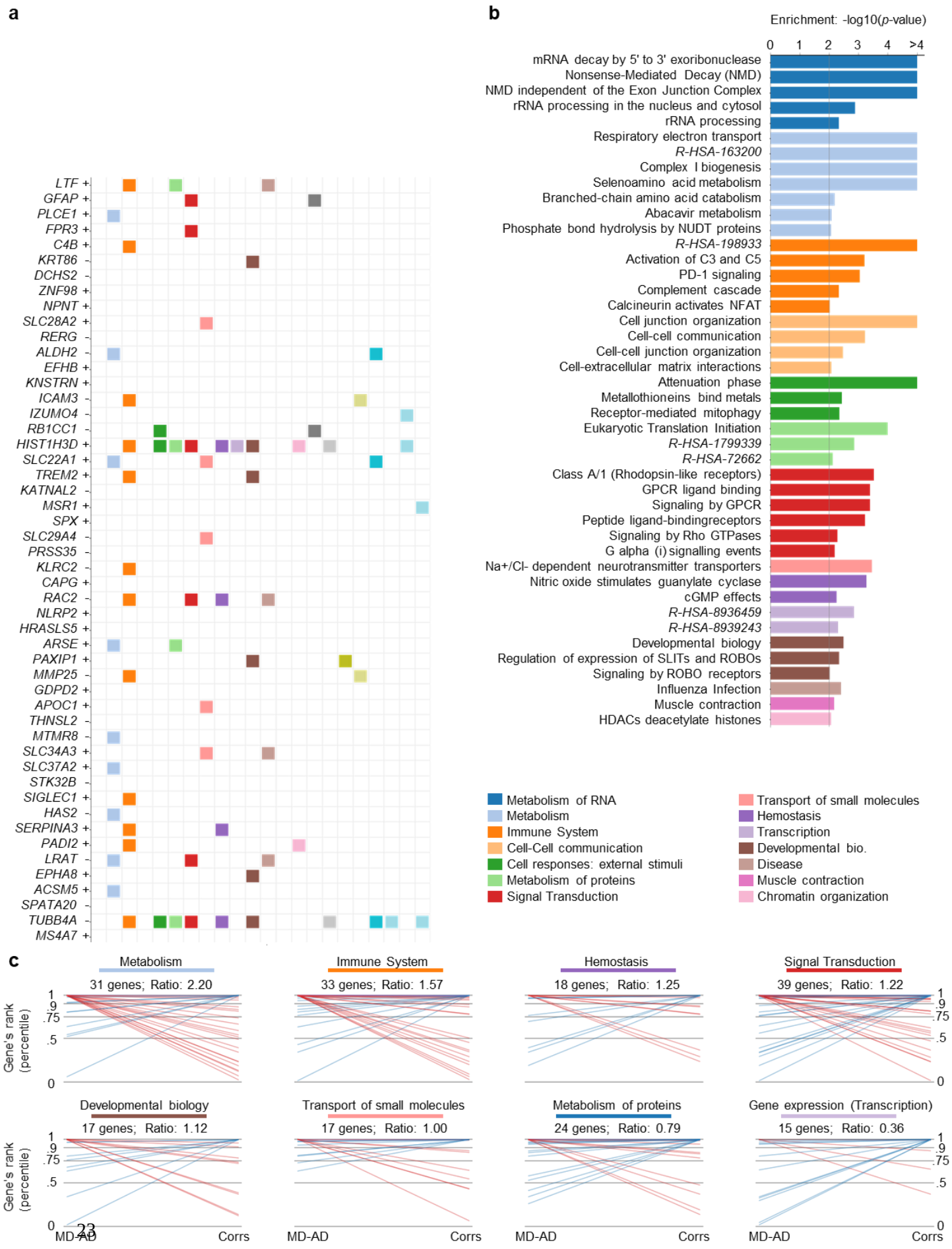

a
LTF +
GFAP +
PLCE1 +
FPR3 +
C4B +
KRT86 -
DCHS2 -
ZNF98 +
NPNT +
SLC28A2 +
RERG -
ALDH2 -
EFHB +
KNSTRN +
ICAM3 -
IZUMO4 -
RB1CC1 -
HIST1H3D +
SLC22A1 +
TREM2 +
KATNAL2 -
MSR1 +
SPX +
SLC29A4 -
PRSS35 -
KLRC2 -
CAPG +
RAC2 -
NLRP2 +
HRASLS5 +
ARSE +
PAXIP1 +
MMP25 -
GDPD2 +
APOC1 +
THNSL2 -
MTMR8 -
SLC34A3 +
SLC37A2 +
STK32B -
SIGLEC1 +
HAS2 -
SERPINA3 +
PADI2 +
LRAT -
EPHA8 +
ACSM5 +
SPATA20 -
TUBB4A -
MS4A7 +
b
Enrichment: -log10(p-value)
0 1 2 3 4 >4
mRNA decay by 5' to 3' exoribonuclease
Nonsense-Mediated Decay (NMD)
NMD independent of the Exon Junction Complex
rRNA processing in the nucleus and cytosol
rRNA processing
Respiratory electron transport
R-HSA-163200
Complex I biogenesis
Selenoamino acid metabolism
Branched-chain amino acid catabolism
Abacavir metabolism
Phosphate bond hydrolysis by NUDT proteins
R-HSA-198933
Activation of C3 and C5
PD-1 signaling
Complement cascade
Calcineurin activates NFAT
Cell junction organization
Cell-cell communication
Cell-cell junction organization
Cell-extracellular matrix interactions
Attenuation phase
Metallothioneins bind metals
Receptor-mediated mitophagy
Eukaryotic Translation Initiation
R-HSA-1799339
R-HSA-72662
Class A/1 (Rhodopsin-like receptors)
GPCR ligand binding
Signaling by GPCR
Peptide ligand-bindingreceptors
Signaling by Rho GTPases
G alpha (i) signalling events
Na+/Cl- dependent neurotransmitter transporters
Nitric oxide stimulates guanylate cyclase
cGMP effects
R-HSA-8936459
R-HSA-8939243
Developmental biology
Regulation of expression of SLITs and ROBOs
Signaling by ROBO receptors
Influenza Infection
Muscle contraction
HDACs deacetylate histones
Metabolism of RNA
Metabolism
Immune System
Cell-Cell communication
Cell responses: external stimuli
Metabolism of proteins
Signal Transduction
Transport of small molecules
Hemostasis
Transcription
Developmental bio.
Disease
Muscle contraction
Chromatin organization
c
Metabolism
31 genes; Ratio: 2.20
Immune System
33 genes; Ratio: 1.57
Hemostasis
18 genes; Ratio: 1.25
Signal Transduction
39 genes; Ratio: 1.22
Developmental biology
17 genes; Ratio: 1.12
Transport of small molecules
17 genes; Ratio: 1.00
Metabolism of proteins
24 genes; Ratio: 0.79
Gene expression (Transcription)
15 genes; Ratio: 0.36
Gene's rank (percentile)
1
.9
.75
.5
0
MD-AD
Corrs

Figure 2.4. Top predictive genes for the consensus MD-AD model. **(a)** Top 50 MD-AD genes and whether they are negatively (-) or positively (+) associated with high neuropathology. Colored squares indicate that the gene belongs at least one pathway in the column-labeled REACTOME category. **(b)** Gene set enrichment -log10(*p*-value) across the final MD-AD gene ranking for REACTOME pathways. Bars are colored by the pathway's REACTOME category. We show all pathways with significant enrichment ($p<.01$). REACTOME pathways with long names are indicated by their REACTOME stable IDs. **(c)** Comparison of top genes from MD-AD vs a linear correlation-based approach. For each ranking method, we identify the top 1% of all genes and check their membership in REACTOME categories. For each REACTOME category with at least 15 genes in the top 1% of MD-AD and/or correlation rankings, we generate the following plot: each line represents a gene, with left endpoint at the percentile rank for MD-AD and right endpoint at percentile rank for correlations. For clarity, we color the line purple if the gene falls in the top 1% of both MD-AD and correlations, red if it is only in the top 1% of MD-AD, and blue if it is only in the top 1% of correlations. Finally, the title indicates the ratio of MD-AD to correlation-based top genes for the given REACTOME category.

transduction pathways were highly ranked for MD-AD but not for correlation-ranking. In contrast, transcription-related genes were more frequently highly ranked for correlation-based rankings compared to MD-AD's rankings. Overall, gene importance scores generated via correlations alone were enriched for more REACTOME pathways (Supplementary Figure 2.9b), whereas MD-AD offered a more specific set of processes for further investigation (Figure 2.5b). We saw similar results when performing the same analyses with KEGG pathways (Supplementary Figure 2.10)[40].

The nonlinear relationships identified by MD-AD can implicitly capture interaction effects with other covariates observable from expression data (e.g., sex, age, medication intake). Leveraging the fact that, if our model captures a nonlinear effect, then two samples with the same expression level for a single gene could receive different IG ("importance") scores by MD-AD (e.g., Figure 2.5d; in contrast, a linear model would have no vertical dispersion), we assessed whether a covariate like sex could explain discrepancy between expression levels and IG scores. (Sex is a major risk factor in AD and has prominent gene expression signatures[41]). In particular, to identify sex-interacting genes relevant to AD we modeled each gene's per-sample IG score as a linear combination of the gene's expression, the individual's sex, and the interaction between them. Of the 14,591 genes in our dataset, 6,465 showed differential MD-AD importance between sexes in an interaction model ($p<0.05$ after FDR correction), demonstrating that sex-specific expression effects in AD may be widespread. When focusing on the top 100 genes with the highest MD-AD scores, we consistently observed high degrees of interaction between sex and immune system genes (as well as reproduction and hemostasis-related genes) (Figure 2.5a-b; we saw similar patterns for KEGG pathways in Supplementary Figure 2.11b-c). To confirm that genes are not sex-differential by chance, we show the distribution of sex-differential genes compared with the same analysis conducted with shuffled sex labels (Supplementary Figure 2.11a).

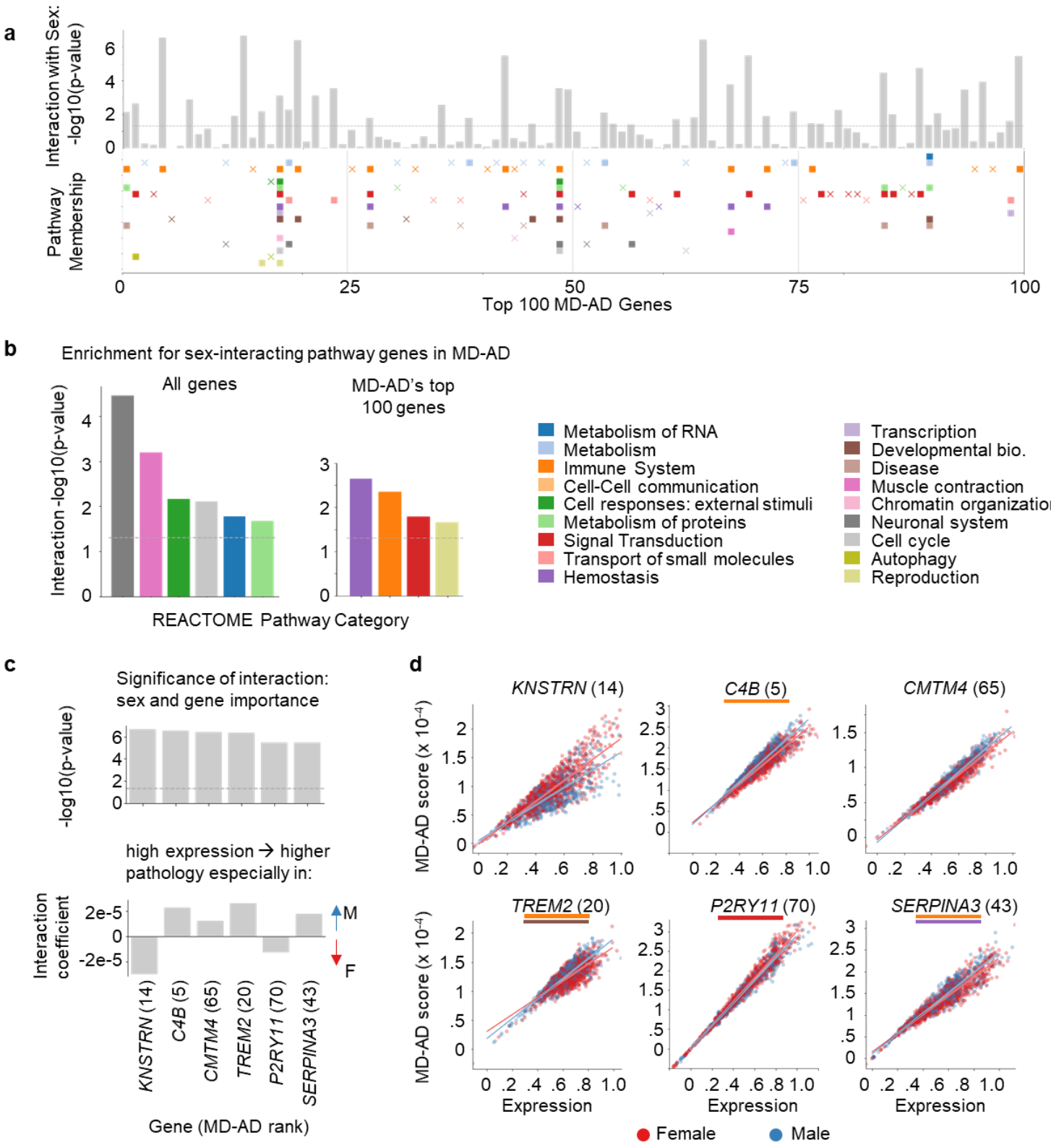

a
Interaction with Sex: -log10(p-value)
Pathway Membership
Top 100 MD-AD Genes
b
Enrichment for sex-interacting pathway genes in MD-AD
All genes
MD-AD's top 100 genes
Interaction -log10(p-value)
REACTOME Pathway Category
Metabolism of RNA
Metabolism
Immune System
Cell-Cell communication
Cell responses: external stimuli
Metabolism of proteins
Signal Transduction
Transport of small molecules
Hemostasis
Transcription
Developmental bio.
Disease
Muscle contraction
Chromatin organization
Neuronal system
Cell cycle
Autophagy
Reproduction
c
Significance of interaction: sex and gene importance
-log10(p-value)
high expression → higher pathology especially in:
Interaction coefficient
M
F
KNSTRN (14)
C4B (5)
CMTM4 (65)
TREM2 (20)
P2RY11 (70)
SERPINA3 (43)
Gene (MD-AD rank)
d
KNSTRN (14)
C4B (5)
CMTM4 (65)
TREM2 (20)
P2RY11 (70)
SERPINA3 (43)
MD-AD score (x 10-4)
Expression
Female
Male

Figure 2.5. MD-AD's top genes and their interactions with sex. **(a)** For the top 100 MD-AD genes, we compute the significance of the interaction between expression and sex for its MD-AD score. The bars indicate the gene's –log10(p-value) of the interaction term with sex (after FDR correction), and pathway categories each gene belongs to are indicated below. A filled square indicates that the gene significantly interacts with sex ($p<.05$ after FDR correction), and an "x" marker indicates that it does not. **(b)** For genes with significant sex interactions, we compute the significance of the overlap between REACTOME category genes and sex-differential genes among: Left: all genes, and Right: the top 100 MD-AD genes only. **(c)** For the top 100 MD-AD genes, we identify the genes with the most significant sex interaction for MD-AD scores. We show the significance of the interaction (Top) and the interaction coefficients (Bottom) for the top 6 most sex-differential genes. Each gene's MD-AD rank is indicated in their x-axis labels **(d)** For the top 6 most-sex differential top 100 MD-AD genes, we display scatter plots of expression by MD-AD score, coloring each sample by sex of the donor.

We next explored specific examples of genes with high MD-AD rankings and strong interactions with sex (i.e., the six genes from the top 100 MD-AD list with the strongest interaction *p*-values; Figure 2.5c-d): *KNSTRN, C4B, CMTM4, TREM2, P2RY11*, and *SERPINA3*. For each of these genes, we observed high expression values associated with higher neuropathology predictions but some stratification across sexes: high expression in females led to especially high neuropathology predictions for *KNSTRN* and *P2RY11*, while the opposite was true for the other four genes. Our finding that immune genes display sex-differential contributions to MD-AD scores appears to be consistent with conclusions from recent studies about sex differences in neuroinflammatory activity and the role these differences may play in neurodegenerative disorders[42].

We note that some of our top sex-interacting genes may play important roles in immune response, particularly in microglia. *TREM2*, which is genetically implicated in AD, interacts with *CD33* (another AD susceptibility gene)[43], is an important contributor in the clearance of toxic Amyloid-β by microglia in mice[44], and is correlated with Aβ deposition in the human brain[43]. Similarly, *KNSTRN* is known to be upregulated in mouse microglial cells' early response to neurodegeneration[45]. These findings indicate that MD-AD may capture patterns related to sex-differential microglia activity. To explore this idea further, we obtain lists of upregulated genes from nine clusters of single cell microglial transcriptomes[46], and compare them to our MD-AD gene rankings. As expected, many top MD-AD genes are upregulated in multiple microglial clusters (Figure 2.6a); correlation-based methods ranked these microglial genes less highly (Supplementary Figure 2.11d). Furthermore, genes upregulated in clusters related to stress, immune function and proliferation tended to be sex-differential in their gene importance (Figure 2.6b), further strengthening the finding that sex differences in immune response and inflammation may be an important factor in the molecular basis of age-related neuropathology.

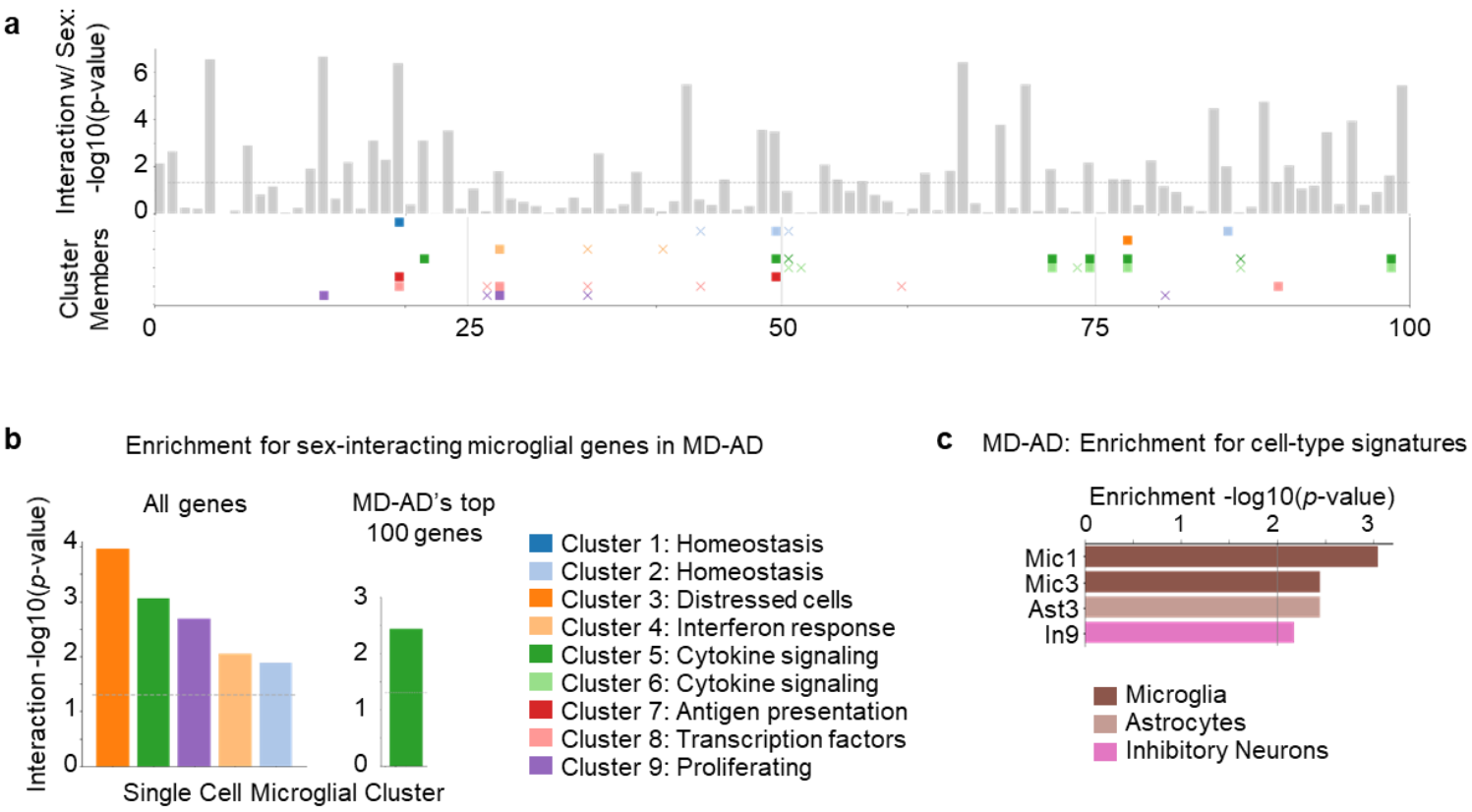


Figure 2.6. MD-AD's reliance on microglial cluster genes and gene set signatures. **(a)** Bars indicate the gene's –log10(p-value) of the interaction term with sex (after FDR correction), and gene membership in microglial cluster gene sets from Olah et al.[46] is indicated below. A filled square indicates that the gene significantly interacts with sex ($p<.05$ after FDR correction), and an "x" marker indicates that it does not. **(b)** For genes with significant sex interactions, we compute the significance of the overlap between microglial cluster genes and sex-differential genes among: Left: all genes, and Right: the top 100 MD-AD genes only. **(c)** Gene set enrichment -log10(*p*-value) across the final MD-AD gene ranking for cell type signatures[20].

To more broadly identify possible cell-type specific effects of MD-AD's important genes, we tested for the enrichment of 41 different cell type clusters (across six cell types) found by a single cell transcriptomic analysis of AD[20]. Here, we found an enrichment of 2 different microglia clusters, as well as astrocytes and inhibitory neuron clusters (Figure 2.6c). Hence, MD-AD's predictions of neuropathology rely on broader transcriptomic events beyond microglia genes, suggesting a heterogeneity in the underlying molecular biology that is predictive of accumulation of AD-related neuropathology.

Complex transcriptomic predictors learned by MD-AD are conserved across tissues.

Although MD-AD was developed for brain gene expression data, we next asked whether the learned transcriptomic signatures generalize to blood. To this end, we applied our brain-trained MD-AD model to gene expression datasets from two batches of the AddNeuroMed cohort, which we called Blood1 and Blood2 (n=711; NCBI GEO database accessions GSE63060 and GSE63061, respectively; [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63060 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE630601] summarized in Supplementary Table 2.5)[47,48]. As shown in Figure 2.7a, MD-AD predicted significantly higher neuropathology scores for individuals with both mild cognitive impairment (MCI) (two-sided t-test: t=7.34, p <.001) and AD dementia (two-sided t-test: t=5.87, p <.001) compared to cognitively normal controls (CTL). Consistent with external brain samples shown in Figure 2.2d and 2f, MD-AD predictions tended to increase with age for cognitively normal individuals, while they were consistently significantly higher for MCI and AD individuals compared to controls for individuals under 80 years old (Figure 2.7b, Supplementary Figure 2.12b). Importantly, we noted that a linear model failed to make meaningful predictions (Figure 2.7a and Supplementary Figure 2.12a), suggesting that complex models like MD-AD have better performance in extracting the true underlying signal transferrable between tissues than linear models.

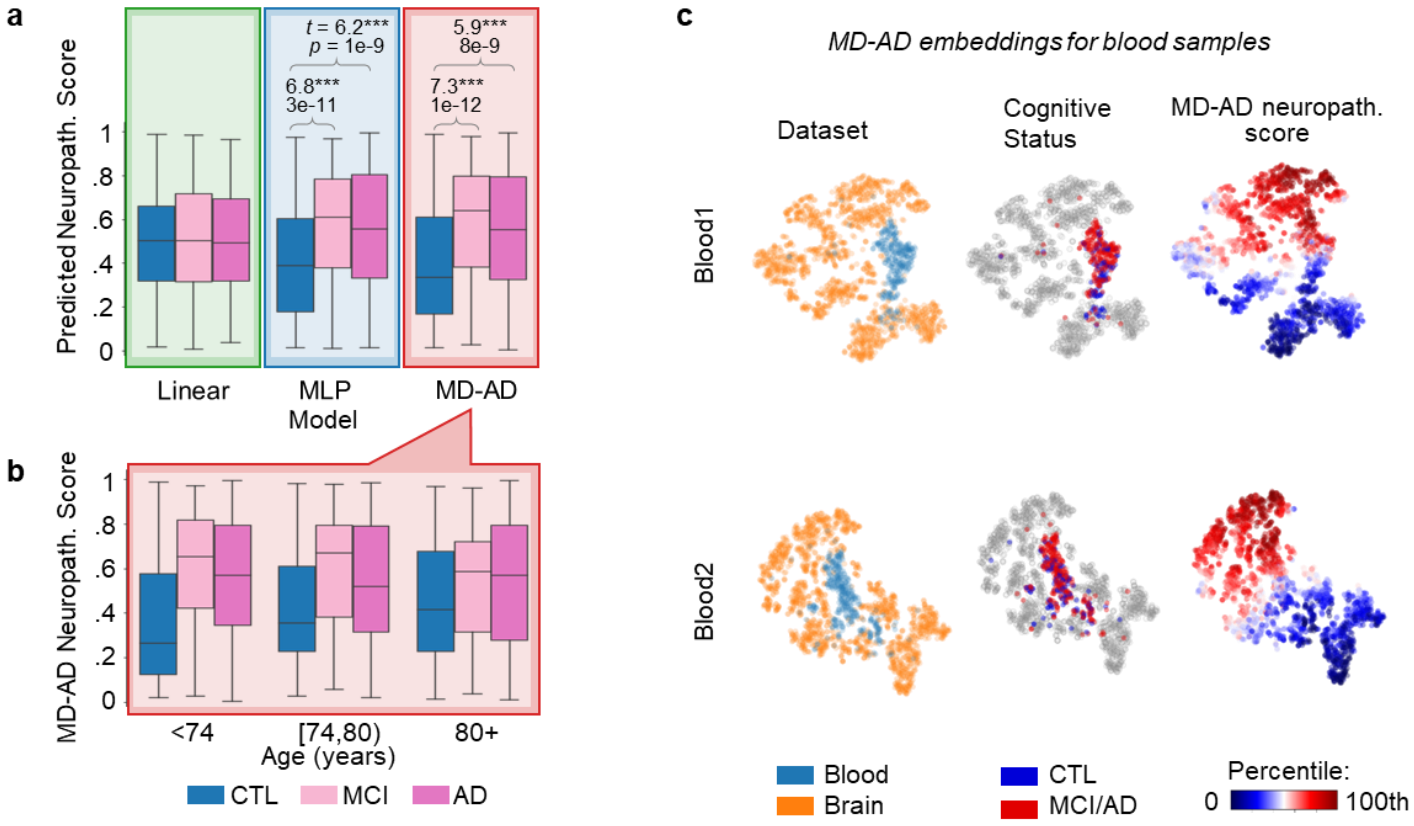


Figure 2.7. MD-AD's transfer performance for blood gene expression data sets. **(a)** Box plots show neuropathology predictions from each method split by cognitive status. From left to right, individuals who are cognitively normal (CTL), have mild cognitive impairment (MCI), and Alzheimer's dementia (AD). *T*-tests highlight significantly different groups within each method (2-sided t-test, ***p<.001). All box plots in this figure indicate median (center line), upper and lower quartiles (box limits), 1.5x interquartile range from quartiles (whiskers), and outliers (points). N=238 CTL, 189 MCI, 284 AD samples. **(b)** Box plots show the differences in MD-AD predicted neuropathology samples from individuals stratified by age group and cognitive status (Significant differences are shown in **Supplementary Figure 2.12b;** n=238 CTL, 189 MCI, 284 AD samples) (same box plot elements as described in part a). **(c)** t-SNE embedding of last shared layer from MD-AD models trained for Blood1 and Blood2 datasets. Samples are colored by their dataset (Left), cognitive status (while brain samples are shown in grey; Middle), and predicted neuropathology score (Right).

Next, we evaluated whether the patterns captured by the MD-AD model were consistent across training brain gene expression samples and blood. To this end, we again visualized MD-AD's learned embedding using the t-SNE algorithm (Figure 2.7c). We noted a clear difference in expression patterns between blood and brain samples (as seen by the clustering of blood samples in Figure 2.7c); however, MD-AD nevertheless produced an embedding for blood data that stratified blood samples along predicted neuropathological phenotypes in a manner highly consistent with the blood donor's cognitive status (Figure 2.7c; Supplementary Figure 2.12c). Together, these analyses indicate that jointly learning the relationship among brain gene expression and several neuropathological phenotypes may allow for learned representations that span tissues. This in turn can open avenues for early identification of individuals at risk, and provide clues into tissue-agnostic molecular mechanisms underlying AD dementia.

Discussion

We introduce MD-AD, a deep neural network approach for jointly modeling the relationship between brain gene expression data and multiple sparsely labeled neuropathological phenotypes in a multi-cohort setting. By exploiting the synergy between deep learning and a multi-cohort, multi-task setting, we demonstrated that MD-AD can capture complex, non-linear feature representations that are not learned using conventional expression data analysis methods. Specifically, we observed that multi-task learning improves prediction performance over singly trained models. Adding data from different cohorts improves performance for various neuropathological phenotypes, even those that lacked labels. When we extended our method to other datasets, it captured AD-related biological signals, showing that MD-AD can transfer effectively to out-of-cohort, out-of-species (mouse), and even out-of-tissue (blood) datasets.

As a neural network framework, MD-AD's last shared layer embedding reveals high-level features of gene expression that are predictive of neuropathology according to the intermediate components of the

model. As expected, due to multi-task supervision, our embedding nodes tend to relate to AD-associated neuropathology far more effectively than do standard unsupervised approaches and earlier reported (unsupervised) module-based approaches. Compared to singly task-supervised neural networks, MD-AD's joint training consistently provided a more stable and coherent AD-related embedding. By exploring the molecular pathways relevant to each node, we identified relevant gene sets contributing to these high-level AD-related features of gene expression.

Finally, we leveraged the complex relationships learned by MD-AD to refine our understanding of the molecular drivers of AD neuropathology. By interpreting genes relevant to our model's predictions, we uncovered that MD-AD relied on many genes not found in earlier linear-based methods, including several immune genes. These findings expand the general narrative established by human genetic studies of AD and now a proteomic study of AD[49]; in particular, we see enrichment for complement pathway genes (Figure 2.4) which likely connect with the role of the complement receptor 1 (*CR1*) gene which harbors an AD susceptibility variant whose functional consequences remain poorly understood but do include an influence on the accumulation of neuritic plaque pathology[50–53]. Thus, MD-AD results converge with human genetic results to emphasize the role of complement in AD; interestingly complement protein C4B emerges as one of the top pathology-related genes that display a strong interaction with sex, with men showing a much stronger association than women (Figure 2.5c). This is similar to the behavior of *TREM2*, another well-validated AD susceptibility gene (Figure 2.5c); however, its relation to amyloid pathology in ROSMAP data was previously reported as being modest[43]. MD-AD was able to uncover its more prominent role in transcriptional data, which is obscured by its sex-dependent nature. Likewise, women reported to have higher expression of a signature of aged microglia in these data[42], and two modules of co-expressed cortical genes enriched for microglial genes and associated with amyloid (module m114) or tau (module m5) pathology are also influenced by sex[54]. However, the role of neither group of genes is explained by sex; this indicates that the role of sex in the impact of the immune system in AD is complex. MD-AD was able to uncover this complexity more effectively, as is illustrated in Figure 2.5c where some genes have greater effects in men and others in women. Thus, it is not the case that role of the immune system is polarized in one of the two sexes; rather, some pathways and perhaps certain cell subsets may have a larger role in women while others are dysfunctional in men. This could explain why the role of immune genes is more prominent in our analyses: reports from simpler linear models often included immune pathways[8,19,55,56] but other pathways usually figured more prominently in these earlier RNA-based network models. A meta-analysis of RNA studies (which include the ROSMAP data) highlighted the larger number of sex-influenced genes among the AD-associated gene modules and noted that microglial cells appear to be enriched for both male and female-specific expression effects. We note that like any other machine learning-based model

applied to observational data, we are unable to directly infer causality in our framework, as both gene expression and neuropathology data used for MD-AD were collected from post-mortem brains. Nevertheless, with our list of results and our careful evaluation of sex effects we now have an important road map with which to guide our exploration of the role of microglia in AD in a sex-informed manner. This perspective will be critical not only for mechanistic studies whose results could be obscured by sex effects but also, more importantly, by guiding the study design of clinical trials as highly targeted therapeutic agents emerge to modulate the immune system in AD.

This is but one of the narratives that has emerged from our initial deployment of the MD-AD approach in the aging brain. As new cohorts are characterized, sample sizes expand and new data such as single nucleus RNA sequencing profiles emerge, our approach will help to facilitate data integration and to uncover insights that would not otherwise emerge. Beyond enabling good predictions, our report may actually highlight a more important contribution of MD-AD in resolving key elements of the data structure in the nodes that we defined: these are more than simple aggregates of factors with predictive power. They are beginning to uncover complex interactions, such as the impact of sex which is involved in both men and women, but in different ways, making it difficult to appreciate the role of certain immune pathways in simpler statistical models. Beyond producing accurate and generalizable neuropathology predictions and improving our biological understanding of AD pathogenesis, MD-AD provides a framework for integrating and analyzing gene expression data from separate cohorts and identifies common underlying relationships among gene expression and phenotypes of interest, which may be expanded as new data sets emerge.

Methods

Data processing

For developing the MD-AD model, we used data from the following RNA-Seq and neuropathology datasets available through the AMP-AD Knowledge Portal: (1) Adult Changes in Thought (ACT) [9], (2) Mount Sinai Brain Bank (MSBB) [10], and (3) Religious Orders Study/Memory and Aging Project (ROSMAP)[6–8]. Details of sample collection and sequencing methods are described in previously published work [6–10]. We pooled together brain gene expression data from the temporal cortex, parietal cortex, hippocampus, and forebrain white matter from ACT, Brodmann areas 10, 22, 36, and 40 from MSBB, and the dorsolateral prefrontal cortex from ROSMAP. To avoid confounding conditions, we excluded samples from individuals

who had neuropathological diagnoses other than AD. Taken together, the studies provide 1,758 gene expression samples.

In all three studies, extensive quality control measures were taken during the original processing of the datasets, as described by the original papers introducing transcriptomic datasets for ACT[9], MSBB[10], and ROSMAP[8] cohorts. All samples which passed quality control checks in these individual studies were included in our study. If Ensemble gene IDs were provided, we mapped them to gene symbols to keep consistent gene identifiers across datasets. In order to compile gene expression samples across the three cohorts, we retain expression levels for genes which are present in all datasets. Within each dataset, we exclude genes with null values for over two-thirds of samples. For ACT and ROSMAP, gene expression measurements were provided in normalized FPKM units, so we log-transformed the RNA-Seq datasets to obtain gene expression datasets that were roughly normally distributed (whereas MSBB gene expression data were already normalized). Then, for all datasets, we normalized values such that each gene's expression measures varied between 0 and 1. We then combined the gene expression datasets and kept all 14,591 genes that are present across all three datasets. Of these genes, 96.3% are autosomal (3.5% on only the X chromosome, 0.1% on only the Y chromosome, and 0.1% on both X and Y chromosomes). Finally, we performed batch effect correction with ComBat to reduce systematic differences across studies (Supplementary Figure 2.1c)[31].

Next, for each gene expression sample, we incorporated the available corresponding neuropathology labels: (1) **Aβ IHC**: amyloid-β protein density via immunohistochemistry, (2) **plaques:** neuritic amyloid plaque counts from stained slides, and (3) **CERAD score:** a semi-quantitative measure of neuritic plaque severity[57], (4) **τ IHC:** abnormally phosphorylated τ protein density via immunohistochemistry, (5) **tangles:** neurofibrillary tangle counts from silver stained slides, and (6) **Braak stage:** a semi-quantitative measure of neurofibrillary tangle pathology [33]. Detailed descriptions for each neuropathological phenotype within each dataset are provided in Supplementary Table 2.1. Because Braak stage and CERAD score are global measurements of neuropathological damage, if an individual had multiple available gene expression measurements from different regions, they each sample was labeled with the same Braak and CERAD values. However, Aβ-IHC and τ-IHC were provided for several brain regions for both ROSMAP and ACT studies. Therefore, each expression sample was labeled with the Aβ-IHC and τ-IHC measurements for the same or nearest region. Because the available plaques label provided by MSBB was averaged over several brain regions, we similarly used ROSMAP's average plaques and tangles labels (aggregated from several regions) for consistency with MSBB's metrics. We provide demographic and neuropathology information about individuals in each cohort in Supplementary Table 2.2.

Finally, for consistency across datasets, we first normalized all neuropathological variables to vary between 0 and 1 before combing datasets.

Computational methods

Review of previous approaches. Post-mortem transcriptomic studies have investigated molecular and neuropathological outcomes in AD. Early work in this domain examined simple correlations among gene expression and AD symptoms[22] or compared gene expression levels across AD-patients versus controls[23]. More recently, more systematic network-based analyses have contributed to the understanding of AD biology. In particular, Zhang et al. [19] constructed molecular networks based on bulk gene expression data separately for individuals with and without AD, and identified modules with remodeling effects in the AD network. More recently, Mostafavi et al. [8] used co-expressed genes in the aging human frontal cortex to build a single molecular network and identified modules related to AD neuropathological and cognitive endophenotypes. Using single-cell RNA sequencing data, Mathys et al. [20] clustered cells within brain cell-types to identify and characterize AD-related cellular sub-populations. Each of these approaches have been applied to single cohorts. Until recently, a unified and robust modeling of AD neuropathology based on brain gene expression has been hindered by relative scarcity and regional heterogeneity of brain gene expression datasets. One possible solution is to combine multiple data sets to gain statistical power. The collection of postmortem brain RNA-sequencing datasets, assembled by the AMP-AD (**A**ccelerating **M**edicines **P**artnership **A**lzheimer's **D**isease) consortium, provides new opportunities to combine multiple data sets. However, such heterogeneous datasets pose challenges to many methods, which must account for inter-study differences. In a recent attempt, Logsdon et al.[21] used a meta-analysis approach to identify co-expressed modules separately for 7 brain regions across 3 datasets, then subsequently applied consensus methods to identify modules that were conserved across multiple regions and studies. As of now, we're not aware of any methods that directly model all data in a unified way.

Modeling assumptions. For modeling gene expression and neuropathology data, we make several assumptions. *Common assumptions of machine learning models.* First, regarding stability of our models, we assume that a single training of each model is representative of all training instances. In order to buffer the potential failure of the assumption, we used cross-validation to select hyperparameters across different splits of our data across MD-AD, singly trained MLPs and linear models. Further, our final model is trained 100 times to generate ensemble predictions and interpretations, as described in future sections. Second, we assume that our samples are "sufficiently" independently and identically distributed (i.i.d.) such that a model trained on these samples should generalize well to new samples from the population of interest.

Third, we assume that the true data distribution is smooth such that samples with very similar gene expression values should display similar neuropathology.

Additional assumptions of MD-AD: Deep learning relies on the assumption that the data is generated by a composition of (learnable) features in a hierarchical manner. This allows neural networks with multiple layers to collapse correlation patterns in the input space to generate intermediate embeddings in a way that is useful for prediction[26]. Unlike linear models, our deep learning framework does not assume that there's a linear relationship between the predictors and outcomes, nor does it require normally distributed predictors, or low multicollinearity. Although deep learning relies on relatively few assumptions, in practice, some of these assumptions do not fully hold. In particular, our samples are certainly not i.i.d,, as some samples are derived from the same brain. Thus, external validation is invaluable for evaluating the effectiveness of our framework in new settings with no information leakage. The observation that MD-AD transfers well to separate datasets (and even species and tissues) implies that our framework is effective regardless of whether these assumptions were fully upheld. Finally, multitask modeling frameworks hinge on the assumption that there is shared common information across neuropathological phenotypes[25]. In combining multiple datasets with different sparsity patterns, we additionally assume that this common representation is consistent across cohorts and is generalizable to new data sets. This assumption appears to hold, as demonstrated by the improved test performance of the multi-task network over singly-trained MLPs, and improved ability to generalize well to external datasets.

The MD-AD Model. MD-AD (**M**ulti-task **D**eep learning for **A**lzheimer's **D**isease neuropathology), is a *unified framework for analyzing heterogeneous AD datasets* to improve our understanding of expression basis for AD neuropathology (Figure 2.1). Unlike previous approaches, MD-AD learns a single neural network by jointly modeling multiple neuropathological measures of AD severity phenotypes, and hence can incorporate data collected from multiple datasets. This *unified* framework has key advantages over separately trained models. First, MD-AD allows sparsely labeled data, which is a natural characteristic of datasets aggregated through consortium efforts (Figure 2.1e). Even if different phenotypes only partially overlap in the measured samples, each sample contributes to the training of both phenotype-specific and shared layers. Predicting multiple phenotypes at once biases shared network layers to capture relevant features of these AD phenotypes at the same time. This is of critical importance: each phenotype represents a *different type* of noisy measurement of the same underlying true biological process, and as we demonstrate by joint training MD-AD is able to average out the noise to extract the true hidden signal. Additionally, the increased sample size enables MD-AD to capture complex non-linear interactions between genes and phenotypes. In contrast, Multi-layer perceptrons (MLPs) offer another powerful approach for directly capturing complex relations between gene expression and a neuropathological phenotype. However,

training separate MLPs for each phenotype (Supplementary Figure 2.1a) has limited scope: it can utilize only the samples measured for a specific phenotype, and it cannot share information across related phenotypes. We demonstrate that these advantages improve MD-AD prediction accuracy, enabling it predictions to generalize across species and tissue types (Figure 2.1b). As illustrated in Figure 2.1a, the MD-AD network jointly predicts six neuropathological phenotypes from gene expression input data via shared hidden layers followed by task-specific hidden layers.

Training & evaluating MD-AD

Pre-processing with PCA. In order to have efficient and robust training and to reduce overfitting, we apply a principal component analysis (PCA) transformation to the data and use resulting top 500 principal components – a 500-dimensional representation of our 14,591 gene expression values – as the input to the MD-AD and all baseline models. This approach is consistent with the use of PCA for pre-processing in other studies that have employed deep learning in gene expression analyses[58–60]. Our choice to use 500 PCs is supported by some preliminary analyses of AD-related signals captured by various PCs. First, as shown in Supplementary Figure 2.1b, the cumulative variance explained by 500 PCs is about 92%, indicating that reducing our input features by a factor of about 30 still retains most of the variation in the data. However, we note that by using only 500 PCs, we may lose some information which may be especially predictive of AD neuropathology. To investigate this potential issue, we sought to predict average neuropathology scores from different sets of PCs using a linear model, and compare predictive performance to the full set of genes. We use cross-validation to tune the alpha parameter and, based on the same training/testing splits used in our main analyses (described below), find that by the time we include up to 500 PCs, we have reached similar predictive performance between a model trained on PCs versus the raw gene expression features (Supplementary Figure 2.1d). This suggests that the linear transformation provided by the first 500 PCs retains features of gene expression data that are almost as linearly predictive of AD neuropathology as the full dataset using all genes.

Construction of Models. For comparison to MD-AD, we generate six analogous MLP networks with un-shared representations, and six linear models containing no hidden layers, to serve as baseline models (see Supplementary Figure 2.1a). All models were built using Tensorflow and Keras packages, and were constructed as consistently as possible, with the same inputs. The MLP baseline model was identical to the MD-AD model except with only a single branch of task specific layers. Similarly, the linear baseline models were identical to the MLP baselines but with all hidden layers removed. Each model was trained on a single Nvidia GeForce GTX 980 Ti GPU. Although training time may vary across machines, we found that training the MD-AD model on the full dataset took about 350 seconds on average.

Internal test-set validation and tuning

Hyperparameters. After some preliminary experiments with single and multi-task neural networks, we decided to train all networks with ReLU activations and drop-out units (with drop-out rates of 0.1), and trained each model for 200 epochs with batch sizes of 20 using adam optimization. These settings were selected because they led to relatively stable and effective predictions. We tended to see some variation in performance based on kernel regularization, and hyperparameters of the optimization method, so for hyperparameter tuning (described below), we performed grid search over the following hyperparameters: kernel regularization parameter (1e-3 vs 1e-5), gradient clip norm (0.1 vs 0.01) for the adam optimizer, and the learning rate (1e-3 vs 1e-4).

Cross-validation and model tuning. For our model training and evaluation, we use a modified cross-validation and testing scheme as illustrated in Supplementary Figure 2.1e, in which we perform five separate rounds of model tuning with cross-validation (CV) followed by evaluation in a test set. For a single round, one-fifth of all samples are assigned to a held-out test set. Then using the remaining 4/5ths of the samples, we perform 5-fold cross-validation to select hyperparameters with the best prediction performance. We then train the selected model using the full training set (4/5ths of the original data) and then report performance on the held-out test set. In order to evaluate the robustness of our evaluation metrics under different splits, we initially split the full dataset into 5 separate groups, and repeated the above process 5 total times, where each one-fifth of the data acted as a held-out test set once. We note that across these iterations, different training sets selected different configurations of hyperparameters, and for each train/test round we trained the full training set on the specific configuration selected by cross-validation in that training set. Thus, our test set evaluations (e.g., in Figure 2.2a) reflect average test performance for the selected models in each round.

For MD-AD, we additionally explored several alternative options for architectures with different amounts of shared and task-specific layers (Supplementary Figure 2.2b-c). We selected the final architecture (shown in Figure 2.1a) because we wanted to have multiple hidden lasyers in both the shared portion and task-specific portion of the network to allow for non-linear interactions to be learned in both the shared representation and in the task-specific branches. However, when we evaluated alternatives to this approach (using the same selected hyperparameters for our original MD-AD model), we found that alternatives to this approach tended to perform similarly or worse (Supplementary Figure 2.2b-c).

Evaluation metrics. As described above, for each round of train/test splits, we use five-fold cross-validation to make modeling choices for the MD-AD model and baselines before training each model with

the full training set and reporting and reporting test $1\text{-}R^2_{CV}$ error (mean squared error divided by the phenotype's variance in the validation set; averaged over all five test splits). We evaluate model performance in two ways: (**1**) standard train and test sets, and (**2**) ROSMAP test performance for different subsets of the available datasets.

First, separately for each of our five cross validation training sets, we calculate the final test MSE on the corresponding hold-out set. To test whether these effects are significant, for each baseline method, we performed one-sided paired t-tests to determine whether there is a significant difference between the baseline method's error and MD-AD's across the five test folds (Figure 2.2a).

Next, in order to evaluate the contributions of each dataset to prediction performance, we performed the above procedure with different subsets of available datasets. Because ROSMAP is the only dataset with all available neuropathological phenotypes, we evaluate performance specifically on ROSMAP. In Figure 2.2b, we show ROSMAP test samples' MSE performance when trained on all subsets of ACT, MSBB, and ROSMAP training samples (following the same cross-validation procedure described above). We additionally repeated the same analysis using MSBB and ACT test samples and computed their prediction performance for available phenotypes (Supplementary Figure 2.3a). In order to evaluate how transfer performance (i.e., training and evaluating with samples from disjoint cohorts) was impacted by the addition of samples, we performed an additional analysis where we trained with ACT samples and varying fractions of MSBB samples to see how additional MSBB samples impacted ROSMAP test performance (and also evaluated the reverse, training on ROSMAP and MSBB and testing with ACT samples) (Supplementary Figure 2.3b). Interestingly, we saw that ROSMAP test performance improves with the addition of MSBB samples during training with ACT, whereas ACT samples generally do not improve with the addition of MSBB samples during training with ROSMAP. This may imply that there are more pronounced distributional differences for the ACT cohort when compared with other cohorts, or that improvements are more apparent when the training set is much smaller (as is the case when training with only ACT samples).

Final model selection. Finally, after our in-depth cross-validation and testing scheme was used to evaluate our methods internally, we constructed "final models" for external validation and model interpretation. First, we selected a single set of hyperparameters for each model by ranking each configuration's prediction performance for each round, and then choosing the configuration with the highest average rank. The selected hyperparameters for each "final model" is provided in Supplementary Table 2.6. We trained "final models" for MD-AD and baselines by each using a single set of hyperparameters on the full dataset.

Evaluating models with covariate-corrected data. Gene expression-related covariates may influence gene expressions in a systematic way, and thus should be critically considered. Indeed, there does seem to be a small but significant correlation between neuropathology scores and both postmortem interval (PMI; r = -0.16, p=1e-9) and RNA integrity number (RIN; r = -0.09, p=0.002), which are both features which may influence measured gene expression. In our study, we chose to leave our expression profiles uncorrected for all covariates, and instead allow MD-AD to learn from the available gene expression patterns so that we can subsequently assess how these covariates interact within our final models.

Although Supplementary Figure 2.6c shows that PMI and RIN had modest residual correlations with nodes in the consensus MD-AD network, and thus likely do not appear to be driving forces in our model, we performed an additional analysis to ensure that gene expression-related covariates were not an important factor in our prediction performance results presented via our cross-validation evaluations. To that end, we use the following method to correct our gene expression data for sequencing-related covariates: we linearly regressed PMI and RIN from our expression inputs by modeling the expression of each gene as a linear regression with PMI and RIN. We then saved the residuals of the predicted expression value as our corrected expression values. We then performed the same model training and evaluation procedures as described above using the corrected gene expression values as inputs, and found that these results were quite similar to our original results with uncorrected gene expression values (Supplementary Figure 2.13a-b). Together, these findings indicate that covariates related to gene expression measurement procedures do not seem to have a large impact on our final results, nor does the MD-AD heavily rely on these covariates, and for that reason, our main results are all based on gene expression data without covariate-correction.

Evaluating models with fully independent CV splits. All internal validation results were presented for the same cross-validation and testing splits. We note that in generating these original splits, we randomly assigned all samples within each cohort. However, because ACT and MSBB datasets provide multiple samples (collected from different brain region) from each individual, there are many individuals in our dataset with samples in both training and test splits. In order to ensure that performance improvements seen for MD-AD versus MLP or linear baselines were not due our splitting choice, we repeated our cross-validation experiments using a new method of splitting samples. Instead of splitting samples completely randomly as was done to generate our main internal test results (i.e., in Supplementary Figure 2.1e), we instead split *individuals* randomly for each dataset to ensure that no samples from the same individual could be split across training and validation sets.

We performed the same cross-validation and hyperparameter selection process as was done for our original dataset splits, and our resulting prediction performance and last shared layer evaluations are shown

in Supplementary Figure 2.13c-d. In these experiments, we find that MD-AD (as well as the baseline methods) provides very similar prediction performance when trained and evaluated on fully separated training and test sets, suggesting that our original results did not seem to hinge on the similarity of samples between the training and test set. We similarly find that MD-AD continues to produce embeddings which capture both neuropathology phenotypes and higher-level AD variables more consistently than alternative approaches. Finally, we note that they hyperparameters selected using the new splits are similar to the original final model selected from our original splits with the exception of a single hyperparameter (kernel regularization of .001 for our original splits, compared with .00001 for the new splits). However, we find that in our analyses with the new splits, a model trained on our originally selected set of hyperparameters has very similar performance to the newly selected set of hyperparameters. Together, these results indicate that our choice to split samples randomly produced very findings to the alternative of splitting samples pseudo-randomly by individual.

External dataset validation (Human)

In order to evaluate MD-AD's ability to generalize to out of sample data, we assessed performance on three datasets: Mount Sinai Brain Bank Microarray (MSBB-M; N=1,047), Harvard Brain Tissue Resource Center (HBTRC; N=338), and Mayo Clinic Brain Bank (N=157). These datasets were collected from AMP-AD (with the exception of HBTRC which was collected from GEO: GSE44772), but were left out of the original MD-AD training because they were microarray samples or lacked many neuropathology labels.

After normalizing gene expression samples from external data sets in the same way as described for the ACT, MSBB RNA Seq, and ROSMAP datasets, we then adjust the expression values to have similar distributions to our batch corrected training data sets. We evaluated the MD-AD model on our new processed data to obtain predictions for all six phenotypes. Because these three external datasets provide a sparse set of neuropathological labels, we do not have access to labels for many of the six MD-AD labels. Instead, we evaluated whether MD-AD's predictions were consistent with the (binary) neuropathological diagnosis of AD, by aggregating MD-AD's various neuropathology predictions into one "neuropathology score". The "neuropathology score" was produced by first calculating percentiles across samples (within each dataset) for each neuropathological phenotype, then averaging over the six phenotypes.

Figure 2.2c shows that MD-AD provides the largest differences in neuropathology scores between individuals with and without neuropathological diagnoses of AD. We further compared neuropathology scores between AD and non-AD individuals split by age group (significance between groups shown in Supplementary Figure 2.4b).

A similar analysis was carried out comparing carriers of the APOE ε4 allele with non-carriers (instead of AD vs control individuals). Results shown in Supplementary Figure 2.5a-c revealed similar patterns, including improved discrimination between groups for MD-AD compared with MLP and linear baselines, and more pronounced differences in predicted neuropathology for younger APOE ε4 carriers versus non-carriers. However, when comparing predicted neuropathology between APOE ε4 carriers versus non-carriers within the same cognitive diagnosis, we do not see a difference in predicted neuropathology for AD-afflicted APOE ε4 carriers and non-carriers (Supplementary Figure 2.5d-e).

Cross-species validation (Mouse)

To evaluate how well expression patterns predictive of neuropathology learned by MD-AD recapitulates neuropathology in mouse models. To that end, we obtained gene expression data from Matarin et al. [35] for 15 TASTPM mice which harbor double transgenic mutation in APP and PSEN1, as well as 37 wild type mice. For each mouse, brain gene expression was measured from two samples collected from the cortex and hippocampus, doubling the total sample size. Data were quantile-normalized and log transformed. For this experiment, we mapped mouse to human genes (via gene symbols) for a total of 7,057 intersecting genes between our training dataset and the mouse expression data, which were again normalized to follow the same distributions as our MD-AD training data. We retrained our MD-AD model on only these 7057 genes for all MD-AD samples and then generated "neuropathology scores" for the mouse samples exactly as described in the previous section. As with external validation experiments described above, we compare MD-AD to MLPs and linear models in separating neuropathology scores between TASTPM and wild type mice (Figure 2.2e). We also show differences in neuropathology scores between different age groups (Figure 2.2d, Supplementary Figure 2.4c).

Supervised embedding validation

The output of an intermediate layer of a neural network can be viewed as lower dimensional embedding of the input features. In this paper, we focus on the last shared layer of the MD-AD network because it is a supervised embedding of gene expression data which is influenced by all six training phenotypes. We evaluate the embedding compared with those generated by both singly-trained MLPs as well as unsupervised methods (i.e., K-Means and principal components analysis (PCA)) in two ways: (1) high level visualization with t-SNE, and (2) evaluating the correspondence between individual nodes and AD-related features.

Visualizations with t-SNE: For each of the MD-AD, MLP, and unsupervised models, we train the models on the full combined dataset. For the deep learning models, we then generate "supervised"

embeddings by obtaining the output of the last shared layer (or analogous layer of the MLP model). For the unsupervised methods, K-Means and PCA, we generate an embedding of 100 dimensions to be consistent with the MD-AD and MLP models. After generating these embeddings for all samples, we then compress them to 2 dimensions via the t-SNE algorithm [37]. T-SNE Visualizations of MD-AD's supervised embedding are shown in Figure 2.3a (left side for each phenotype), and the figure is replicated with six times, with each plot showing samples colored by neuropathological phenotype severity for each of the six phenotypes. For comparison, t-SNE visualizations for the singly-trained MLPs and unsupervised methods are shown in Figure 2.3c (colored by CERAD Score only) and colored by other characteristics and covariates of interest in Supplementary Figure 2.7.

Node-phenotype correlations: To test whether MD-AD's embedding generalizes more to AD phenotypes than the alternative methods, we compare the nodes that best capture each phenotype among MD-AD, MLPs, and unsupervised methods. We perform the following analysis with the same five training and test splits described earlier: for each of the six phenotypes used in MD-AD's training, we identify the node in MD-AD's last shared layer whose output is most significantly correlated with that phenotype in the training set. We then report the –log10(*p*-value) (after FDR correction over nodes) for the correlation between that node's output and the training phenotype in the test set, averaged across the train/test splits. (Figure 2.3a, right side for each phenotype).

We also perform a similar analysis with higher-level AD phenotypes not used during model training: dementia diagnosis (binary variable available in all datasets), last available cognition score (controlling for age, sex, and education; only available for the ROSMAP dataset), and AD duration (i.e., time between dementia diagnosis and death; available for the ACT and ROSMAP datasets). For this analysis, we report the highest –log10(p-value) after FDR correction between nodes and the high-level phenotypes, average over the five test sets (Figure 2.3b).

Model interpretation

Integrated Gradients (IG). Although deep learning models have shown promise in biological and health applications, they've been limited by the difficulty of explaining their predictions. Fortunately, the development of interpretability methods for "black box" models such as deep neural networks have helped researchers derive understanding from complex models[61]. In particular, Integrated Gradients (IG) is a method for assigning sample-specific importance scores for inputs of a model on the output based on the gradients of neurons' weights across the network. As described in detail by Sundararajan et al. [30], the IG score calculated for a specific sample is generated for each input dimension on each output dimension by

accumulating gradients along the path from the input to output. Thus, applying IG to MD-AD allows us to achieve sample-specific gene importances for each neuropathological phenotype predicted by MD-AD. Additionally, by treating the last shared layer as the "output" of the MD-AD model (i.e., by temporarily removing all subsequent layers), IG is also able to identify gene-level importances for nodes in MD-AD's last shared layer. As described next, we use IG applied both to the phenotype predictions and last shared layer nodes to interpret MD-AD's learned representations.

Obtaining IG scores. We note that for each MD-AD model (of the 100 re-trainings), we apply the IG algorithm for each sample, which generates sample-specific IG scores for gene on each output. Thus, for each MD-AD model, we generate a (# samples × # genes × # output nodes) matrix providing sample-level gene importances for output nodes. Using the standard approach for a single MD-AD model, this provides sample-specific importances for each gene on each output phenotype. We additionally generate a modified MD-AD network with all layers beyond the last shared layer removed to obtain sample-specific IG scores for genes on all nodes in the last shared layer. Thus, for each of the 100 MD-AD models, we have a (# samples × # genes × # output phenotypes) matrix of gene importance for neuropathology predictions, as well as a (# samples × # genes × # last shared layer nodes) matrix of gene attributions for the last shared layer of the network. As described in the following section, we derive insights from the consensus MD-AD model by aggregating these IG values in various ways.

Aggregating gene importance scores for nodes. For both the output nodes (6 neuropathological phenotype predictions) and last shared layer nodes, we have (# samples × # genes × # nodes) IG matrices for each MD-AD run as described above. Now, we describe how we are able to aggregate across samples (and ultimately runs) to obtain a final gene ranking for each (output or last shared layer) node. First, for each MD-AD run, we generate a gene ranking for each of these nodes using a weighted average. Our weighted average uses the following weights: +1 for samples from individuals with high Braak and CERAD scores, -1 for samples from individuals with low Braak and CERAD scores, and 0 otherwise. Thus, the genes with the highest aggregated IG scores are those for which high IG scores coincide with high node outputs. This approach is used for both ranking genes' relevance to neuropathology in the MD-AD framework, and for annotating last shared layer nodes, as described in the next sections.

Constructing and annotating MD-AD consensus nodes

Because deep neural networks have non-convex loss functions, randomness in our training procedure produces networks with different weights from run to run. In order to capture robust nodes and highly relevant genes, we repeat our training procedure 100 times, in order to simulate a "consensus network". As

shown in Supplementary Figure 2.8a, we construct "MD-AD consensus nodes" by clustering nodes from many runs: (1) we train 100 MD-AD networks, (2) we obtain last shared layer node outputs for all samples and normalize them (0-mean, unit variance), (3) we combine all nodes across all runs and then cluster them using k-means (where the dimensions used to calculate similarity are samples) with k=50, (4) we summarize each cluster of nodes by their medoid. Thus, for each sample, the MD-AD consensus embedding is made up of 50 nodes which are medoids of clusters generated from 100 re-trainings.

In Supplementary Figure 2.6b, we provide a visual overview of the MD-AD consensus embedding generated as described above. To provide a simple view of clusters, we select a subset of samples for which we have clear high or low pathology, excluding ambiguous cases. We include (1) individuals with Braak stage of at least 5 and CERAD scores at least 3 (i.e., "moderate"), or (2) individuals with Braak stage of 3 or lower and a CERAD score of 1 (i.e., "absent") who are at least 85 years old and have no dementia. Case 1 captures all individuals with pathologic AD diagnoses (with and without dementia), whereas case 2 captures all individuals considered "resistant" to AD due to their old age but lack of cognitive or neurological decline (consistent with previous literature, e.g. Latimer et al. (2019)). To annotate each node in the consensus embedding, we display their correlations with various phenotypes and covariates, as well as their enrichment for REACTOME pathways.

Correlations: For each variable (neuropathological phenotypes, high-level AD phenotypes, and covariates), we compute the correlation –log10(p-value) between the variable and each consensus node output. In Supplementary Figure 2.6c, a high –log10(p-value) indicates that a node captures (or is highly linearly related to) a variable.

Pathway enrichment: Beyond relationships between nodes and various phenotypes, we annotated nodes with which gene sets are relevant to their outputs. First, in order to identify relevant genes to each consensus node, we use IG scores. As described in Section A, for each run, we aggregate IG scores across samples to obtain a weighted average of gene importance scores for each last shared layer node. Because our consensus last shared layer nodes are actually individual nodes sourced from various runs of MD-AD, we simply combine the aggregated IG scores from the relevant nodes across these runs. For each MD-AD consensus node, this method therefore provides us with a ranking over all genes by their importance. We then test for enrichment of REACTOME pathways [63] in these gene rankings via gene set enrichment analysis (GSEA) [38,39] to identify whether certain pathways seem to be involved in the activation these nodes. Enriched pathways for the MD-AD consensus nodes are shown in Supplementary Figure 2.6d. Supplementary Table 2.3 provides detailed annotations for each node.

Identifying MD-AD's top genes

In order to identify genes that drive MD-AD predictions, we used integrated gradients (IG) [30] to provide importance estimates of each gene on the predicted outcomes. In order to improve model stability, we calculate gene rankings based on 100 re-trainings. As described in Section A, after each run of training, we take our trained model and apply IG for each sample to get the importance of each gene on each neuropathological phenotype prediction. We next aggregate our IG scores into gene rankings by calculating the ranks of each gene (for each phenotype) in each run, and then averaging across runs to obtain consensus gene ranks. For each phenotype (see Supplementary Figure 2.8 for illustration). Thus, the gene with the highest consensus IG score (i.e., score close to 1) is the gene with the highest average rank across runs (most positively associated with the neuropathological phenotype), and the gene with the lowest consensus IG score (i.e., score close to 0) is the gene with the lowest average rank across runs (most negatively associated with neuropathology). While we generate these consensus rankings separately for each phenotype, we again average across the six phenotypes to obtain our final MD-AD consensus IG scores. We note that 100 re-trainings are more than enough to converge to a stable gene ranking (Supplementary Figure 2.8c). The top genes for MD-AD are shown in Figure 2.4a, and enriched REACTOME pathways in the top ranked MD-AD genes (via GSEA) are shown in Figure 2.4b. The full gene ranking, generated separately for each neuropathological phenotype, is provided in Supplementary Table 2.4.

For comparison with a linear gene ranking method, we also generate correlation-based gene rankings as follows: we calculate the correlation coefficients between each gene's expression level and each neuropathological phenotype (across all samples in our dataset), and then percentile-rank the genes by their average correlation coefficients across all six phenotypes (with 0 for the most negatively correlated and 1 for the most positively correlated gene with high pathology). Our final correlation-based gene ranking is the average over the phenotype-specific rankings. Comparisons between REACTOME categories represented in the top MD-AD vs correlation-based rankings are shown in Figure 2.4c.

Calculating nonlinear effects for MD-AD genes

As a deep learning method MD-AD has the capacity to identify non-linear relationships among genes' expression levels and neuropathological phenotypes. These non-linear relationships may reveal an implicit capture of interaction effects with other covariates observable from expression data. Thus, we sought to investigate the presence of interactions between sample-level covariates and specific genes in their contributions to the MD-AD predictions.

Generating sample-level gene importance scores. To simplify our analyses, we generate consensus IG scores for each gene within each sample as follows: for each sample and gene, we average over the gene's IG weights across both neuropathological phenotypes and runs in order to obtain its average importance for general neuropathology across all runs.

Measuring interaction effects. To monitor the presence of interaction effects in gene importance scores, we modeled the consensus per-sample IG scores as a linear combination of a gene's expression level, a covariate of interest, and the interaction of the two. Specifically, $score_{g,i} = \boldsymbol{a}\, expr_{g,i} + \boldsymbol{b}\, feat_i + \boldsymbol{c}\, expr_{g,i}\, feat_i + \boldsymbol{d}$, where $score_{g,i}$ is the consensus IG value for gene g and sample i, $expr_{g,i}$ is the sample i's expression level for gene g, and $feat_i$ is sample i's value for the covariate. Based on this representation, we consider there to be an interaction effect between a gene and feature on its importance in the MD-AD model if the learned $\boldsymbol{c}$ coefficient is statistically significant ($p<.05$, after FDR correction over all genes). We primarily focus on identifying an interaction effects with sex ($feat_i = 1$ if sample i comes from a male), and rank interactions between genes and sex for MD-AD based on the –log10(p-value) of the interaction term.

Gene set enrichment: We evaluated whether sex-differential genes were enriched for the following gene sets: (1) REACTOME pathways[63] and (2) microglial cluster gene signatures from a recent single cell RNA Seq analysis of microglial cells from autopsied aging brains[64]. To evaluate whether the list of sex-differential MD-AD genes are enriched for gene sets of interest, we use Fisher's exact tests to evaluate the significance of overlap between all sex-differential genes and members of each gene set. Next, to evaluate whether the top MD-AD sex-differential genes are enriched for the same gene sets, we perform Fisher's exact tests again, but this time only consider the top 100 MD-AD genes in the calculations.

Blood gene expression validation

To evaluate the ability of MD-AD to transfer to blood gene expression data, we downloaded publically available AddNeuroMed cohort data from GEO (GSE63060 and GSE63061, which we refer to as Blood1 and Blood2, respectively). Details about the AddNeuroMed samples are provided in Supplementary Table 2.5**.** As with the other validation datasets, each blood dataset was normalized such that each gene's expression values have the same mean and variance as the processed MD-AD expression data. Because each blood dataset had a different set of available genes, for each dataset, we re-trained MD-AD consensus models for brain samples with only the genes available between them and blood samples (12,104 and 11,392 genes for Blood1 and Blood2 respectively). Because these blood samples came from living participants, we do not have access to the many neuropathology variables available across the brain samples. Instead,

we assess whether MD-AD's predictions align with individuals' cognitive diagnosis of cognitively normal (CTL), mild cognitive impairment (MCI), or dementia.

We evaluate the effectiveness of the MD-AD model by comparing predicted MD-AD pathology scores between CTL and MCI individuals, and between CTL individuals and individuals with dementia via two-sided t-tests (together, and split by age). To evaluate the MD-AD embedding for blood samples, separately for each blood dataset, we obtain the last shared layer embeddings of both the MD-AD brain expression samples and blood samples from the first round of training.

Chapter 3. An automatic integrative method for learning interpretable communities of biological pathways

Although knowledge of biological pathways is essential for interpreting results from computational biology studies, the growing number of pathway databases complicates efforts to efficiently perform pathway analysis due to high redundancies among pathways from different databases, and inconsistencies in how pathways are created and named. We introduce the PAthway Communities (PAC) framework, which reconciles pathways from different databases and reduces pathway redundancy by revealing informative groups with distinct biological functions. Uniquely applying the Louvain community detection algorithm to a network of 4,847 pathways from KEGG, REACTOME and Gene Ontology databases, we identify 35 distinct and automatically annotated communities of pathways and show that they are consistent with expert-curated pathway categories. Further, we demonstrate that our pathway community network can be queried with new gene sets to provide biological context in terms of related pathways and communities. Our approach, combined with an interpretable web tool we provide, will help computational biologists more efficiently contextualize and interpret their biological findings. [†]

Introduction

Many computational biology studies aim to identify groups of genes that are associated with or differentially expressed with respect to phenotypes of interest[11,12]. Performing pathway analyses to associate these gene-level findings with biological processes is a crucial step in providing both biological context for genes and a systems perspective to the analysis. Researchers have constructed *pathways* (i.e., networks of genes that interact in various ways to perform certain biological tasks[65]) using a variety of methods; these range from large-scale computational analysis over literature (e.g., Gene Ontology [66]) to hand-curation by experts (e.g., KEGG[67–69] and REACTOME[63]). These pathway gene sets are then commonly used in *pathway enrichment analysis*, where a new list of genes is compared with previously established pathway gene sets (via statistical tests to measure whether their genes overlap at higher than random rate) in order to identify which pathways are most related to the gene set of interest.

Since many pathway gene set databases offer unique advantages, it has become common practice to perform pathway enrichment analysis across multiple databases. Unfortunately, the difficulty of interpreting pathway-level results is increasing as the number of available pathways and databases grows.

[†] This paper was joint work with Ayse B. Dincer and Su-In Lee. It has been published in *NAR Genomics and Bioinformatics*[130].

Figure 3.1a shows that pathways across different databases and even within a database often have highly overlapping gene sets, causing redundancy in enrichment results. In fact, 50% of all pathways in Figure 3.1a have some corresponding pathway in a separate database, with at least 70% of genes in common. Therefore, a single query of genes can return multiple enriched pathways (e.g., KEGG's oxidative phosphorylation and Parkinson's disease pathways, which have 71% overlap), and it may be difficult to assess how the pathways are related, especially if they are from varying sources. Systematically analysing these phenomena can provide context for the simultaneous capture of multiple pathways.

Many studies have attempted to solve the problem of pathway redundancy. One common approach collapses pathways from multiple sources into a condensed set of high-level pathways – such as PathCards [70], MSigDB Hallmark [71] or GO slim [72] – to simplify enrichment tests; however, by condensing many pathways, these approaches tend to remove smaller nuanced pathways describing specific biological functions. Pathway network visualization tools, such as Enrichment Map, also highlight the relatedness of pathways [73] but do not map pathways to functional categories. Although standard clustering algorithms have been useful for identifying groups of related pathways [74,75], they are limited in that these approaches have not been systematically validated with respect to expert labels. Furthermore, previous clustering-based approaches have relied on manual annotation of clusters and thus do not scale well to the growing numbers of pathways and databases [73–75].

In this paper, we address the problems of heterogeneity and redundancy across literature-derived pathways by introducing the Pathway Communities (PAC) framework. Our framework uniquely relies on the Louvain community detection algorithm to cluster pathways into communities based on their gene-level similarities (Figures 3.1-3.2). Further, we enhance the biological interpretability of cross-database pathway analysis by 1) characterizing learned communities with respect to pre-defined categories (Figure 3.3), 2) devising a method to algorithmically annotate communities (Figure 3.4a), 3) applying interactive visualization techniques to investigate newly revealed connections within and across pathway communities (Figure 3.4), and 4) providing a tool to help researchers investigate novel gene sets in the context of our learned communities and member pathways, which we demonstrate on a breast cancer gene expression example.

Material and Methods

Pre-processing of pathways and generation of curated categories

For our analyses, we constructed a pathway graph comprised of pathways from KEGG [69], REACTOME [63], and Gene Ontology (GO) [66] gene sets, downloaded from MSigDB v7.0 [39,76]. Each pathway database

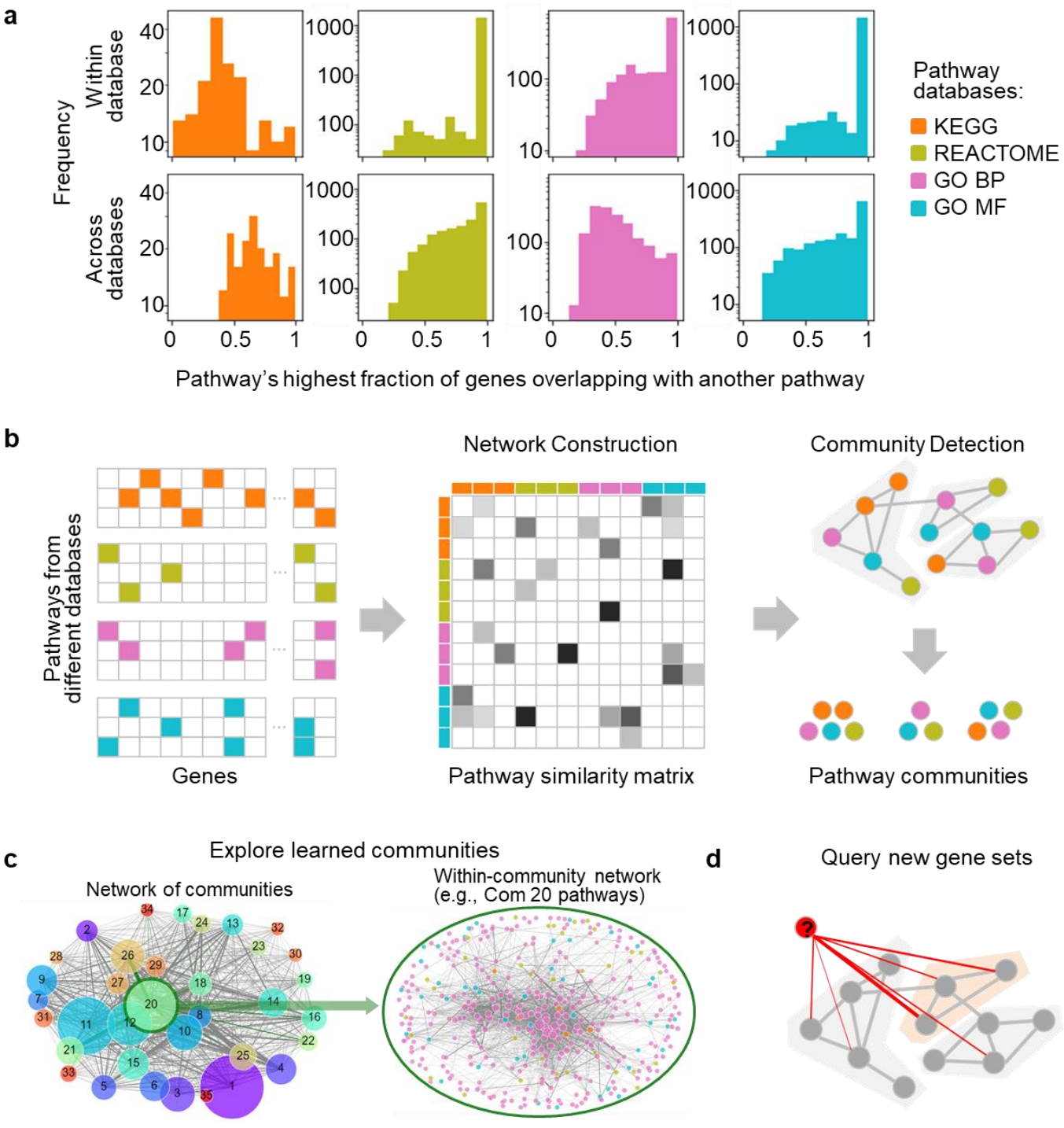


Figure 3.1. Overview of the pathway community approach. (A) The pathway overlap problem in pathway enrichment analyses. For each pathway database, we show the distribution of each pathway's maximum fraction of overlap with other pathways, both within the same database (top) and with other databases (bottom). (B) Our approach for learning pathway communities: we first construct a pathway network based on gene overlaps, and we then perform community detection to produce pathway communities. (C) Schematic highlighting of functionality on our webpage showing detailed views of each learned community. (D) Schematic highlighting of our proposed method for querying a new gene set against our learned communities to identify relevant processes enriched in a query gene set.

contains a list of named pathways that each have set of associated genes. For each database, we additionally pre-processed a provided set of curated categories or hierarchical relationships among pathways to identify higher-level categories associated with each pathway (see Supplementary Methods). These simple mappings from pathways to curated categories were treated as ground-truth labels for evaluating our method. Although these curated categories share some common themes across databases, it is not possible to combine them, so initial validation experiments were performed separately for each database. Finally, we downloaded MSigDB Hallmark gene sets [71], a collection of refined pathways meant to summarize thousands of founder gene sets (some of which are KEGG and REACTOME pathways in our analysis), which we use for additional validation.

Generating the pathway graph and learning communities

Using the PAC method for identifying communities of related pathways involves two steps: 1) construction of the pathway network, and 2) detection of pathway communities (Figure 3.1b). For the first step, we represented pathways from multiple databases as a large graph, where each node is a pathway (with an associated set of genes). For each pair of pathways, we performed Fisher's exact test [77] to evaluate the significance of their gene overlap (Supplementary Figure 3.1a). We added edges between all pairs of pathways, with edge weights corresponding to the –log10(*p*-value) from Fisher's exact test measuring the significance of gene overlap (*p*-values were Bonferroni corrected, and edges between pathways with $p>0.01$ were set to a weight of 0). This process generated a sparsely connected graph in which each node represents a pathway, and edges indicate similarity of pathways with respect to shared genes.

For the second step, we identified communities of related pathways from this network using the Louvain community detection algorithm [78,79] using the Community API in Python. To our knowledge, ours is the first work to use an approach based on graph modularity to cluster pathways. The Louvain algorithm learns well-connected communities from a network by greedily optimizing for graph modularity, namely, a measure of the density of intra- vs inter-community edges. During our evaluation, we first performed the PAC method separately for each pathway database (generating 4 separate pathway graphs and performing community detection separately). We ran the Louvain algorithm with a resolution parameter of 0.4 and conducted several experiments to verify the stability of our approach with different initializations, hyperparameters, and alternative methods for graph construction and evaluation (Supplementary Methods; Supplementary Figures 3.2-3.6). We also compared other graph clustering approaches with the Louvain algorithm: agglomerative clustering, spectral clustering, and the Clauset-Newman-Moore (CNM) algorithm

(another modularity-based approach); for all methods, we evaluated several hyperparameter options. Figure 3.2 reports the highest NMI achieved between the learned clusters and curated categories (i.e., the reduction uncertainty for the curated category label when the algorithmically learned community label is known).

Finally, after confirming the consistency of our learned communities with curated categories within each database (Figure 3.2), we learned an integrative set of communities from the pathway graph that combined all four pathway databases, resulting in 35 learned communities across 4,847 pathways. The full list of pathways' community assignments is provided in Supplementary Table 3.1. Our code for producing the pathway graph and learning communities is available at https://gitlab.cs.washington.edu/nbbwang/P-COM.

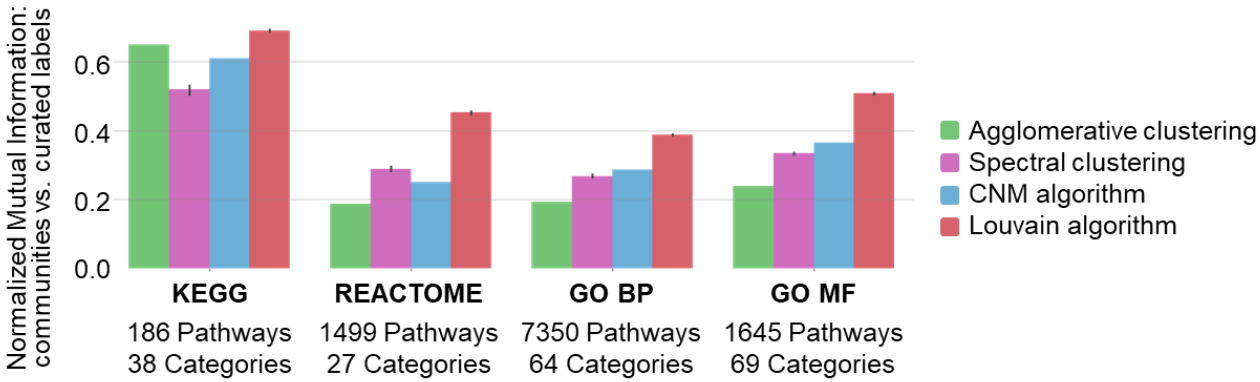


Figure 3.2. Comparison of graph clustering and community detection algorithms we evaluated. Pathway graphs and communities were learned separately for each database and then compared with curated category labels from each source database via normalized mutual information.

Automatic labelling of pathways

To automatically generate descriptions for our communities that were independent of expert-curated category labels, we developed a method based directly on the names of member pathways. For each community, we pre-processed member pathway names from MSigDB and identified commonly appearing terms. After pre-processing these names to remove database identifiers and common stop words (using the Natural Language Toolkit Python package), we counted instances of all 3-mers across community members and used the most commonly appearing 3-mer (with at least three appearances) as a label to describe our community. In the case of a tie, we selected the 3-mer whose source pathways had the highest average hubness within their community. When no 3-mers made at least three appearances, we repeated the process

with 2-mers. Supplementary Figure 3.7 shows an example of this procedure, and Supplementary Table 3.2 provides the top ten terms for each community.

Querying new gene sets

As described above, the PAC method creates a graph with edges based on gene set overlaps among pathways and learns communities via the Louvain algorithm, which greedily optimizes the modularity of graph partitions. Therefore, using the following process, it is straightforward to query a new gene set in our learned community network. First, we calculate Fisher's exact test *p*-values between the new gene set and each pathway already in the graph (Supplementary Figure 3.1b), and we then temporarily add the gene set as a new node in the pathway graph. Next, holding the previously learned partition constant and initially considering the new gene set as a single-member community, we calculate the modularity change associated with moving the gene set to each of the other communities. We then rank the candidate communities for the new query gene based on their associated increase in modularity.

To ensure that this approach is useful for any arbitrary list of genes (e.g., from a differential gene expression analysis), we first validated it by again using the MSigDB Hallmark pathways, because some of their founders are pathways in our graph. Thus, we evaluated whether, when queried, these Hallmark gene lists tended to be assigned to communities containing many of their founder gene sets. Finally, as described below, we created an interactive web tool to help users query a list of genes and identify candidate communities most closely associated with it.

Interactive web tool

Because the pathway graph contains thousands of pathways and learned communities contain up to 492 pathways, our integrative communities are not amenable to static visualizations. Therefore, we created an interactive web tool, available at https://nicasia.github.io/PAC, using the Plotly Dash framework. The tool lets users explore the community and pathway networks to gain a clearer understanding of how pathways relate to each other and the higher level processes learned by each community. It offers several views. First, a *community-level view* shows users the community graph (i.e., Figure 3.4a) along with detailed annotations, including the top automatically generated labels, pathway members, and most commonly appearing genes. Users can query specific biological processes (i.e., automatically learned labels) to identify which communities contain relevant pathways. Second, a *pathway-level view* reveals sub-graphs containing all pathway members for a selected community. Here, users can highlight a specific pathway in the graph to visualize how it relates to other pathways in its community. Third, a *gene-level view* helps users query any gene to see which communities contain pathways with that gene and whether the gene appears

disproportionately often in certain communities (Supplementary Table 3.3 provides the full list of genes with significant overrepresentation in any community as described in Supplementary Methods). Finally, we provide a page that lets users *query a new gene set* against our pathway network and visualize both enriched pathways and top related communities (based on the method described above). Together, these different pages for examining the learned communities of pathways, along with the ability to query a new gene set, may allow users to more effectively interpret their biological findings in the context of known pathways.

Biological example with breast cancer data

To demonstrate an application of our method to new gene sets, we used an example of a differential gene expression analysis in breast cancer samples. Data was provided by the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database. The database includes breast cancer sequencing data from 2,509 patients; gene expression measurements from biopsied breast cancer samples, including 24,368 measured genes; and phenotypic or treatment labels associated with each sample (e.g., estrogen receptor status and whether the patient was treated with chemotherapy).

We restricted our analysis to 2,469 sampled profiles for which estrogen receptor status was reported, of which 74% of samples were positive for estrogen receptors. For each of the 24,368 genes measured, we compared expression levels for ER+ and ER- samples using two-sided independent *t*-tests and identified 8,984 genes significantly differently expressed between groups ($p<.05$ after Bonferroni correction; Supplementary Methods describe data processing). Consistent with prior knowledge [80], this indicates that expression differences between ER+ and ER- cancers are widespread across genes. To refine our understanding of the top genes, we identified the top 1% (243) of genes with the most significant differential expression (Supplementary Figure 3.8) and examined them in the context of our community network, thus enhancing the interpretability of a set of differentially expressed genes.

Results

The PAC framework uses a Louvain community detection method that outperforms alternative approaches

To evaluate whether we could learn informative and meaningful communities from pathway networks, we separately applied the pathway network construction step to each pathway database and then applied various community detection and clustering algorithms to each individual graph. We then computed the normalized mutual information (NMI) to compare the resulting communities with the curated categories

(see Methods for details) from each database. Figure 3.2 shows that the Louvain community detection method achieved high NMI scores when comparing the automatically learned communities with ground-truth curated labels, exceeding NMI scores from all alternative approaches across all pathway databases (p<0.001 for all t-tests; alternative evaluation metrics revealed similar results, as shown in Supplementary Figure 3.2). This indicates that the Louvain community detection method generates communities that are consistent with expert-curated labels and offers a promising approach for automatically categorizing communities based on their shared genes.

PAC combines four major pathway databases, resulting in communities that are consistent with hand-curated categories

To learn a unified set of communities across all 4,847 gene sets from the four different sources (KEGG, REACTOME, GO BP and GO MF), we constructed a joint network and used the Louvain algorithm with a resolution of 0.4 (selected because it provided high NMI scores across all four datasets; Supplementary Methods & Supplementary Figure 3.4). This approach generated 35 communities, ranging in size from 7 to 492 pathways (Figure 3.3a) and effectively integrated pathways from different sources (Figure 3.3b), which indicates that the communities are likely driven by function rather than database-specific signals.

Like the separately generated ones, communities found from our combined pathway graph tend to be very consistent with their own curated categories (NMIs of 0.62, 0.43, 0.32, and 0.41 for KEGG, REACTOME, GO BP, and GO MF, respectively; NMI of 0.30 when all curated categories were concatenated; Supplementary Figures 3.9-3.13 show comparisons between these communities with curated categories). However, differing curated categories from each database are not easily reconcilable into common biological themes. Thus, to explore whether the learned communities capture meaningful common processes, we analysed our results with respect to the MSigDB's Hallmark pathways (v7.1), developed to summarize pathways across MSigDB's various sources[71].

All Hallmark pathways have associated sets of *founder pathways*, sets of pathways from which the Hallmark pathways are derived, including 745 pathways in our combined graph. We initially observe that our automatically learned communities are highly consistent with grouping pathways by Hallmark pathways for which they are founders (NMI=0.60; pathways that were not Hallmark founders were not considered in this calculation). Furthermore, purple cells in Figure 3.3c illustrates the distribution of Hallmark pathway founders within each community; most of our communities are associated with few Hallmark pathways, signifying that Hallmark pathways provide insight into coherent biological processes within communities. For example, there is a nearly one-to-one mapping between community 23 and the

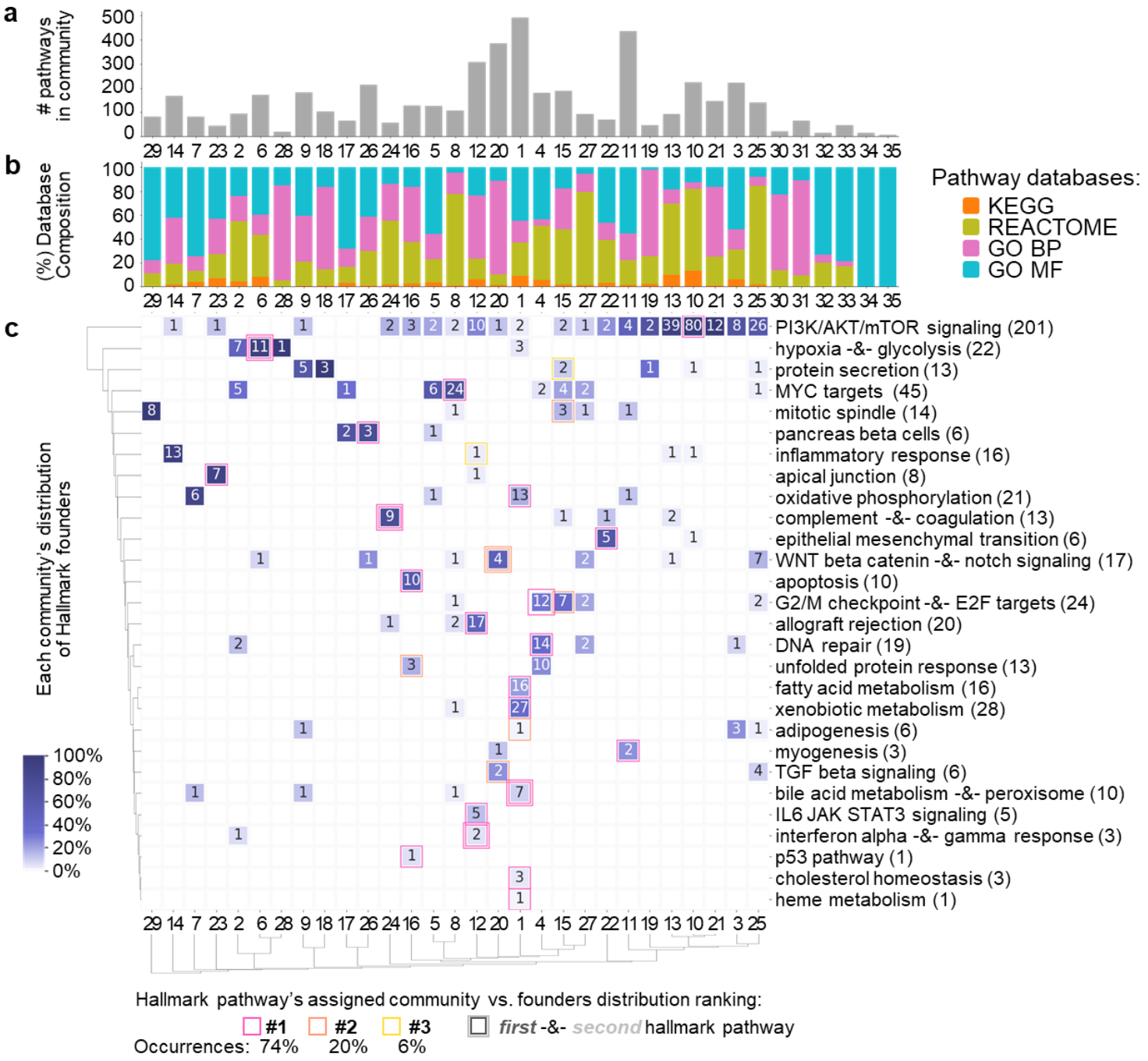


Figure 3.3. Overview of learned communities and their association with known processes. (a) Communities' number of member pathways (ranging from 7 to 492 members). (b) Composition of pathway database members in each community. (c) Communities' associations with MSigDB Hallmark pathways. Each Hallmark pathway has a list of associated "founder" gene sets, some of which are in the REACTOME and KEGG databases. Heatmap annotations indicate the number of founders in each community associated with each Hallmark pathway. Cells are colored by each community's frequency of founders distributed across Hallmark pathways (e.g., the darkest purple indicates that all Hallmark founders contained in a community are mapped to a single Hallmark pathway). We note that some Hallmark pathways have identical founder gene sets, and we include both Hallmark pathways in the same row of the heatmap (separated by "-&-" in the label). Finally, we use PAC's gene set querying method (described in Methods) to assign each Hallmark pathway to a community based on its member genes. The assigned communities are indicated by squares in the heatmap, and these squares are colored by how closely they agree with the top community based on founders.

Hallmark 'apical junction' pathway (i.e., genes involved in adherens and tight junctions between cells [71]), suggesting that this Hallmark pathway may be an appropriate annotation for community 23. This is further supported by the fact that most pathways in the community relate to cell-to-cell adhesion (Supplementary Table 3.1), consistent with the biological function of the Hallmark pathway. Similarly, all Hallmark apoptosis founders are in community 16, consistent with the fact that many central members of the community relate to apoptotic signalling (Supplementary Table 3.1).

Although the examples illustrated above and in Figure 3.3c are promising, because Hallmark pathways are based on only a small subset (15.4%) of our original pathways, they cannot be used to annotate smaller communities with no Hallmark founders. Furthermore, this interpretation may fail to reveal highly specific processes occurring within the communities. For example, the PI3K/AKT/mTOR Signalling Hallmark Pathway, broadly important in regulating the cell cycle, is split across several of our communities, showing that these communities may each represent more specific sub-processes. In the next section, we address this problem by automatically generating descriptive labels for each community.

Finally, we also use Hallmark pathways to validate our method for querying new gene sets in the PAC framework. Because each Hallmark pathway has a set of associated founder gene sets (some of which are in our pathway network), we use our querying method to assign each hallmark pathway to a community in our network; we can then determine if there is agreement between the community assigned via our method and communities containing its founders. For 74% of Hallmark pathways, Figure 3.3c shows that the assigned community based on our querying method is the same community that contains the most founders (see pink cell outlines in Figure 3.3c); for all Hallmark pathways, the assigned community is in at least the top three communities based on founder membership.

PAC's Automatically generated labels and visualizations provide high-level overviews of learned communities

To highlight high-level relationships among our pathway communities, we visualize our pathway community graph with their automatically learned labels (Figure 3.4a). First, we observe that many of our automatically generated label terms are consistent with our Hallmark pathway analyses (Figure 3.3c), e.g., Community 16, whose top label was "Apoptotic signaling pathway," is consistent with its most closely related Hallmark pathway, "apoptosis." Importantly, our labelling approach also provides insights unavailable from Hallmark pathway analysis alone. For example, Communities 9, 19, and 19 are all strongly associated with the Protein Secretion Hallmark pathway (Figure 3.3), but our automatically generated labels reveal separate phenomena (i.e., lipoproteins, synaptic processes, and viral activity for Communities 9, 18,

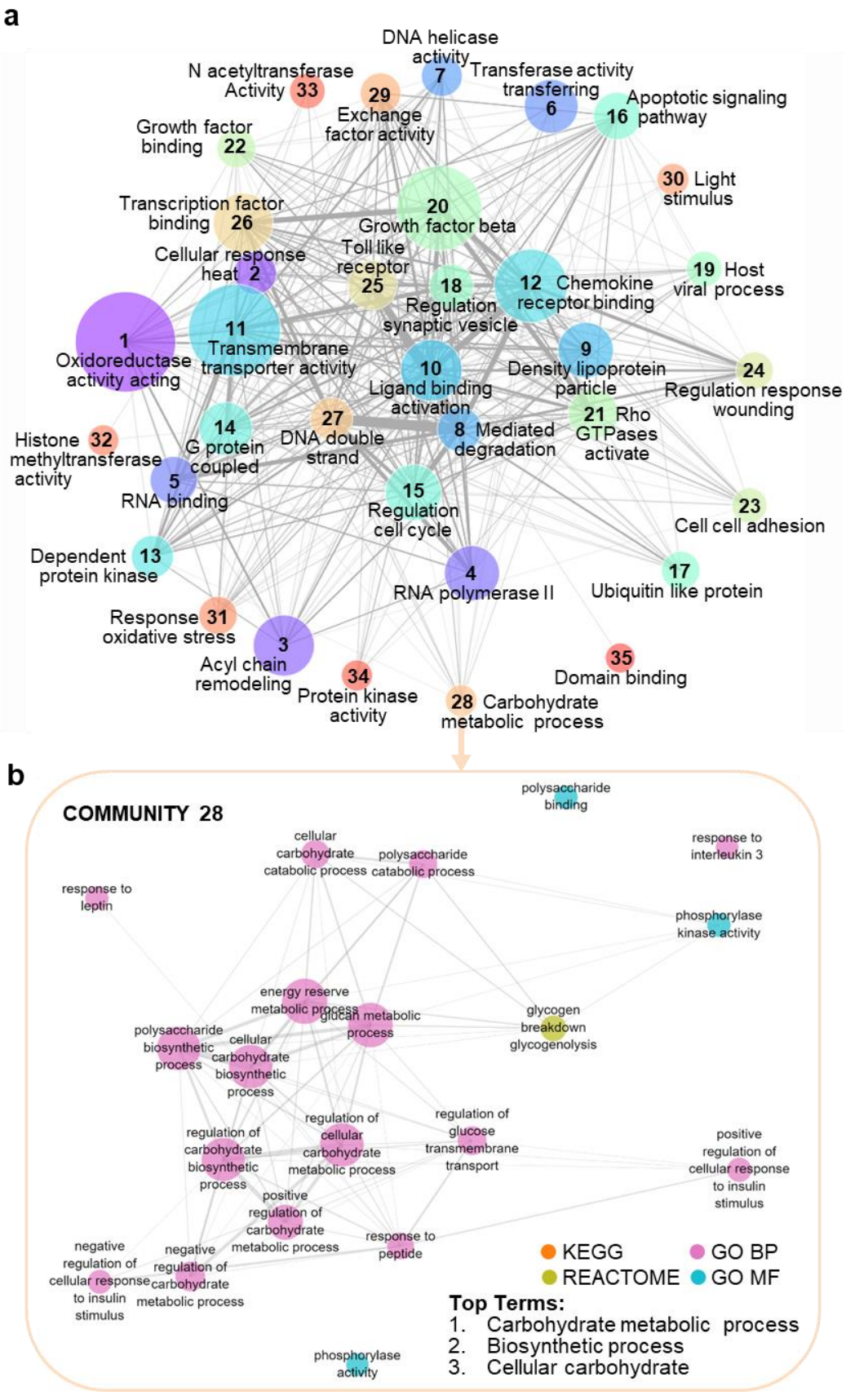

a
DNA helicase activity
7
N acetyltransferase Activity
33
29
Exchange factor activity
Transferase activity transferring
6
Apoptotic signaling pathway
16
Growth factor binding
22
30 Light stimulus
Transcription factor binding
26
20
Growth factor beta
Cellular response heat
2
Toll like receptor
25
18
Regulation synaptic vesicle
12 Chemokine receptor binding
19 Host viral process
1
Oxidoreductase activity acting
11
Transmembrane transporter activity
10
Ligand binding activation
9
Density lipoprotein particle
24
Regulation response wounding
27
DNA double strand
8 Mediated degradation
21 Rho GTPases activate
Histone methyltransferase activity
32
14
G protein coupled
5
RNA binding
15
Regulation cell cycle
23
Cell cell adhesion
Dependent protein kinase
13
4
RNA polymerase II
17
Ubiquitin like protein
Response oxidative stress
31
3
Acyl chain remodeling
34
Protein kinase activity
35
Domain binding
28 Carbohydrate metabolic process
b
COMMUNITY 28
polysaccharide binding
cellular carbohydrate catabolic process
polysaccharide catabolic process
response to interleukin 3
response to leptin
phosphorylase kinase activity
energy reserve metabolic process
glucan metabolic process
glycogen breakdown glycogenolysis
polysaccharide biosynthetic process
cellular carbohydrate biosynthetic process
regulation of cellular carbohydrate metabolic process
regulation of carbohydrate biosynthetic process
regulation of glucose transmembrane transport
positive regulation of cellular response to insulin stimulus
positive regulation of carbohydrate metabolic process
response to peptide
negative regulation of cellular response to insulin stimulus
negative regulation of carbohydrate metabolic process
KEGG
GO BP
REACTOME
GO MF
Top Terms:
1. Carbohydrate metabolic process
2. Biosynthetic process
3. Cellular carbohydrate
phosphorylase activity

Figure 3.4. **(**a) The final learned community network. We display all community nodes, with sizes proportional to the number of pathway members, and edge widths proportional to the average weights among all pairs of edges between members of each community. Each community is labeled with automatically generated labels as described in Methods. (b) An example view of a single community. We display all members of community 28 as nodes with sizes proportional to their hubness in the subnetwork containing only community 28 pathways, and colored by the database for each pathway. Edge widths are proportional to the –log10(p-value) for Fisher's exact test measuring gene overlap between pathways.

and 19, respectively). Overall, Figure 3.4a demonstrates that our communities' labels cover a broad range of cellular activities while remaining sufficiently specific to describe well-defined biological functions, while network edges highlight the interrelatedness of biological processes across these communities.

Finally, to demonstrate that our approach produces informative pathway communities, we explored individual networks within communities. As an example, Figure 3.4b visualizes Community 28, one of multiple communities associated with the Hallmark hypoxia and glycolysis pathways. Pathway members are consistent with the top label "carbohydrate metabolic process," and the community integrates pathways related to this biological function from multiple databases. These promising results highlight that our process can overcome database bias and capture functional similarities.

Supplementary Table 3.1 shows community pathway membership, and Supplementary Table 3.2 shows the top labels for each community. Our web tool provides detailed interactive visualizations for each community; see https://nicasia.github.io/PAC.

A biological example: genes differentially expressed between estrogen receptor positive vs. negative breast cancers relate to relevant pathway communities

Breast cancers are commonly classified by their estrogen receptor (ER) status – i.e., whether estrogen receptors are expressed in cancer cells (ER+) or not (ER-). These cancer subtypes manifest in markedly different ways and require different treatment regimens because ER+ cancers rely heavily on estrogen to grow and reproduce, whereas ER- cancers do not [81,82]. Thus, ER status is associated with a wide range of transcriptional differences between ER+ and ER- cancers [83]. In this example, we conduct a simple differential expression analysis to identify the most differentially expressed genes and then highlight how our method clarifies these findings. In particular, as described in Methods, we identify the top 1% of the most significantly differentially expressed genes between 1,825 ER+ and 655 ER- samples from the

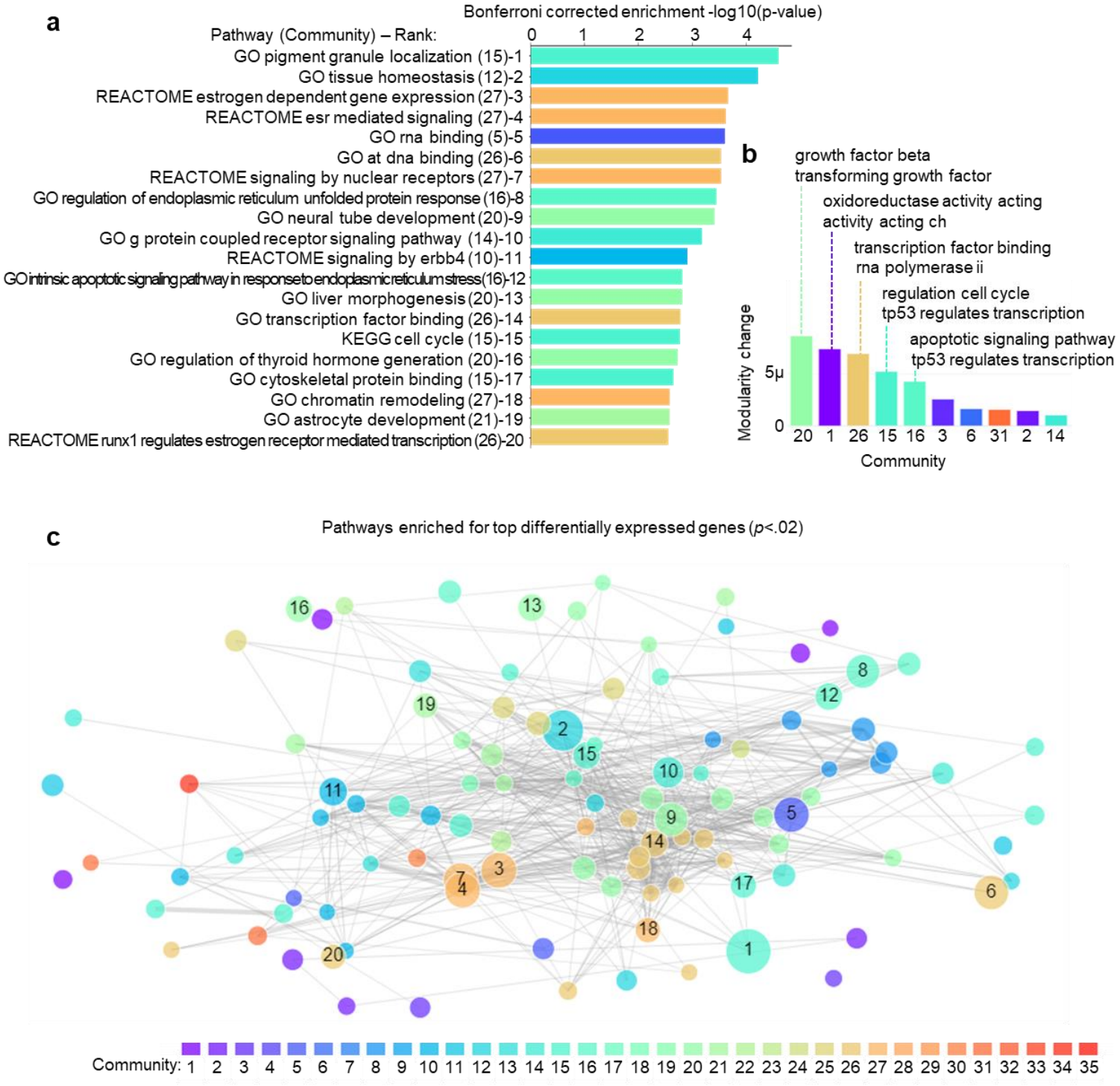


Figure 3.5. **(**A) Enrichment of top pathways for the 243 most significantly differentially expressed genes biopsied samples for ER+ vs. ER- breast cancers. Each bar shows the –log10(p-value) for enrichment based on Fisher's exact test after Bonferroni correction over all pathways. Each pathway is labeled with its assigned community and enrichment rank. (B) Using the PAC's gene set querying method (described in Methods), we query the 243 most significantly differentially expressed genes against the learned communities and display the modularity change associated with assigning the query gene set to each community. The top 10 communities are shown, with the top two automatically generated labels (Methods) for each of the top 5 communities, indicating processes that are generally related to the query gene set. (C) Network overview for the top 108 enriched pathways (p<.02) for the 243 genes described above. Each node is colored by its community and sized proportionally to the enrichment –log10(p-value), and we annotate the top 20 pathways with their rank as indicated in (A).

METABRIC dataset (Supplementary Figure 3.8) and use these genes as a query gene set to explore relevant pathways in the context of our learned communities.

Figure 3.5a shows the top enriched pathways (as determined by Fisher's exact test of overlap for the top genes and all pathways), which would be the final outcome of a standard pathway analysis. By additionally coloring each bar with the community to which the pathway was assigned, we see that some communities appear frequently in the top 20 enriched pathways.

Next, we query the top genes against our communities (see Methods). This approach considers not only the top enriched pathways, but the overall association of all pathways in each community with the query gene set. This analysis reveals that the top communities related to differentially expressed genes between ER+ and ER- broadly relate to growth factors and cell differentiation (community 20) and metabolic activity (community 1) (Figure 3.5b). Interestingly, although community 27 contains multiple estrogen-related pathways that are highly enriched in the top genes (e.g., #3 and 4 in Figure 3.5a), most pathways in community 27 relate to broader cell-cycle activity (Supplementary Tables 3.1-3.2); therefore, that community is not favored in the community-level analysis. Thus, although our approach identifies specific pathways relevant to our gene sets, use of the community network query reveals some broader patterns of processes captured in the top genes related to ER-status. Finally, Figure 3.5c shows a network visualization that highlights 108 pathways enriched at the $p<.02$ level in our gene set query. This helps the user visualize not only which pathways are enriched but also how they relate to each other (note that interactive versions of Figure 3.5b-c, which provide more detail, are available at https://nicasia.github.io/PAC).

Discussion

Our approach is not without limitations. In particular, because our graph is constructed using Fisher's exact test-based edges, our method for querying new gene sets also relies on the use of Fisher's exact test for pathway enrichment to be consistent with the rest of the graph's structure (since querying a new gene set involves simulating its addition as a node to the pathway network). Thus, our tool does not currently support the use of an alternative approach to compute pathway enrichment (e.g., GSEA); we may implement this functionality in the future. Additionally, although we empirically found that our automatic community annotation approach was consistent and interpretable, it relies primarily on the quality of pathway names. If the analysis were repeated with a new set of pathway databases that used uninformative pathway names, our approach for automatically labelling communities would not be applicable.

In summary, we contribute the PAC framework, an automated approach for identifying communities of closely related pathways across several databases that successfully recapitulates expert-curated categories. By conducting separate analyses on pathway databases, we verified that the learned communities were consistent with curated ground truth labels. When scaled up to an integrative analysis across four databases, we verified that learned communities were consistent with Hallmark pathways. We also found that maintaining a pathway-level understanding of our communities provides additional nuance and context that is lost by consolidated gene sets (e.g., using Hallmark pathways alone for pathway enrichment analysis). Further, unlike previous methods focused solely on visualization of pathway enrichment results, we leverage an automatic labelling approach to yield additional insights about biological pathway relationships. Finally, we believe that our tool to query new gene sets against our analyses (demonstrated with a breast cancer gene expression data example analysis) and our interactive webpage for examining relationships among pathways and communities will help computational biologists contextualize meaningful new findings for a wide variety of biological processes.

Chapter 4. Efficient and Explainable Risk Assessments for Imminent Dementia in an Aging Cohort Study

As the aging US population grows, scalable approaches are needed to identify individuals at risk for dementia. Common prediction tools have limited predictive value, involve expensive neuroimaging, or require extensive and repeated cognitive testing. None of these approaches scale to the sizable aging population who do not receive routine clinical assessments. Our study seeks a tractable and widely administrable set of metrics that can accurately predict imminent (i.e., within three years) dementia onset. To this end, we develop and apply a machine learning (ML) model to an aging cohort study with an extensive set of longitudinal clinical variables to highlight at-risk individuals with better accuracy than standard rudimentary approaches. Next, we reduce the burden needed to achieve accurate risk assessments for those deemed at risk by (1) predicting when consecutive clinical visits may be unnecessary, and (2) selecting a subset of highly predictive cognitive tests. Finally, we demonstrate that our method successfully provides individualized prediction explanations that retain non-linear feature effects present in the data. Our final model, which uses only four cognitive tests (less than 20 minutes to administer) collected in a single visit, affords predictive performance comparable to a standard 100-minute neuropsychological battery and personalized risk explanations. Our approach shows the potential for an efficient tool for screening and explaining dementia risk in the general aging population.[‡]

Introduction

Alzheimer's disease (AD), a degenerative brain condition, affects an estimated 5.8 million Americans. As the world's older population grows at an unprecedented rate, the number of individuals with dementia is projected to more than double, making it an increasingly pressing health concern[84]. Significant advances in diagnostic predictions are essential to curb the devastating effects of dementia worldwide. We believe these advances will be enabled by large-scale aging cohort studies and machine learning (ML) innovations.

Although no currently known treatment can cure or retard AD progression, identifying AD cases before severe neurological damage ensues is crucial. Predicting onset can promote treatment efficacy once successful interventions are developed and swiftly identify individuals who may benefit from drug trials. It

[‡] This paper was joint work with co-first-author Alex Okeson, along with Tim Althoff and Su-In Lee. It has been published in the *IEEE Journal of Biomedical and Health Informatics*[131].

will also help families plan for patient care and patients to receive resources to help make personal decisions about their care before they lose the autonomy to do so[85].

Although studies have demonstrated the possibility of identifying individuals who already have dementia[86], such diagnoses occur beyond the critical window for effective interventions or end-of-life planning[85]. Other studies have predicted the onset of dementia in advance of a clinical diagnosis, but often involve costly data collection using neuroimaging or in-depth neuropsychological batteries over multiple years [87–92]. The use of repeated cognitive testing may help to model and predict an individual's cognitive decline[90]; however, given that only 16% of American seniors receive regular cognitive assessments in primary care settings [93], this approach may be impractical for the general population. Furthermore, decreasing the required window of repeated testing would enable earlier diagnostic predictions because predictions would be made using fewer (and therefore earlier) observations.

Our goal is to find a balance between accurate but costly tests and efficient but relatively inaccurate predictions. In particular, we assess and explain an individual's risk for dementia multiple years into the future using relatively easy-to-collect measures that may scale well to large aging populations.

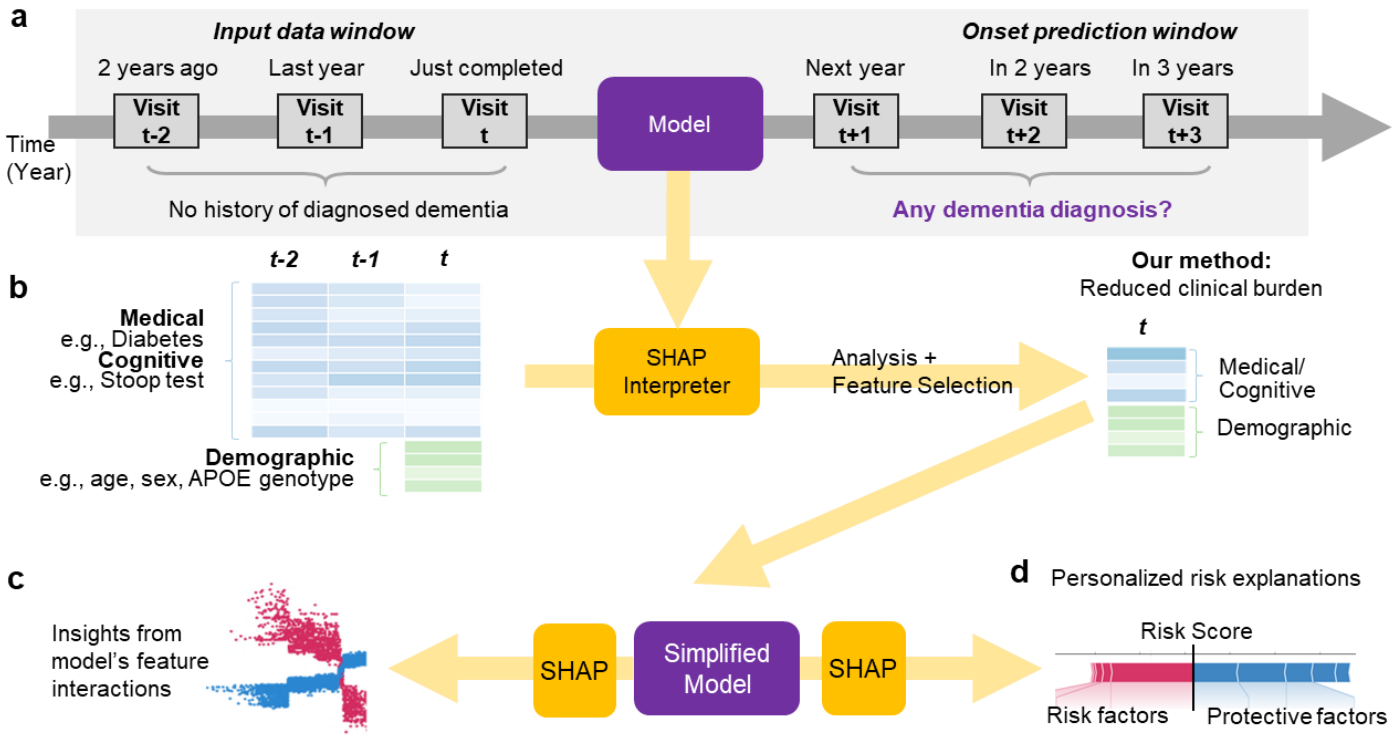


Figure 4.1. Overview of our approach to producing efficient and explainable dementia onset risk predictions. We link figure components to research questions (RQs) and in-text discussion. (a) RQ1, (b) RQ2, and (c) RQ3.

To this end, we address the following three research questions (RQs), encapsulated in Figure 4.1 and linked to in-text discussion. **RQ1:** Using longitudinal clinical and cognitive data from an aging cohort study, can we effectively predict whether an individual will develop dementia? **RQ2:** To what extent can we reduce the need for burdensome data collection while still maintaining predictive performance? We explore this question with respect to both repeated cognitive testing over multiple years and the number of required tests. **RQ3:** Using complex models that learn interactions among features and risks, can we leverage interpretability methods to provide personalized dementia risk explanations?

Our approach makes several noteworthy contributions. First, by exploring multiple classes of ML models, we find that dementia onset (within three years) can be predicted robustly and requires cognitive measurements from *only a single session*. Second, by using an interpretability method to measure sample-level feature importance, we can identify a *small subset of tests* that provide similar predictive value to a standard battery, while only taking one fifth of the time to administer. Third, each dementia risk prediction estimate is accompanied by *individual explanations of risk*, which may aid clinicians in tailoring care to their patients.

Related work

State-of-the-art dementia diagnosis. Many studies have sought to predict the presence of dementia based on brain scans and other metrics. For example, deep learning has improved AD classification using both magnetic resonance imaging and positron emission tomography scans[87,88]. Adding lifestyle and cognitive factors has additionally improved prediction performance for AD onset[92]. Although these studies have shown success in AD diagnosis and risk prediction, neuroimaging data requires significant amounts of time and funding, making it intractable for widespread use. In contrast, we develop imminent dementia predictions based on inexpensive measures. Our approach also complements current approaches by highlighting high-risk individuals who might benefit most from more extensive testing.

Basic risk factors. Without expensive brain imaging, it is common to predict the onset of dementia from age, sex, education, and genetic factors[7,94,95]. In particular, variations in APOE, the gene encoding Apolipoprotein E protein, are thought to be the main genetic factor impacting AD risk [94,96]. However, using only these basic risk factors produces non-robust predictions[97]. Here, we augment these primary risk predictors by adding cognitive and medical variables.

Modeling cognition trajectories. Because dementia is characterized by a rapid decline in cognitive functioning, studies have used cognitive variables to predict its onset[90]. Johnson *et al.* (2009) characterized cognitive trajectories for elderly individuals with and without AD and found that precipitous drops in

cognition tend to occur between one and three years prior to dementia diagnosis [91]. Based on this result, we use up to three years of past data to predict imminent dementia onset. Unlike these longitudinal cognition studies, however, we evaluate the need for repeated testing and attempt to reduce the burden on both clinicians and participants of required study visits to achieve accurate, but efficient predictions of dementia onset.

Diagnosing cognition status. Some research has focused on assessing whether an individual already has dementia[98] or mild cognitive impairment (MCI)[99] via short questionnaires. Multiple cognitive assessments have been developed to efficiently diagnose MCI[86,100], such as the Mini-Mental State Examination (MMSE). Further studies have used MCI diagnoses made by clinicians[101] and MMSE test scores[102] to predict future dementia onset. Building on these successes, we highlight a set of easily administered tests that significantly outperform the sole use of these clinical tests.

Results

We use data from the Religious Orders Study and Rush Memory and Aging Project (together known as ROSMAP)[6,7], two longitudinal aging cohort studies, to build dementia onset risk prediction models (see *Dataset* sub-section of Methods). During each yearly visit, individuals provide medical information and undergo extensive cognitive testing (Table 4.6). We generate samples with at least three years of consecutive visits and no dementia history and then build models to predict imminent dementia onset (i.e., a diagnosis within the next three years). Results described below are based on 9,103 samples from 1,597 individuals, split into stratified training and test sets.

Preliminary analyses reveal feature interactions

Preliminary data exploration comparing imminent dementia and control cases reveals many significant differences in the outcome variable among demographic and cognitive variables (Table 4.6). Additionally, strong correlations are seen between many features and the outcome variable, as well as among features themselves. This is expected since many of the cognitive tests assess the same cognitive domains. Together, these observations suggest that we could train an effective imminent dementia classifier from the available features. Furthermore, the high inter-relatedness of features indicates that some may provide redundant information and may therefore be reduced.

We also explore non-linear and interaction effects in our data to identify appropriate model classes. From these analyses, we observe two notable complex interactions, shown in Figure 4.2. First, having a single APOE e4 allele seems to modulate dementia risk in particular groups: males (Figure 4.2a), people

under 85 (Figure 4.2b), and relatively low cognition-scorers (Figure 4.2c). We observe similar modulation among carriers of two APOE e4 alleles (e.g., females), although they represent less than 2% of our sample (Table 4.6). Second, we see a strong interaction between overall cognition and many demographic features. Having a high cognition score may buffer dementia risk regardless of demographic factors, while demographic features might confer more information about risk when they coincide with low cognition. For example, APOE e4-carriers (Figure 4.2c), females (Figure 4.2d), older individuals (Figure 4.2e), and highly educated individuals (Figure 4.2f) seem to exhibit especially high risk if they are also low cognition-scorers. Due to such non-linear effects among our features, a complex model may be useful for capturing interactions among features and risk.

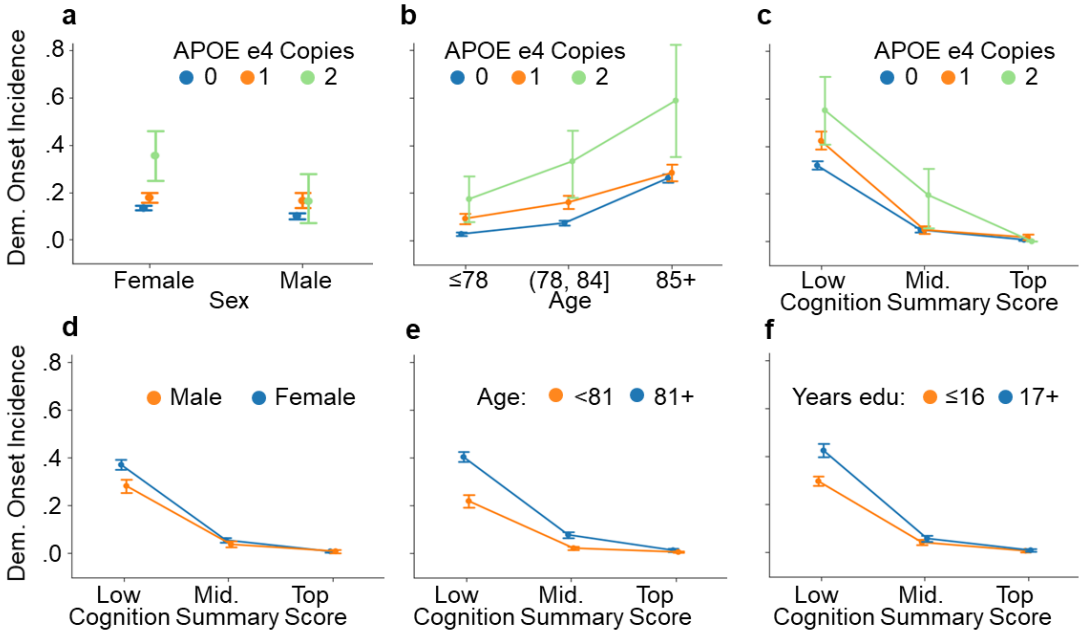


Figure 4.2. Average imminent dementia onset rates (with 95% confidence intervals) by demographic and cognitive factors, highlighting non-linear and interaction effects.

Multivariate models enable dementia risk prediction

To answer our first research question, we initially aim to build an ML model that can accurately predict dementia onset. To do so, we evaluate the prediction performance of multiple model classes and techniques

to address *class imbalance* and *time-series data* using stratified cross validation (CV) within our training set. Due to class imbalance in our dataset (13.7% rate of dementia onset), we consider various downsampling options. Due to the data's longitudinal nature, we explore the use of time encodings to pre-process input data (e.g., moving averages; described further in Methods). We find that models trained without downsampling or specialized time encodings had similar or better CV accuracy, AUROC, and AUPRC scores across all model classes described below, and thus proceed with these selections for all subsequent model tuning.

For our prediction task, we compare the performance of four classes of ML models: (1) regularized logistic regression (LR), (2) XGBoost (XGB), (3) multi-layer perceptron (MLP), and (4) long short-term memory network (LSTM). For each model class, we perform extensive hyperparamter selection across five stratified cross-validation (CV) splits (within the training set). Table 4.1 shows the top-performing models in each class (Methods provides tuning procedure details).

Table 4.1. Average cross-validation (CV) performance statistics for each model (± standard error).

Model	CV Accuracy	CV AUROC	CV AUPRC
XGB	0.9046 ± 0.0045	0.9163 ± 0.0044	0.6763 ± 0.0132
LR	0.9045 ± 0.0048	0.9205 ± 0.0044	0.6893 ± 0.0110
MLP	0.9036 ± 0.0056	0.9186 ± 0.0050	0.6694 ± 0.0144
LSTM	0.9021 ± 0.0050	0.9047 ± 0.0168	0.6691 ± 0.0189

In general, we find that many of the model classes achieve similar predictive performance. MLP, LR, and XGB models perform similarly (within the standard error ranges) with respect to AUROC and AUPRC. Among complex model classes, we chose the XGB model because the neural network methods (MLP and LSTM) exhibit unstable performance, as shown by their large error bars in Figure 4.3 (particularly when trained on a single year of data). We opt for an XGB final model over a linear (LR) one because: (1) Unlike linear models, XGB is able to learn non-linear and interaction effects like those found in our data. Prior meta-analyses of dementia risk prediction suggest that the linearity assumption does not hold for critical risk factors (consistent with our observations in Figure 4.2) [103]. Another study found that, even when producing equally accurate predictions, linear methods applied to non-linear data sets tend rely on irrelevant features[104]. Thus, XGB may learn a richer representation of the true complex relationships among features. (2) Due to the non-linear and interaction effects learned by XGB, we can obtain personalized risk explanations for each individual via interpretability methods (e.g., SHAP, see Methods),

whereas linear models place the same importance on each feature across individuals. Thus, we elect to perform final analyses with an XGB model but compare these results to a linear model for completeness.

Recent, not cumulative, observations are needed for effective dementia onset prediction

We answered our first research question by successfully predicting future dementia onset from three years of consecutive ROSMAP study visits. However, the use of repeated visits may be unrealistic for predicting dementia onset in the general population since only 16% of American seniors receive regular cognitive assessments[93]. Therefore, we turn to RQ2 to evaluate whether we can reduce the burden of repeated cognitive testing (i.e., do we need multiple years of cognitive measurements to make an accurate prediction?). To that end, we evaluate our model's CV AUROC when we reduce the number of consecutive visits in the inputs (described further in Methods). As we reduce the number of cumulative years the model sees during training (Figure 4.3, circular markers), we find no major changes in model performance across all four model classes. This suggests that requiring multiple years of consecutive data is not necessary for accurate predictions, which may reduce the burden of regimented follow-up testing in the clinical setting.

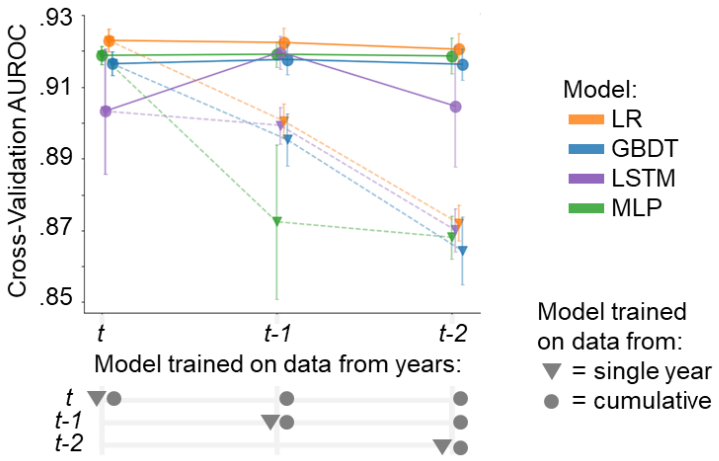


Figure 4.3. Average cross-validation area under the receiver operating curve (AUROC) for our four models trained on different combinations of yearly visits. Circle marks show that cumulative data has limited value, while triangle marks highlight the importance of recent data.

Next, we identify the relative importance of recent data by evaluating the model trained on a single year of past data alone. As expected, we see a decline in prediction performance for models trained on older data (Figure 4.3, triangular markers). Although the most effective prediction models were trained with the

most recent year of observations (*t*), we evaluated the stability of our final conclusions by repeating analyses shown in Figure 4.5 and Table 4.2 using data from *t-1* and data from *t-2* (the same data shown by triangle markers in Figure 4.3). In both cases, models show slight drops in all performance metrics, but our models outperform baselines by similar margins on time *t-1* and *t-2* data compared to *t* data. Together, these results imply that recent cognitive measurement are vital for predicting imminent dementia status, but that repeated testing is not needed for accurate dementia onset prediction since recent data may supersede outdated cognitive information obtained in past years.

Finally, we apply SHAP[105], a local feature attribution method, to our XGB model trained on all three years of consecutive data to ascertain whether the model relies on previously collected data. We find that the model's top ten features consist of demographic data or tests from the most recent year: even when provided access to measurements from prior years, our model still tends to focus on more recent data. Based on these CV results, we decided to train the final models using only the current year (*t*) of data, a decision that enables earlier, more efficient predictions that need not wait for additional years of cognitive tests before generating a dementia prediction.

Table 4.2. Test performance of final models (± standard error from bootstrap re-sampling).

	Test Accuracy	Test AUROC	Test AUPRC	Relative IDI (simplified with APOE vs. row)
Final models				
All Features (XGB)	0.8975 ± 0.0002	0.8977 ± 0.0003	0.6387 ± 0.0010	−0.0571
Simplified (with APOE) (XGB)	0.8947 ± 0.0002	0.8903 ± 0.0003	0.6236 ± 0.0010	-
Simplified (no APOE) (XGB)	0.8964 ± 0.0002	0.8896 ± 0.0003	0.6184 ± 0.0010	0.0084
Baseline models in our study				
Linear Selected Features (LR)	0.8825 ± 0.0002	0.8224 ± 0.0004	0.4907 ± 0.0011	0.6012
Linear Selected Features (XGB)	0.8781 ± 0.0002	0.8050 ± 0.0005	0.4771 ± 0.0011	0.7432
Baseline feature sets in the literature				
Demographics + MCI (XGB)[101]	0.8770 ± 0.0002	0.8203 ± 0.0005	0.4449 ± 0.0011	0.5058
Normalized Cognitive Features Sum (LR)	0.8737 ± 0.0002	0.8128 ± 0.0005	0.4473 ± 0.0011	0.9804
Demographics + MMSE30 (XGB)[102]	0.8748 ± 0.0002	0.8124 ± 0.0005	0.4273 ± 0.0011	0.6707
Demographics (XGB)[6]	0.8593 ± 0.0003	0.7215 ± 0.0005	0.2660 ± 0.0008	2.9291

Efficient and effective dementia onset predictions can be made with a small subset of features

After extensive cross-validation experiments, we settle on a final XGB model using the hyperparameters selected based on CV performance. This final "All Features" XGB model is trained on all training data using all available features from year *t* only. Table 4.2 shows held-out test set performance metrics for this model. To drive further insights, we use SHAP local feature explanations[105] to interpret the final model. To see which features our XGB model relies on, we aggregate the local explanations of our training samples to obtain global insights (described in Methods). Figure 4.4 shows the top 20 most important features (ranked by their average SHAP importance magnitude across all samples).

First, we note that the feature attributions are consistent with findings in the literature, validating our modeling approach and SHAP interpretations. For example, nearly all previous work[6] has found females, older individuals, and carriers of an APOE e4 allele to be at higher risk of dementia, consistent with Figure 4.4 SHAP explanations. Similarly, as expected, low performance on all cognitive tests contributes to a higher risk score. In contrast, our years of education feature attributions are not consistent with the literature (which find negative associations between high education and dementia incidence). We discuss this result further in the Discussion Section.

As we move down the list of top-ranked features, we see a dramatic drop in the magnitude of SHAP values (i.e., relative influence of a given feature on the final prediction). We therefore hypothesized that future dementia onset can be predicted using only the most informative features. For evaluation, we choose the top four demographic features and top four cognitive tests and use them to train a simplified prediction model. The top demographic features (age, sex, education, and APOE genotype) are widely cited as being important[6] and are simple to measure. The four top cognitive tests chosen are: categorical fluency (Cat Flu; 2 minutes; semantic memory); symbol digit modality test (SDMT, ≤5 minutes, perceptual speed); word list test (WL, 3 minutes, episodic memory), and mini-mental state exam (MMSE30, 5-10 minutes, general cognition); Table 4.3 describes these tests. Interestingly, each test lies in a different cognitive domain in Table 4.6, indicating that the model relies on diverse and non-redundant cognitive attributes. From our simplified feature set, we train two "simplified" final models on our full training set: one including and one excluding the APOE genotype (which, though commonly used in prior studies, is not always available in clinical settings). Although cognitive diagnostic status (MCI diagnosis) was ranked among the top influential features, we excluded it from our simplified models: it is very time consuming to obtain in the ROSMAP study (since it is based on all cognitive tests and a clinician examination), and it may be difficult to obtain in the general aging population.

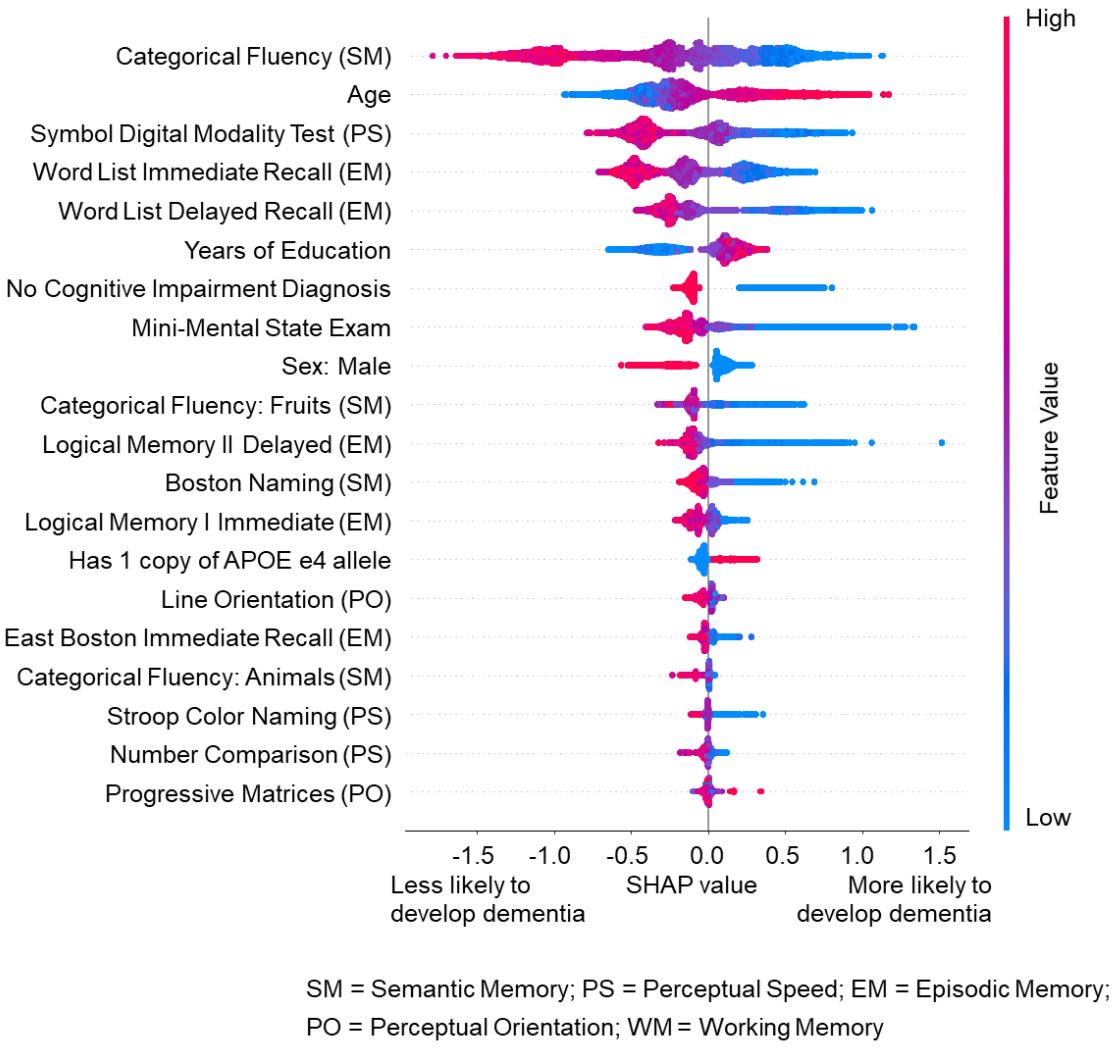


Figure 4.4. SHAP summary plot: violin plot of the 20 most informative features of the XGBoost current year model, ordered by importance. Each point is a training sample colored by its feature value. The point's x-axis position is the feature's contribution to the final risk prediction.

Unlike the above use of SHAP-based feature selection from our XGB model, feature selection for linear models involves choosing those with the highest magnitude regression coefficients. For comparison with our SHAP selection method above, we use standard feature selection based on the final LR model's coefficient magnitudes. For consistency, we use the same four demographic features as above and then select the cognitive features with the highest-magnitude regression coefficients: digits forward, digits backward, digits ordering (all working memory), and the East Boston Test (episodic memory) (21 minutes total; See Table 4.3).

Table 4.3. Selected cognitive tests from XGBoost (XGB) and linear regression (LR) models (cognitive domains shown in table 4.6). Full cognitive battery: 98 minutes.

Test (Domain)	Time (min)	Description
Selected cognitive tests from XGB model:		
Categorical fluency (SM)	2	Subject names as many items in a category as they can in a minute (Rounds: animals, fruits).
Symbol digit modality (PS)	≤5	Subject learns a symbol-to-digit mapping, then must substitute digits when symbols are shown.
Word list (EM)	3	Subject hears a list of 10 words, then is tested on immediate recall, delayed recall, and recognition (selecting correct words from distractors).
Mini-mental state exam	≤10	Short diagnostic general cognition test for dementia.
Selected cognitive tests from LR model:		
Digits Forward (WM)	5	Given a list of numbers, subject repeats them in the same order as given.
Digit Ordering (WM)	5	Given a list of numbers, subject repeats them in numerical order.
East Boston Test (EM)	6	After hearing a short story, subject recalls story units immediately and after distractor-filled delay.
Digits backward (WM)	5	Given a list of numbers, subject repeats them in the reverse order as given.

We compare our final simplified XGB models to the XGB model trained on the full feature set, an LR and XGB model trained using the features from linear feature selection, and a baseline XGB trained on multiple commonly used clinical baseline feature sets. Figure 4.5 shows the held-out test set's receiver operating curves for all models, highlighting the sensitivity and specificity based on all decision boundaries on the test set. Using a decision cut-off of 0.5, we report true negatives, false positives, false negatives, and true positives for our top models and the top performing baseline model in Table 4.4. For each model, Table 4.2 lists the area under the receiver operating curves (AUROC), precision recall curves (AUPRC), and the accuracy at a 0.5 decision cut-off point. Additionally, we calculate the relative integrated discrimination improvement (IDI), comparing the "Simplified (with APOE)" model's discrimination ability to every other model[106].

Together, these results show that computing SHAP feature importances for our XGB model allows us to identify measures that are particularly useful in our model and thus dramatically improve prediction performance over more basic clinical baselines by including a few short cognitive tests. These tests are standardized and simple to administer; any primary care physician or assistant could conduct them during

a patient's annual physical exam (taking a total of 15-20 minutes to administer compared to the 98 minutes required for all tests in the ROSMAP neuropsychological battery in Table 4.3).

Table 4.4. Using a 0.5 decision cut-off, we report the number of true negatives (TN), false positives (FP), false negatives (FN) and true positives (TP) in the test set.

	TN	**FP**	**FN**	**TP**
All Features (XGB)	1510	50	135	110
Simplified (with APOE) (XGB)	1513	47	143	102
Demographics + MCI (XGB)[101]	1479	81	141	104

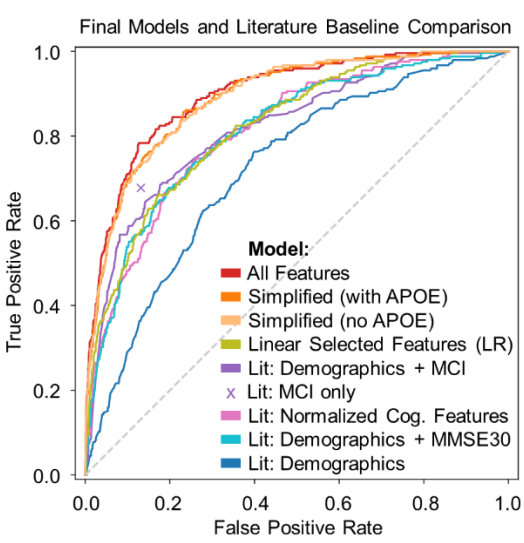


Figure 4.5. Receiver operating curves: final models and baselines from the literature (Lit). Area statistics are shown in Table 4.4.

Cross-cohort generalizability. Due to differences in study design and measured features, it is uncommon for dementia prediction studies to validate findings with external datasets[97,103]. While our data is comprised of pooled ROS and MAP samples, the studies recruit participants from different groups (clergy from Catholic religious organizations across the US and individuals in retirement facilities throughout

northern Illinois, respectively)[6,7], and these studies differ in demographic and lifestyle factors and outcomes. Thus, we seek to evaluate the cross-study generalizability of our final models. Using our previously defined training and test splits, we retrain our Simplified (with APOE) model separately for ROS and MAP training samples and evaluate each model's performance separately for ROS and MAP held-out test samples. In both cases, the "external" and "internal" test set AUROCs are within 0.01 of each other (Table 4.5). Furthermore, similar tests would be selected if we were to perform feature selection based on models trained separately from each cohort (the same top four tests for ROS, and three of four top tests--with the number comparison test replacing MMSE--for MAP). Together, these findings indicate that the model generalizes stably and effectively across cohorts, both in terms of predictive performance and selected features.

Table 4.5. Cross-study test set performance for ROS vs. MAP models.

		AUROC for Test Samples	
		ROS (N=1156)	MAP (N=649)
Training Samples	ROS (N=4506)	0.8848	0.8948
	MAP (N=2792)	0.8792	0.8851

Missing data experiments. As shown in Table 4.6, many features have missing values and are imputed for all analyses (see Methods). In particular, some features are missing at significantly different rates for control versus dementia onset cases. To ensure that our promising results were not driven by a confounding effect of imputing features at different rates between case and control groups, we experiment with removing potentially confounded samples and features as follows. We first exclude features with one-fifth of samples missing (Stroop color naming and Stroop word reading tests). We next exclude all samples with a missing observation for any of the remaining 10 features with significantly different rates of missingness between control and dementia onset cases, resulting in a new dataset with 8,392 samples (92% of the original dataset). First, we note that final model performance on this filtered dataset (test AUROC=0.8952) is very similar to performance from the full dataset (test AUROC=0.8977). Importantly, our SHAP feature rankings (generated via average SHAP importance magnitude) result in the same top four selected cognitive tests as the original dataset. Further, we observe similar performance for the final simplified model (test AUROCs of 0.8865 and 0.8903, respectively, for filtered and original datasets). Together, these experiments indicate that imputing missing features had little effect on our final models.

SHAP provides personalized risk explanations

We turn now to our third research question, which addresses personalized dementia risk explanations. Because XGB learns complex relationships among features (unlike linear models), we examine SHAP interaction values among pairs of features (see Methods). For example, according to XGB interactions,

having one copy of the APOE e4 allele impacts an individual's XGB risk prediction, particularly if he or she has a low cognition score (Figure 4.6a, consistent with Figure 4.2c) or is younger (Figure 4.6b, consistent with Figure 4.2b). Finally, males, especially those younger than 80, are at particularly low risk for developing dementia (Figure 4.2c, consistent with earlier findings[6]). Thus, by using SHAP to interpret our simplified XGB model, we find that aggregating non-linear feature effects across samples reveals relevant interactions learned by the model, and that these interactions are consistent with our data's structure (Figure 4.2) and prior literature.

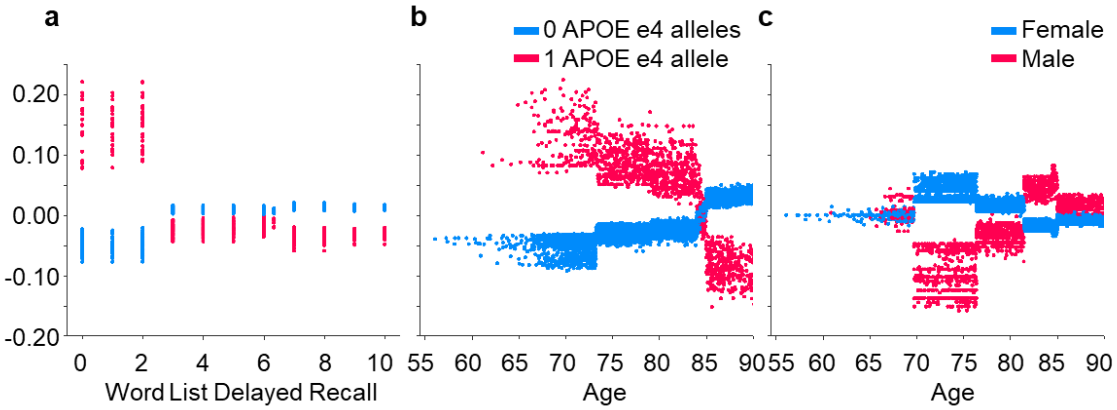


Figure 4.6. SHAP interaction values for selected pairs of features in our final Simplified (with APOE) XGBoost model.

Beyond receiving a risk score, using XGB and SHAP feature attributions gives patients and their medical practitioners a personalized explanation of risk (i.e., how particular features drive the XGB's prediction). To illustrate how this benefit works, we generate a synthetic sample that represents the “typical sample” in our dataset (with mode- or average-valued features) and display the risk score and explanation in Figure 4.7a. We show perturbations to single features of APOE (where we change the APOE e4 allele count from zero to one) in Figure 4.7b and the word list delayed recall (WLDR) score from the average value to two standard deviations below the average in Figure 4.7c. In both examples, we see that the perturbed variable becomes the primary risk factor driving up the risk score compared to the “typical individual."

Finally, in Figure 4.7d, the effect of both risk factors from parts b and c shows that the combined risk of having one APOE e4 allele and a low WLDR score substantially increases risk. In particular, the jump in risk from both risk factors (a 0.21 increase over the "typical individual") far exceeds the additive effects of each single risk score alone (0.05 and 0.07 increase, respectively). A linear model, in contrast, would have produced additive predictive importance values and therefore would have failed to identify a compounding effect of these features. This example highlights the ability of our XGB model with SHAP interpretations to provide personalized risk explanations based on a combination of feature values. This ability may prove to be powerful in clinical settings because it would help clinicians discuss the unique configuration of risk factors relevant to individual patients.

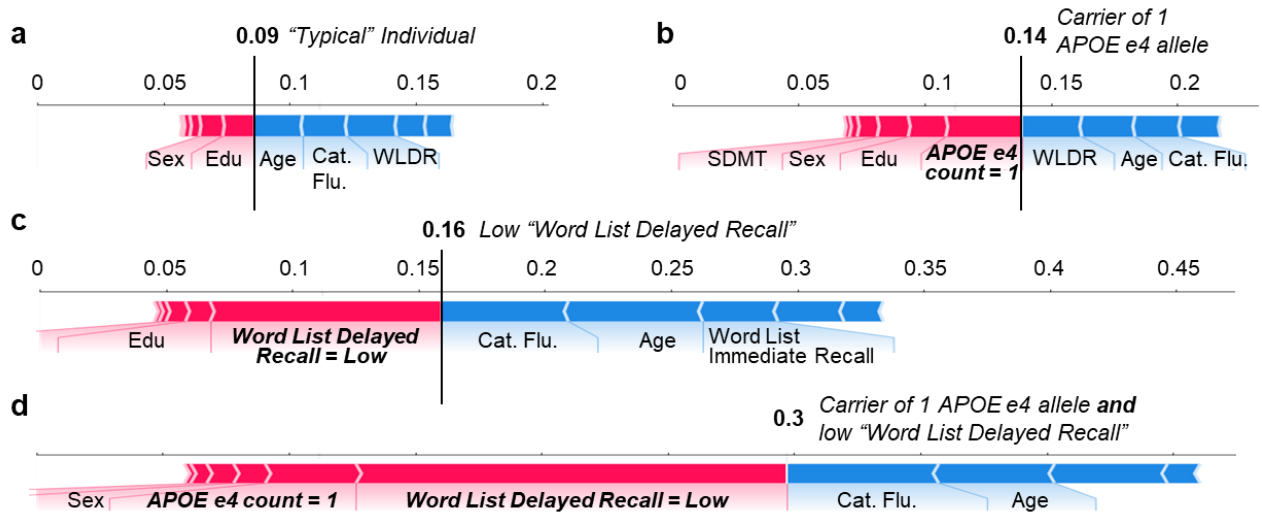


Figure 4.7. Feature explanations for synthetic samples: (a) risk and explanations for a "typical individual" in the ROSMAP data, (b, c) perturbations to single features (bolded), (d) the combined effects of both risk factors.

Discussion

Comparison with previous findings. Reviews of dementia prediction studies have found that combinations of cognitive tests have aided in the prediction of dementia onset[97,107]. In particular, for predicting conversion from MCI to dementia, combining episodic memory tests with executive functioning or language tests tended to produce high predictive accuracy[107]. A review of community-based aging cohort studies (consistent with our approach) also found that using three or four tests spanning multiple cognitive domains

led to improved predictions of dementia onset for 2.5 to 5 year follow-ups[97]. Compared to our results, these studies reported similar or lower AUROCs (ranging from 0.83 to 0.88); however, each study was based on samples from different cohorts (ranging from 478 to 551 total participants) and with different follow-up periods, so direct comparison may not be appropriate. Importantly, despite being performed on a larger cohort (1,597 individuals) and using a non-linear XGB model (unlike the previous studies, which all relied on linear analyses), our approach identified a small number of tests spanning multiple cognitive domains (Table 4.3) as predictors of dementia, consistent with these prior studies.

Longitudinal input data. Curiously, our analyses show the modest value of longitudinal measurements. Because dementia is an acquired condition marked by cognitive decline, one might expect to see gradual changes in cognition prior to dementia onset. In fact, our choice of a three-year input data window was based on observed cognitive changes preceding dementia in prodromal cases[91]. However, because our goal is to predict a *future* dementia diagnosis (not a current one), changes in cognition scores may be less useful than expected. Our results seem consistent with other studies, which reported limited value for cognitive changes in predicting future dementia onset. In particular, one study found that reliable change indices (RCIs) for MMSE had low predictive accuracy for dementia onset[108]. Furthermore, because longitudinal input data inherently requires rarer datasets with multiple cognitive assessments, RCI-based studies have often failed to achieve the same predictive accuracy as single-observation studies[97].

Limitations. Our final dataset contained 9,103 samples from 1,597 individuals, of which, 521 developed dementia. Although our study is based on a larger dataset than prior studies mentioned above[97], future studies should replicate our findings in other populations. Because we rely on samples from the ROS and MAP cohorts, our findings are subject to potential bias introduced by each cohort's procedures. In particular, for our ROSMAP samples, approximately three quarters come from females and two thirds from participants with 16 or more years of education. The unusually high education levels in our data may explain why some feature explanations for education level are inconsistent with findings in the literature. Future studies should especially explore sex- and education-based dementia risk in a more balanced dataset.

It is uncommon for dementia prediction studies to validate their findings externally due to prohibitive differences in study design, populations, and measured features[97]. According to a recent review[103], less than a quarter of examined ML studies externally validated their findings (the majority of which were imaging studies with harmonized measurements). As with many prior studies, we could not directly assess our findings on an external dataset. Nevertheless, despite differences between the ROS and MAP cohorts, our models generalized well between them when trained separately.

Additionally, the Stroop color naming and word reading tests had high levels of missingness in our dataset (Table 4.6), so it is possible that those tests may have been more highly ranked if they were observed in more samples. However, most features had relatively low rates of missingness (Table 4.6), and we found that there was not a significant relationship between feature missing rates and their SHAP importance for our final XGB model (Pearson correlation r=-0.11, p=0.46). Furthermore, analyses described in *Missing Data Experiments* sub-section showed that our imputation methods did not significantly affect our findings.

Finally, our initial choice of time window (three input years and three years of onset monitoring) limited the samples that were included from the ROSMAP dataset, biasing our sample against individuals with fewer than six yearly visits. We made this decision based on prior work, which suggested at least one-to-three years of cognitive data are useful for modeling cognitive decline in prodromal dementia patients[91]. We also viewed this as a necessary drawback in order to evaluate whether longitudinal data is needed for accurate prediction.

Conclusion

We conducted an in-depth analysis of many ML models, sampling techniques, and usages of time-series data to obtain models that predict imminent dementia onset more accurately than basic demographics-based or single-test approaches and more efficiently than predictions from a full neuropsychological battery.

Importantly, we can accurately predict imminent dementia diagnoses using data from just one clinical visit consisting of only demographic information and four easily measured cognitive tests that can be conducted in less than 20 minutes (five times shorter than the standard cognitive battery in the ROSMAP study). By using complex non-linear models and leveraging ML interpretability methods, we also generate personalized explanations of risk predictions that account for non-linear and interaction effects. These findings may provide substantial clinical value given the growing aging population and low rates of routine medical assessments. Our method could be scaled to explain and highlight at-risk individuals for additional dementia screenings, preventative treatments (when they become available), and enable planning for a potential imminent diagnosis. Our study takes important steps toward using complex models to generate explainable dementia risk predictions from relatively cheap metrics. While our findings highlight the effectiveness of our approach, more studies are needed to provide further validation for use in clinical practice. Nevertheless, we provide a framework with which others may replicate our experiments and construct models tailored to other cohorts and their measured cognitive tests.

Table 4.6. Between-group baseline (time *t*) statistics. We provide summary statistics for each group (including missingness rates and indicators of significantly higher rates of missingness for one group). (*p<.05, **p < .01, ***p < .001 for statistical tests.)

	All samples: Between-group test statistic	Controls: No impending dementia (N = 7866)	Dementia onset within 3 years (N = 1244)
Demographics			
Age	t = −29.60***	80.12 ± 6.77 (0%)	86.19 ± 6.37 (0%)
Sex: % male	χ^2 = 15.43***	29.5% (0%)	24.0% (0%)
Years of education	t = 1.45	16.91 ± 3.61 (0.1%)	16.75 ± 3.56 (0.2%)
Race (White/Black/Native American/Asian)	χ^2 = 12.72**	94.0%/5.5%/0.3%/0.2% (0%)	94.2%/4.9%/0.2%/0.6% (0%)
Ethnicity: % Hispanic	χ^2 = 0.10	3.1% (0%)	3.3% (0%)
# APOE e4 copies (0/1/2)	χ^2 = 54.59***	78.6%/20.3%/1.1% (1.5%)	70.3%/27.0%/2.8% (1.4%)
Episodic Memory (EM)			
Word list: immediate (1min)	t = 40.10***	20.57 ± 4.55 (2.4%)	15.01 ± 4.08 (2.3%)
Word list: delayed (1min)	t = 45.43***	6.76 ± 2.16 (2.4%)	3.69 ± 2.34 (2.4%)
Word list: recognition (1min)	t = 32.82***	9.85 ± 0.56 (2.3%)	9.02 ± 1.74 (2.7%)
East Boston test: delayed (3min)	t = 26.25***	9.89 ± 1.73 (0.3%)	8.46 ± 2.03 (1.0%**)
East Boston test: immediate (3min)	t = 34.59***	9.64 ± 1.90 (0.5%)	7.41 ± 3.07 (1.2%**)
Logical memory I (3min)	t = 37.73***	14.34 ± 4.10 (2.3%)	9.51 ± 4.44 (2.1%)
Logical memory II (3min)	t = 40.47***	13.23 ± 4.45 (2.4%)	7.60 ± 4.81 (2.4%)
Perceptual Orientation (PO)			
Line orientation (15min)	t = 13.54***	10.59 ± 2.97 (3.8%)	9.32 ± 3.02 (6.1%***)
Progressive matrices (20min)	t = 22.82***	11.65 ± 2.82 (5.1%)	9.61 ± 2.79 (8.2%***)
Perceptual Speed (PS)			
Symbol digits modality test (5min)	t = 37.91***	41.77 ± 10.09 (3.9%)	29.72 ± 9.87 (7.2%***)
Number comparison (3min)	t = 26.12***	26.22 ± 7.23 (3.7%)	20.29 ± 7.12 (6.2%***)
Stroop color naming (3min)	t = 22.30***	20.19 ± 7.34 (65.2%***)	12.34 ± 6.55 (60.0%)
Stroop word reading (3min)	t = 13.66***	48.87 ± 13.53 (65.3%***)	39.74 ± 14.55 (60.1%)
Semantic Memory (SM)			
Boston naming (5min)	t = 29.15***	14.19 ± 0.98 (3.0%)	13.22 ± 1.51 (4.0%)
Categorical fluency: animals (1min)	t = 33.21***	18.25 ± 5.45 (0.1%)	12.90 ± 3.96 (0.4%)
Categorical fluency: fruits (1min)	t = 37.98***	18.26 ± 5.13 (0.2%)	12.44 ± 4.15 (0.6%*)
Categorical fluency (combined)	t = 40.00***	36.51 ± 9.42 (0.1%)	25.33 ± 7.08 (0.4%)
National adult reading test (2min)	t = 5.15***	8.49 ± 1.94 (3.6%)	8.17 ± 2.14 (6.7%***)
Working Memory (WM)			
Digits backward (5min)	t = 16.94***	6.61 ± 2.05 (0.4%)	5.56 ± 1.82 (0.9%*)
Digits forward (5min)	t = 11.92***	8.43 ± 1.98 (0.2%)	7.70 ± 1.99 (0.6%)
Digit ordering (5min)	t = 21.30***	7.60 ± 1.56 (1.0%)	6.57 ± 1.67 (2.4%***)
Global Cognition			
Mini-mental state exam (5-10min)	t = 45.16***	28.59 ± 1.51 (2.2%)	26.20 ± 2.71 (1.4%)
Medical history/lifestyle factors			
MCI (No/Yes/Yes-other)	χ^2 = 1685.26***	86.9%/12.8%/0.3% (0%)	37.1%/60.7%/2.3% (0%)
Medical conditions sum	t = −3.77***	1.68 ± 1.16 (2.0%)	1.82 ± 1.21 (2.0%)
Vascular disease burden	t = −7.02***	0.45 ± 0.66 (2.0%)	0.59 ± 0.75 (2.0%)
Vascular disease risk	t = −1.31	0.87 ± 0.81 (1.2%)	0.90 ± 0.77 (1.0%)
Any history of:			
cancer	χ^2 = 2.48	40.2% (2.0%)	37.8% (2.0%)
claudication	χ^2 = 19.62***	22.4% (2.0%)	28.2% (2.0%)
diabetes	χ^2 = 0.44	11.6% (2.0%)	12.3% (2.1%)
diabetes medication	χ^2 = 1.97	15.6% (1.2%)	17.2% (1.0%)
head injury with loss of consc.	χ^2 = 0.03	9.7% (2.0%)	9.8% (2.0%)
heart disease	χ^2 = 10.04**	12.7% (2.0%)	16.0% (2.0%)
hypertension	χ^2 = 7.24**	56.9% (2.0%)	61.0% (2.0%)
stroke	χ^2 = 36.05***	9.7% (0.9%)	15.3% (0.5%)
thyroid disease	χ^2 = 1.51	24.1% (2.0%)	25.8% (2.0%)

Methods

We now describe in detail how we produced the results described in this paper. Additionally, our code for reproducing these results is available at: github.com/suinleelab/EEDRP.

Dataset

The Religious Orders Study (ROS)[7] and Memory Aging Project (MAP)[7] are complementary epidemiological studies that each enroll persons without dementia who agree to annual evaluations and eventual organ donation. ROS enrolls clergy living communally from 40 Catholic groups across the US (primarily employed or retired nuns, priests, and brothers). This study group was selected because communal living provided both high follow-up rates and relative consistency in life experiences and socioeconomic factors. However, as a volunteer cohort of Catholic clergy, the samples are not representative of a wider population of elderly individuals[7]. As a complementary study, MAP recruited participants from a wider range of life experiences throughout northeastern Illinois. Participants are primarily enrolled from continuous care retirement communities (ranging in care levels from independent living to nursing on campus). To reduce participant burden and facilitate high follow-up rates, data was collected via home visits. Clinical data collection procedures were consistent between both studies to allow the data to be merged for analyses[7]. Due to their recruitment strategies, follow-up rates of survivors reached around 95% for both studies. Compared to ROS samples, MAP samples were obtained from relatively fewer males (23.6% vs 31.9%) and from individuals who were older (83 vs 80 years on average) and less educated (15 vs 18 years on average). MAP samples also had higher rates of MCI (21.2% vs 19%) and a higher incidence of dementia onset within three years (15.3% vs. 12.7%).

Upon entering the study, participants share demographic information (e.g., sex, age) and blood samples for genotyping. At each yearly visit, they provide updated medical information and undergo a battery of cognitive tests, resulting in repeatedly measured variables. We predict dementia onset from 41 separate variables (per time point; note that categorical variables were one-hot encoded, leading to 49 total features), which we list in Table 4.6. In total, the data contains 3,194 individuals with one to 23 annual visits. Of all participants with at least two years of visit data and no original dementia diagnosis, 619 (23.7%) were eventually diagnosed with dementia.

Data processing: generating samples

Our prediction task (Figure 4.1) is: Using data from his/her three most recent practitioner visits, does an individual with no history of dementia experience dementia onset within the next three years? In particular,

our selected time-frame was based on prior findings that a precipitous drop in cognitive abilities is usually observed one-to-three years prior to a dementia diagnosis[91]. To construct the appropriate dataset, we narrow our analyses from the 3,194 existing participants to 1,597 individuals for whom we have enough observations.

Many participants had more than six consecutive yearly visits, so we applied a sliding window of six years over their available consecutive visits, thereby generating at least one sample, but often more, per participant. Each sample is split into an input window (consisting of the first three consecutive visits *t-2*, *t-1*, and *t*) and onset prediction window (consisting of the next three consecutive visits: *t+1*, *t+2*, and *t+3*), as illustrated by positive (dementia onset) and negative (no dementia onset) examples in Figure 4.8. Because the goal is to predict future dementia onset in individuals who do not yet have dementia, we exclude all samples in which dementia is already present during visits *t-2*, *t-1*, and *t* (e.g., Figure 4.8, Example Participant A, samples 2 and 3). Finally, we applied sliding windows of four and five years to identify any additional positive onset cases (e.g., Figure 4.8, Participant B, Samples 4 and 5), which helped to mitigate our class imbalance. This procedure could not be used to find negative dementia onset samples because all three future years must be known to definitively rule out a dementia diagnosis.

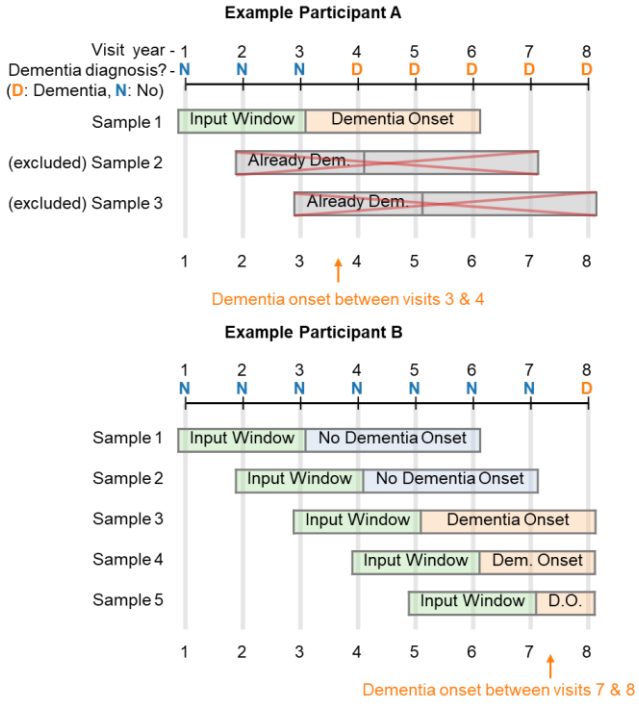


Figure 4.8. Examples of samples from sliding windows. Our samples have no history of dementia during the first 3 years, and either no onset for all of the next 3 years (negative case) or a dementia diagnosis in any of the next 3 years (positive case).

Data processing: pre-processing for all models

After combining all valid six-year windows (and four- and five- year windows where appropriate), we have a sample size of 9,103 samples, of which 13.7% were labeled as positive dementia onset cases (derived from 1,597 individuals, of which 521 developed dementia). For each model next described, our model inputs consist of variables obtained during the first three visits (*t-2*, *t-1*, and/or *t*), called the input data window, and the outputs are a prediction of whether the individual was diagnosed with dementia at any of visits *t+1*, *t+2*, or *t+3*. Table 4.6 shows all demographic, cognitive, and medical features from our 9,103 samples (at time *t*), split by dementia onset label, and associated between-group differences.

Since some downstream analyses require variables to be on the same scale, we standardize all continuous variables for our input data and use z-scores as features for all models. To maintain consistent scores across time points, z-scores are calculated based on time *t* observations (and the same re-scaling procedure based on time *t* is applied to observations at *t-1* and *t-2*). For categorical variables, we apply one-hot encoding. We note in Table 4.6 that most variables have some missing observations across our samples. We impute all missing samples using the mean for continuous variables and the mode for categorical variables (across all samples). Using chi-square tests of independence, we find that some cognitive tests have significantly different missingness rates between dementia onset and control groups, although the rates tend to be low (between 0.2% and 6%, except for Stroop test variables). Additional analyses described in the *Missing Data Experiments* Results sub-section confirm that the effects of imputing values did not impact our final results (compared with filtering out the affected cases).

We next describe model selection with cross-validation and then evaluation on a test set. For each analysis, we use the same stratified training and test sets. To avoid contaminating our test set with training examples, we split our data by participants so that all samples from a single individual fall into the training set or test set, but not both.

Of our 1,597 participants, we assigned one fifth of them to the test set (1,805 associated samples) and the remaining individuals (7,298 associated samples) to our training set. Next, we randomly divide our training set participants into five stratified cross-validation splits. All splits were performed in a stratified manner to maintain consistent ratios of AD to control cases.

Building and evaluating prediction models

We evaluated modeling options under several domains: sampling techniques to address class imbalance, time encoding techniques, and model class. Our modeling choices were based on average accuracy, areas

under the receiver operating curve (AUROC), and areas under the precision recall curve (AUPRC) across five cross-validation (CV) folds.

Downsampling. The dataset has a class imbalance of 13.7% positive labels since few individuals experience dementia onset in any given 3-year window. Therefore, we experimented with four different downsampling techniques: (1) no downsampling, (2) class re-weighting (incorporated into loss functions during training), (3) random downsampling (randomly selecting as many negative as positive samples), and (4) matched pairs downsampling. In (4), for each positive sample, we select the closest negative sample based on sex, age, and education (greedily, without replacement). Due to equal or better prediction performance across five-fold CV, all final models are trained with no downsampling (see Methods).

Time-series encoding. Because of the longitudinal nature of many features, we evaluated methods for incorporating repeatedly observed variables: (1) all data (no special encoding), (2) moving averages, and (3) slopes (see Table 4.7). Training with all data yielded similar or better CV performance, and thus was used for all subsequent models.

Table 4.7. Encoding methods used for time-series features.

Name	**Description**
All data	Unaltered data from *t, t-1, t-2*
Moving averages	(1) Unaltered features from *t*, (2) One simple moving average feature derived from *t, t-1,* and *t-2* features, and (3) Three exponential moving average features with half-life values of 1, 2, and 3 years derived from *t-2* features
Slopes	(1) Unaltered features from *t*, (2) the change in features from each year to the next (i.e., $v_t - v_{t-1}$ and $v_{t-1} - v_{t-2}$ for variable v), and (3) the overall change in scores from the earliest year to the current year (i.e., $v_t - v_{t-2}$)

Model Type. We compared the performance of four classes of ML models: (1) logistic regression (LR; implemented with Scikit-Learn[109], (2) gradient-boosted decision trees via the XGBoost algorithm (XGB; known for handling mixed feature types and medical data well[110]), (3) multi-layer perceptrons (MLP; deep learning approach), and (4) long short-term memory networks (LSTM; time series aware deep learning approach). Both deep learning approaches were implemented in Keras[111] and tensorflow[112]. For each model class, we evaluated several hyperparameter settings and selected the setting with the highest average CV AUROC (reported in Table 4.1). We share our final hyperparameters, along with average CV

performance across modeling choices described in this section, in our code repository: github.com/suinleelab/EEDRP.

Training with fewer input years. We next evaluate whether we can reduce the burden of repeated cognitive testing (i.e., do we need multiple years of data to accurately predict dementia?). We compare performance of models trained on the last 3 year's visits with models trained on fewer time points: the last 2 years' visits (*t* and *t-1*) and the most recent visit (*t*) (circular markers in Figure 4.3). We also evaluate the importance of recent data for impending dementia predictions: in addition to evaluating the model trained on the most recent visit alone (*t*), we also train models on data from single visits one and two years earlier (*t-1* alone and *t-2* alone) (triangular markers in Figure 4.3). Results (Figure 4.3) indicate that recent, but not repeated, measurements are needed for accurate prediction.

To further explore whether the model relies on past data, we perform feature importance analysis using SHAP (described in the next section) on our XGB model trained on the last 3 years of data. The model's top ten features are from time *t* (including demographic features), which provides further evidence that relying on past measurements is not necessary.

Model interpretation with SHAP explanations

To explore what the model is learning and drive further insights, we use SHAP local feature explanations applied to our XGB model (trained on the full feature set with current year, *t*, data). To obtain global feature importances, we aggregate local feature attributions across training samples. Features with higher global importances have more impact on model predictions across samples (Figure 4.4). Next, we select a subset of available features based on their global SHAP ranking: the top 4 demographic features (age, sex, education, APOE genotype) and the top 4 cognitive tests (with their sub-tests; Table 4.3). Our final feature set excludes the variable “No cognitive impairment diagnosis” because it is a cognitive diagnosis that is inefficient to obtain (based on both the full 98-minute neuropsychological battery and a medical review from a physician). Finally, to compare feature selection using SHAP global importances to the more typical global feature selection method in linear models, we use the same demographic features and select the four cognitive tests with the highest-magnitude coefficients from the linear model (Table 4.3).

Measuring final model performance

First, we compare final test performance of XGB trained on the full feature set compared with 2 simplified feature sets: (1) the top four demographic features and the top 4 cognitive tests, and (2) the same set of features but excluding APOE genotype, which may be expensive to obtain for those without existing

genotype data. To compare selected features from SHAP to those from a simple linear method, we also report performance for XGB and LR models trained on the features selected via LR coefficients, described above (Table 4.3).

Finally, for comparison with our methods, we also generate multiple baseline XGB models trained on features commonly used as risk indicators in the literature (Figure 4.5 and Table 4.2): (1) demographic features (above)[6]; (2) MCI diagnosis and demographic variables[101]; (3) the MMSE30 and demographic variables[102]; and (4) the sum over all normalized cognitive test scores controlled for age, sex, and education.

Figure 4.5 displays ROC curves, showing the performance of models at all possible decision cut-off points. We also show confusion matrices for the top-performing baseline and final models using an example cut-off of 0.5 (Table 4.4). Table 4.2 summarizes all performance metrics, including confidence intervals from bootstrap resampling of the test set (repeated 1,000 times). Per Figure 4.5 and Table 4.2, the features selected from the XGB model result in similar AUROCs compared with the full feature set (and outperform the linearly selected features). While the full cognitive battery requires 98 minutes of cognitive testing, we achieve similar predictive value using only four tests that take under 20 minutes.

Examining SHAP explanations in the final model

Feature interactions. To explore the complex interactions learned by the XGB model, we examine SHAP interaction values among pairs of features in our final simplified model. For each sample in our training set, the SHAP interaction value for two features represents the remaining combined feature effect after removing individual main effects of both features.

Figure 4.6 shows feature interactions in the XGB model: each point is a training sample colored by one feature and placed on the x-axis according to its value for the second feature. The y-axis indicates the sample's SHAP interaction value (refer for more detail to Lundberg et al. 2020 [104]). In parts b and c, samples with ages over 90 were censored due to privacy requirements. Higher absolute value y-axis values in these plots indicate that the XGB model makes risk predictions based on feature combinations rather than independently based on single features.

Personalized explanations. For any sample, we can generate a SHAP force plot to explore personalized risk explanations provided by SHAP applied to our final XGB model (e.g., Figure 4.7). These plots indicate both the model's dementia onset risk prediction and the SHAP values for the highest-contributing features impacting the prediction (pink arrows for risk factors, and blue arrows for protective ones).

To clarify the variations in explanations in a controlled setting, we generate four synthetic examples. First, we show a SHAP force plot for a “typical individual” in our dataset (i.e., a sample with mean or mode values for all features; Figure 4.7a). A “typical individual" has a low risk of developing dementia in the next three years since the diagnostic rate for dementia is low in any 3-year period. Next, we show perturbations to single feature values for APOE (where we change the APOE e4 allele count from zero to one; Figure 4.7b) and word list delayed recall (WLDR) score (from the mean value to two standard deviations below the mean, i.e., from six words remembered to just one; Figure 4.7c). Finally, we simultaneously perturb both risk factors above and show that the combined risk of having both one APOE e4 allele and a low WLDR score leads to a large, non-linear jump in risk that exceeds the combined single effects of each feature alone (Figure 4.7d).

Chapter 5. Explanation-guided dynamic feature selection for medical risk prediction

In medical risk prediction scenarios, machine learning methods have demonstrated an ability to learn complex and predictive relationships among rich feature sets. However, in practice, when faced with new patients, we may not have access all information expected by a trained risk model. We propose a framework to simultaneously provide flexible risk estimates for samples with missing features, as well as context-dependent feature recommendations to identify what piece of information may be most valuable to collect next. Our approach uses a fixed prediction model, a local feature explainer, and ensembles of imputed samples to generate risk prediction intervals and feature recommendations. Applied to a myocardial infarction risk prediction task in the UK Biobank dataset, we find that our approach can more efficiently predict risk of a heart attack with fewer observed features than traditional fixed imputation and global feature selection methods.[§]

Introduction

In many medical decision-making scenarios, there is a tradeoff between collecting more features at a cost, versus making a decision about a patient based on what is currently known. In general, observing more information about the person can lead to more accurate and confident predictions, and current machine learning (ML) prediction models often expect a rich feature set, which may not be available at all times in practice. To that end, previous work has focused on designing models that can efficiently choose features dynamically, tailored to the context of a specific sample; however, they often rely on specific modeling/architecture innovations (e.g., decision trees[113,114]). Another line of research has involved sensitivity based approaches that aim to select features based on the unknown features' influence on model predictions predictions[115–117]; however, these approaches have also tended rely on a specific class of model (linear models or neural networks[118,119]). For a fixed supervised ML model (about which we make minimal assumptions), we propose a general framework to simultaneously provide both flexible risk estimates for individuals with missing features and personalized feature recommendations to identify which missing features may be most informative to select next given the context.

Our approach (Figure 5.1) relies on three main components: (1) a conditional feature imputer for sampling possible values of missing features given the observed ones, (2) a fixed supervised ML model, and (3) local feature explainer for the model. Empirically, we find that an approach using KNN-based

[§] This project was done in collaboration with Wei Qiu and Su-In Lee

imputation, an XGBoost model, and a SHAP explainer is able to more efficiently predict 10-year risk for myocardial infarction than traditional single value imputation and fixed (global) feature selection.

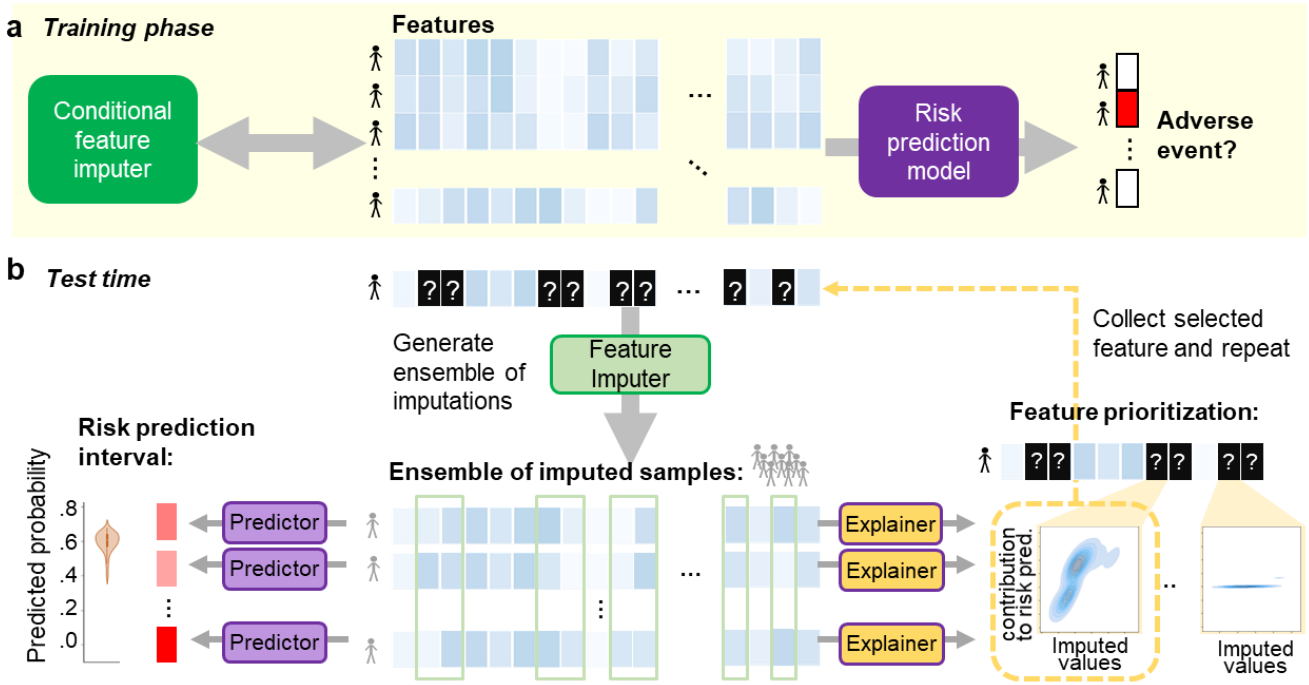


Figure 5.1. Overview of our approach. (a) Our method relies on a fixed conditional feature imputer and risk prediction model fit to a training set. (b) For each incomplete sample at test time, we generate an ensemble of imputed samples and use the prediction model to produce a risk interval (left). The next feature is dynamically selected based on model explanations, and if observed, the process may be repeated.

Methods

For our approach, given an individual with partially observed features (e.g., a sparsely populated medical record), we stochastically impute their missing features–ideally, drawing from the conditional distribution of their missing features given observed ones–to generate examples of a complete record given the currently available information. We then use this "ensemble" of imputations to: (1) provide a risk interval (rather than just a single point estimate) around the adverse event, and (2) select a feature to collect next, guided by model explanations, which may best help reduce uncertainty with respect to the model's prediction. Such an approach may enable clinicians to make better informed decisions by providing them with flexible risk intervals (regardless of how much information is already known about the patient) and suggesting follow-up tests to improve their understanding of a patient's risk profile.

Our approach is model agnostic, and relies on three main components: (1) A conditional feature imputer for generating ensembles of imputed samples. Our goal for the stochastic imputer is to conditionally sample missing features given observed ones such that, by sampling an "ensemble" of multiple imputations

for a given individual, we obtain a distribution of possible complete samples for that person. These imputation ensembles could then be fed into the predictor and explainer to generate a distribution of predicted risk scores and explanations conditioned on the observed context. In practice, developing generative models for conditional imputation is challenging, and an area of active research (e.g., neural network methods such as GAIN[120] and GI[121]; however, we find that a simple approach of sampling nearest neighbors works well in practice (see Results). (2) A supervised ML model. For our approach, we assume that the model has already been trained on a labeled dataset consisting of a fully observed feature set and the clinical label of interest. We make minimal assumptions about the type of model used (for our experiments, we use an XGBoost[110], and only require that the model has an associated (3) feature explainer, which can estimate each feature's contribution to the model's prediction for a given sample (for our experiments, we use SHAP[105]).

For our experiments, we used a fixed training set for all components. However, we note that because the stochastic imputer and prediction model are trained independently, their training sets do not need to be identical. This may be particularly advantageous if there are many more unlabeled samples available (which may be used to train the imputer).

Putting it all together: Our approach

At test time (Figure 5.1b), from the imputation ensemble generated by the conditional feature imputer, we simply use the risk model's predictions to generate a risk distribution (whose spread reflects variation in the model's output with respect to imputed features). We further propose using the distributions of feature attributions to inform feature recommendations. In particular, we hypothesize that using variance in SHAP values across the imputation ensemble will be an effective metric for selecting the next feature. Intuitively, large variations in SHAP values would indicate that the model's predictions are sensitive to our simulated variations in the missing features. In contrast, if a feature has high variance in the imputed values but not SHAP values, that would indicate that these variations are not relevant–according to the prediction model–to risk. Similarly, a feature with high-magnitude but low-variance SHAP values may also be a poor choice, since given the current context, we are already confident about how the unknown feature would impact the model (for example, if we have two redundant features and have already observed one of them).

Experiments

We first evaluate our approach on a toy dataset with a known conditional distribution where we can compare feature selection approaches independently of imputation methods. We then apply our approach to data from the UK Biobank (www.ukbiobank.ac.uk), a biomedical database containing data from individuals

across the UK, including hundreds of features collected during an initial visit and detailed long-term health outcomes. For our analyses, we focus on a randomly selected subset of 100,000 samples, along with 252 features (described further in Supplementary Methods), and we observe that about 2% of individuals have a reported myocardial infarction within 10 years after their initial screening.

Results

In Supplementary Methods, we show the advantages of our dynamic feature selection approach on a synthetic dataset where we directly impute missing features from a known conditional distribution, and thus can evaluate our feature selection approach in isolation. As shown in Supplementary Figure 5.1, compared with a fixed global ordering, our explanation-guided dynamic feature selection strategy more efficiently identifies relevant features given context from observed ones.

We now turn to the task of 10-year myocardial infarction prediction in the UK Biobank (UKB) dataset. We first fit a supervised ML model and imputer to our training set. We then use a test set to simulate an interactive process in which we alternate between (1) generating a prediction given the current observed information, and (2) selecting the next feature to un-mask. We then observe how the prediction model's average performance and prediction variation progress as features are selected and observed.

Initial model training

Supervised risk model and explainer. For the 10-year myocardial infarction prediction task, we fit an XGBoost model to the training set to establish baseline performance expectations for a complete feature set, for which we obtain test AUROC and AP scores of 0.768 and 0.088, respectively (Supplementary Figure 5.2). We use SHAP as our model's explainer, and note that the top 20 features in the training set ordered by mean absolute SHAP value (listed in Figure 5.3b and shown in Supplementary Figure 5.2c) cover a range of feature types, from demographic features to medical lab tests.

Conditional feature imputation. As described in Methods, our approach relies on a conditional feature imputer which samples missing features conditioned on the observed ones. We empirically find that a k-nearest neighbors (KNN)-based approach, with an ensemble size of 100, works well in practice (Supplementary Methods). Across different rates of induced missingness, we find that this approach leads to the most accurate risk predictions (averaged across our XGBoost model's outputs for the ensemble of imputed samples) compared with standard approaches such as mean-value imputation and imputing from the marginal distribution, or recently proposed deep learning imputation models[120,121] (Supplementary Figure 5.2b).

Iterative feature selection experiments

We now demonstrate the value of our approach for iteratively predicting risk intervals and dynamically selecting features. For each test sample, we run the following procedure (illustrated in Figure 2) beginning with 0 observed features, until 20 features have been observed: (1) Generate an ensemble of 100 imputed samples using the KNN-based approach described above, (2) use the supervised risk model to generate predictions for each of the 100 imputed samples, resulting in a predicted "risk interval", (3) apply SHAP to obtain feature contribution scores for each feature across each imputed sample, and finally, (4) choose the feature with the highest SHAP value variance and uncover the true value of that feature.

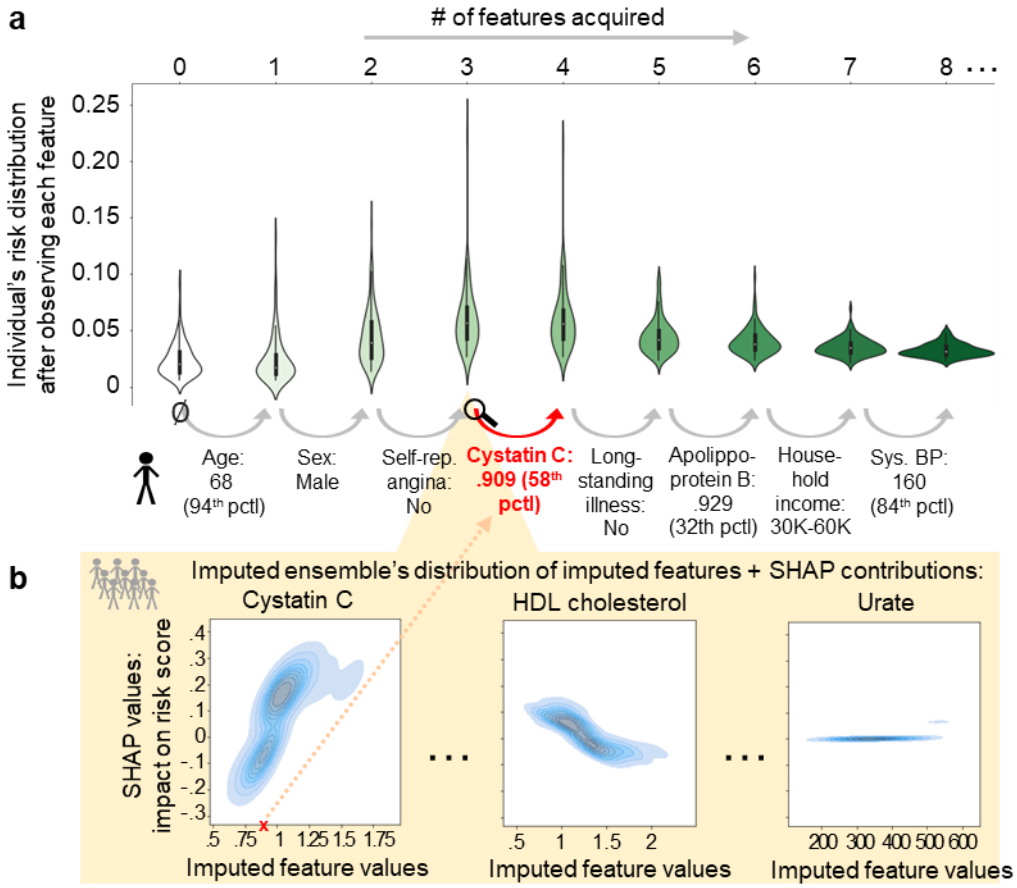


Figure 5.2. An example of how our approach is applied to a single person in the UKB cohort. (**a**) We show how the distribution in risk scores predicted from imputation ensembles changes as each successive feature is collected. (**b**) Zooming into a specific time-point (here, we know just that the individual is a 68-year old male without self-reported angina), we show examples of imputed values vs. their resulting SHAP values for 3 candidate features. While we use the imputed features (x-axis) for generating the distribution of model predictions shown in part **a**, we use the distribution of SHAP values to select the feature with the highest SHAP-value variance.

To illustrate the potential use of our approach in practice, we show an example for a single participant in the UKB study in Figure 5.2. In Figure 5.2a, we first demonstrate the individual's prediction interval evolves as each new feature is collected (resulting in a new ensemble of imputations, and subsequent risk scores). As more features are observed, we see that the prediction intervals tend to shrink indicating that uncertainty of the model predictions with respect to missing features' imputations is decreasing. In Figure 5.2b, we show an example of imputed values vs. their contribution to the model's risk scores (according to SHAP values) after observing three initial features. In this particular example, imputed values for Cystatin C have the highest-varying impact on model predictions given the context at that point, and is thus selected next. Our approach, as highlighted in this example, provides an explanation-guided recommendation (based on ensembles of imputations) for features, and thus may provide users with context-dependent insight into why a feature may be useful to observe.

In Figure 5.3a, we report the performance of our approach, aggregated across the test set, as features are iteratively selected and observed. For comparison, we consider two baselines using fixed feature orderings provided by our global feature ranking (based on mean absolute SHAP value; Supplementary Figure 5.2c). First, we consider the same procedure described above, but replace the dynamic feature selection strategy (steps 3 and 4) with simply choosing features in the fixed global order. Second, a more traditional approach of using global feature selection involves re-training the prediction model on the selected features. Thus, we also consider a collection of models trained with feature budget (i.e., a single-feature model trained on age, a two-feature model trained on age and sex, ..., and a 20-feature model trained on the top 20 features), as such a model is tailored to the specific expected use-case.

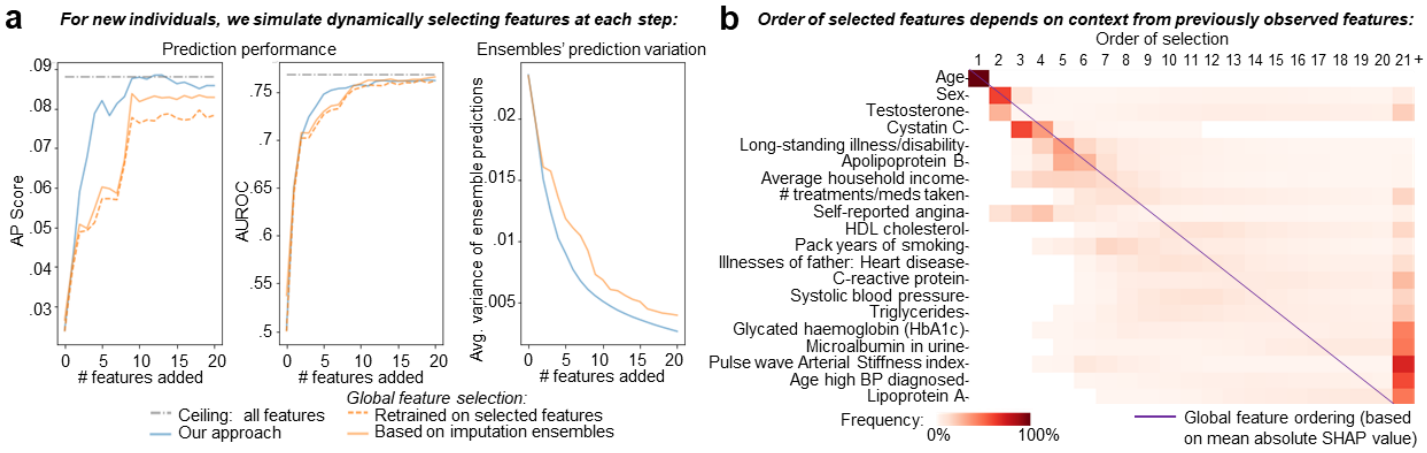


Figure 5.3. (a) Comparison between our dynamic feature selection approach and global feature selection for 10-year myocardial infarction prediction in the UKB dataset. (b) Overview of our feature selection orders compared with fixed global feature selection (in y-axis).

From our experiments, we find that our approach more efficiently prioritizes features than fixed global feature selection strategies, leading to improved prediction accuracy with fewer features observed (Figure 5.3a). In Figure 5.3b, we show that our approach does indeed lead to different feature orderings tailored to samples' context of previously observed features. We see that age is always selected first, but that the participant's age informs the choice of the second feature, whose value informs the next feature choice, and so on. Supplementary Figure 5.3 provides a detailed view of feature orderings across the test set. Anecdotally, we note that for younger individuals, our approach tends to choose testosterone next; sex is often not selected till much later (perhaps because sex can be easily inferred from the testosterone values and thus would provide redundant information).

Discussion

In this work, we propose an approach to leverage conditional imputation and explainability methods to provide flexible health risk estimates and context-dependent feature recommendations for a fixed ML prediction model. Applied to a 10-year myocardial infarction prediction task in the UKB dataset, our approach (implemented with KNN-based imputation, XGBoost, and SHAP) led to more efficient feature prioritization compared with a static approach and single value imputation. One key consideration is that our method relies on having a reliable way to conditionally impute missing features. While our KNN-based approach worked well in practice, it is not guaranteed to be effective, and may be particularly limited in the case of small datasets where neighborhoods may be particularly sparse.

There are several natural extensions of our work to consider. First, our feature selection strategy does not take into account the fact that those features tend to have varying costs. One simple extension is to adjust our feature collection policy to balance our current metric (i.e., SHAP variance) with the cost of the feature. Second, our experiments considered collecting features over a fixed budget of 20 features. In practice, it may be valuable to leverage some notion of uncertainty (contained in risk prediction intervals) in a triage setting, where we may allocate a feature budget unevenly across samples (e.g., terminating a sample's collection procedure once its risk interval is sufficiently narrow).

Although extensive experimentation is needed to validate our method in a real-world setting, initial experiments demonstrate its potential to provide flexible risk estimates despite incomplete medical information, along with context dependent feature recommendations which may aid clinicians in choosing tests to gain additional clarity.

Chapter 6. Conclusion

In this work, we described projects, ranging from systems biology to healthcare applications, with a common goal of leveraging machine learning and explainability methods to better predict and interpret complex biomedical problems in the face of various real-world data limitations. With our MD-AD framework (Chapter 2), we jointly modeled gene expression and multiple related AD neuropathologies, performing the first multi-cohort neural network analysis of GE data in AD. We demonstrated the ability of a joint-learning approach to effectively capture relationships among gene expression and neuropathology, and were able to further interrogate our model with integrated gradients to clarify nuanced sex-specific relationships between immune response genes and AD. Inspired by interpretation challenges from MD-AD, in our PAC project (Chapter 3), we developed a method and tool for understanding relationships among biological pathway gene sets across several databases, which may help computational biologists to contextualize and interpret their biological findings.

The computational systems biology field is growing rapidly, and alongside advancements in computational methods, emerging biotechnologies (e.g., spatial transcriptomics) are enabling the study of cells from new perspectives. While our analyses focused on a single modality of bulk RNA-sequencing data, ML methods will need to evolve to meet the unique challenges of these new emerging forms of data (e.g., as proposed by Weinberger*, Lin* & Lee[122]), and may be particularly useful in helping to unify views across several modalities collected together[123]. For our projects, we also note that we focused on a cases with limited labels across datasets; yet, in some cases, labels may be altogether unavailable. Nevertheless, researchers may want to identify and characterize underlying sources of variation in fully unlabeled datasets, and new ML and explanation methods, such as work by Janizek et al.[124], are already progressing towards such goals.

In the area of clinical ML, in Chapter 4, we presented a method to predict imminent dementia onset with fewer time points and features than previously reported. By identifying globally important features, we were able to identify a small subset (one-fifth of the time required for a full standard neuropsychological battery), thus reducing the burden of clinical testing and possibly creating a more scalable approach for the general population. However, one key limitation of this global feature selection approach is that such a method does not accommodate missingness in the requested features, nor does it take advantage of already known features that were not selected by the global feature selection method. In Chapter 5, we proposed an approach to leverage sparse pre-existing information about a person to provide health risk estimates as well as context-dependent feature recommendations. For both of these projects, we assume that there exists a fixed ML model that has already been trained on a rich feature set, and use post-hoc methods to adapt the

method for limited resources at test time. While the ability to apply our methods post-hoc to an existing risk models may be advantageous when a fixed model is already in use, complementary lines of research have explored the use of modified architectures or approaches to train more flexible models in the first place[118,119]. Joint progress in both of these directions may lead to the largest strides in the field of efficient and flexible risk estimation. Finally, adoption of clinical ML models in practice has been notoriously limited, and will hinge on effective collaboration between ML researchers and clinicians to produce tools that are truly helpful[125].

Together, these projects – spanning in applications from systems biology to healthcare – illustrate the power of adapting machine learning and explainability methods towards the unique specifications of biomedical applications where limited data and resources preclude the out-of-the-box use of more traditional ML methods. In the coming years, rapid progress in biotechnology and healthcare will lead to new advancements, and also new challenges to address. I am grateful to my wonderful collaborators for joining and guiding my exploration of these topics, and look forward to seeing and contributing to the exciting future of AI for biomedicine.

Bibliography

1. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proc. IEEE Int. Conf. Comput. Vis.* 1026–1034 (2015).

2. Popel, M. *et al.* Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nat. Commun.* **11**, 1–15 (2020).

3. Devlin, J., Chang, M.-W., Lee, K., Google, K. T. & Language, A. I. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proc. NAACL-HLT 2019* (2019).

4. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).

5. Gulshan, V. *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).

6. Bennett, D. A., Schneider, J. A., Arvanitakis, Z. & Wilson, R. S. Overview and findings from the religious orders study. *Curr. Alzheimer Res.* **9**, (2012).

7. Bennett, D. A. *et al.* Overview and Findings from the Rush Memory and Aging Project. *Curr. Alzheimer Res.* **9**, 646–663 (2012).

8. Mostafavi, S. *et al.* A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. *Nat. Neurosci.* **21**, 811–819 (2018).

9. Miller, J. A. *et al.* Neuropathological and transcriptomic characteristics of the aged brain. *Elife* **6**, 1–26 (2017).

10. Wang, M. *et al.* The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Sci. Data* **5**, 1–16 (2018).

11. Nevins, J. R. & Potti, A. Mining gene expression profiles: Expression signatures as cancer phenotypes. *Nat. Rev. Genet.* **8**, 601–609 (2007).

12. Oshlack, A., Robinson, M. D. & Young, M. D. Oshlack_From RNA-seq reads to differential expression results_Genome Biol (2010). *Genome Biol.* **11**, 1–10 (2010).

13. Lundberg, S. M. *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).

14. De Jager, P. L., Yang, H. S. & Bennett, D. A. Deconstructing and targeting the genomic architecture of human neurodegeneration. *Nat. Neurosci.* **21**, 1310–1317 (2018).

15. Gaiteri, C., Mostafavi, S., Honey, C. J., De Jager, P. L. & Bennett, D. A. Genetic variants in Alzheimer disease-molecular and brain network approaches. *Nat. Rev. Neurol.* **12**, 413–427 (2016).

16. Marioni, R. E. *et al.* GWAS on family history of Alzheimer's disease. *Transl. Psychiatry* **8**, 0–6 (2018).

17. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).

18. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).

19. Zhang, B. *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707–720 (2013).

20. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer ' s disease. *Nature* **570**, 332–337 (2019).

21. Logsdon, B. A. *et al.* Meta-analysis of the human brain transcriptome identifies heterogeneity across human AD coexpression modules robust to sample collection and methodological approach. (2019). doi:10.7303/syn17114455

22. Blalock, E. M. *et al.* Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 2173–2178 (2004).

23. Katsel, P., Li, C. & Haroutunian, V. Gene expression alterations in the sphingolipid metabolism pathways during progression of dementia and Alzheimer's disease: A shift toward ceramide accumulation at the earliest recognizable stages of Alzheimer's disease? *Neurochem. Res.* **32**, 845–856 (2007).

24. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).

25. Caruana, R. A. Multitask Learning: A Knowledge-Based Source of Inductive Bias. *Mach. Learn. Proc. 1993* 41–48 (1993). doi:10.1016/b978-1-55860-307-3.50012-5

26. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

27. Westermann, F. *et al.* Classification and diagnostic prediction of pediatric cancers using gene expression profiling and artificial neural networks. *GBM Annu. Fall Meet. Halle 2002* **2002**, 673–679 (2002).

28. Lee, T. & Lee, H. Prediction of Alzheimer's disease using blood gene expression data. *Sci. Rep.* **10**, 1–13 (2020).

29. Pirooznia, M., Yang, J. Y., Qu, M. Q. & Deng, Y. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics* **9**, 1–13 (2008).

30. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. *34th Int. Conf. Mach. Learn. ICML 2017* **7**, 5109–5118 (2017).

31. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

32. Mirra, S. S. *et al.* The consortium to establish a registry for Alzheimer's disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. *Neurology* **41**, 479–486 (1991).

33. Braak, H. & Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* **82**, 239–259 (1991).

34. Allen, M. *et al.* Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Sci. Data* **3**, 1–10 (2016).

35. Matarin, M. *et al.* A Genome-wide gene-expression analysis and database in transgenic mice

during development of amyloid or tau pathology. *Cell Rep.* **10**, 633–644 (2015).

36. Cummings, D. M. *et al.* First effects of rising amyloid-β in transgenic mouse brain: Synaptic transmission and gene expression. *Brain* **138**, 1992–2004 (2015).

37. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

38. Daly, M. J. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).

39. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

40. Qiu, Y.-Q. KEGG Pathway Database. in *Encyclopedia of Systems Biology* (eds. Dubitzky, W., Wolkenhauer, O., Cho, K.-H. & Yokota, H.) 1068–1069 (Springer New York, 2013). doi:10.1007/978-1-4419-9863-7_472

41. Trabzuni, D. *et al.* Widespread sex differences in gene expression and splicing in the adult human brain. *Nat. Commun.* **4**, (2013).

42. Olah, M. *et al.* A transcriptomic atlas of aged human microglia. *Nat. Commun.* **9**, 1–8 (2018).

43. Chan, G. *et al.* CD33 modulates TREM2: Convergence of Alzheimer loci. *Nat. Neurosci.* **18**, 1556–1558 (2015).

44. Wang, Y. *et al.* TREM2 lipid sensing sustains the microglial response in an Alzheimer's disease model. *Cell* **160**, 1061–1071 (2015).

45. Mathys, H. *et al.* Temporal Tracking of Microglia Activation in Neurodegeneration at Single-Cell Resolution. *Cell Rep.* **21**, 366–380 (2017).

46. Olah, M. *et al.* Single cell RNA sequencing of human microglia uncovers a subset associated with Alzheimer's disease. *Nat. Commun.* **11**, (2020).

47. Lovestone, S. *et al.* AddNeuroMed - The european collaboration for the discovery of novel biomarkers for alzheimer's disease. *Ann. N. Y. Acad. Sci.* **1180**, 36–46 (2009).

48. Sood, S. *et al.* A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status. *Genome Biol.* **16**, 1–17 (2015).

49. Johnson, E. C. B. *et al.* Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nat. Med.* **26**, 769–780 (2020).

50. Lambert, J. C. *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat. Genet.* **41**, 1094–1099 (2009).

51. Farfel, J. M. *et al.* Relation of genomic variants for Alzheimer disease dementia to common neuropathologies. *Neurology* **87**, 489–496 (2016).

52. Chibnik, L. B. *et al.* CR1 is associated with amyloid plaque burden and age-related cognitive decline. **69**, 560–569 (2011).

53. Thambisetty, M. *et al.* Effect of complement CR1 on brain amyloid burden during aging and its modification by APOE genotype. *Biol. Psychiatry* **73**, 422–428 (2013).

54. Patrick, E. *et al.* A cortical immune network map identifies distinct microglial transcriptional programs associated with beta-amyloid and Tau pathologies. *Press*

55. Wan, Y. W. *et al.* Meta-Analysis of the Alzheimer's Disease Human Brain Transcriptome and Functional Dissection in Mouse Models. *Cell Rep.* **32**, (2020).

56. Wang, M. *et al.* Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. *Genome Med.* **8**, 1–21 (2016).

57. Mirra, S. S. *et al.* The Consortium to Establish a Registry for Alzheimer's Disease (CERAD): Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. *Neurology* **41**, (1991).

58. Basavegowda, H. S. & Dagnew, G. Deep learning approach for microarray cancer data classification. *CAAI Trans. Intell. Technol.* **5**, 22–33 (2020).

59. Zhang, D., Zou, L., Zhou, X. & He, F. Integrating Feature Selection and Feature Extraction Methods with Deep Learning to Predict Clinical Outcome of Breast Cancer. *IEEE Access* **6**, 28936–28944 (2018).

60. Fakoor, R., Ladhak, F., Nazi, A. & Huber, M. Using deep learning to enhance cancer diagnosis and classification. *30th Int. Conf. Mach. Learn. WHEALTH Work.* (2013).

61. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 20170387 (2018).

62. Latimer, C. S. *et al.* Resistance and resilience to Alzheimer's disease pathology are associated with reduced cortical pTau and absence of limbic-predominant age-related TDP-43 encephalopathy in a community-based cohort. *Acta Neuropathol. Commun.* **7**, 9 (2019).

63. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).

64. Olah, M. *et al.* Single cell RNA sequencing of human microglia uncovers a subset that is associated with Alzheimer's disease. (2020).

65. Viswanathan, G. A., Seto, J., Patil, S., Nudelman, G. & Sealfon, S. C. Getting Started in Biological Pathway Construction and Analysis. *PLoS Comput. Biol.* **4**, (2008).

66. Carbon, S. *et al.* The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

67. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951 (2019).

68. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595 (2019).

69. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

70. Belinky, F. *et al.* PathCards: Multi-source consolidation of human biological pathways. *Database* **2015**, 1–13 (2015).

71. Liberzon, A., Birger, C., Ghandi, M., Mesirov, J. P. & Tamayo, P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).

72. Harris, M. A. *et al.* The Gene Oncology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, 258–261 (2004).

73. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: A network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **5**, (2010).

74. Wang, L. L. & Gennari, J. H. Similarity metrics for determining overlap among biological pathways. *CEUR Workshop Proc.* **2137**, 1–6 (2017).

75. Li, Y., Agarwal, P. & Rajagopalan, D. A global pathway crosstalk network. *Bioinformatics* **24**, 1442–1447 (2008).

76. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

77. Raymond, M. & Rousset, F. An Exact Test for Population Differentiation. *Evolution (N. Y).* **49**, 1280 (1995).

78. De Meo, P., Ferrara, E., Fiumara, G. & Provetti, A. Generalized Louvain method for community detection in large networks. *Int. Conf. Intell. Syst. Des. Appl. ISDA* 88–93 (2011). doi:10.1109/ISDA.2011.6121636

79. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, (2008).

80. Gruvberger, S. *et al.* Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.* **61**, 5979–5984 (2001).

81. Aaltomaa, S. *et al.* Hormone receptors as prognostic factors in female: Breast cancer. *Ann. Med.* **23**, 643–648 (1991).

82. Parl, F. F., Schmidt, B. P., Dupont, W. D. & Wagner, R. K. Prognostic significance of estrogen receptor status in breast cancer in relation to tumor stage, axillary node metastasis, and histopathologic grading. *Cancer* **54**, 2237–2242 (1984).

83. Sotiriou, C. & Pusztai, L. Gene-Expression Signatures in Breast Cancer. *N. Engl. J. Med.* **360**, 790–800 (2009).

84. 2020 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **16**, 391–460 (2020).

85. M. Prince, R. Bryce, and C. Ferri. World Alzheimer Report 2011: The benefits of early diagnosis and intervention. *Alzheimer's Dis. Int.* 1–68 (2011).

86. Folstein, M., Folstein, S. & Mchugh, P. Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* **12**, 189–198 (1975).

87. Suk, H. Il & Shen, D. Deep learning-based feature representation for AD/MCI classification. *Int. Conf. Med. image Comput. Comput. Interv.* 583–590 (2013). doi:10.1007/978-3-642-40763-5_72

88. Suk, H. Il, Lee, S. W. & Shen, D. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *Neuroimage* **101**, 569–582 (2014).

89. Cui, R. & Liu, M. RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. *Comput. Med. Imaging Graph.* **73**, 1–10 (2019).

90. Lee, S. *et al.* Episodic memory performance in a multi-ethnic longitudinal study of 13,037 elderly. *PLoS One* **13**, 1–17 (2018).

91. Johnson, D. K., Storandt, M., Morris, J. C. & Galvin, J. E. Longitudinal study of the transition from healthy aging to Alzheimer disease. *Arch. Neurol.* **66**, 1254–1259 (2009).

92. Barnes, D. E. *et al.* Predicting risk of dementia in older adults: The late-life dementia risk index. *Neurology* **73**, 173–179 (2009).

93. Gaugler, J., James, B., Johnson, T., Marin, A. & Weuve, J. 2019 Alzheimer's Disease Facts and Figures. *Alzheimer's Dement.* **15**, 321–387 (2019).

94. Chouraki, V. *et al.* Evaluation of a Genetic Risk Score to Improve Risk Prediction for Alzheimer's Disease. *J. Alzheimer's Dis.* **53**, 921–932 (2016).

95. Naj, A. C. *et al.* Age-at-Onset in Late Onset Alzheimer Disease is Modified by Multiple Genetic Loci. *JAMA Neurol.* **71**, 1394–1404 (2014).

96. Liu, C. C., Kanekiyo, T., Xu, H. & Bu, G. Apolipoprotein e and Alzheimer disease: Risk, mechanisms and therapy. *Nat. Rev. Neurol.* **9**, 106–118 (2013).

97. Stephan, B. C. M., Kurth, T., Matthews, F. E., Brayne, C. & Dufouil, C. Dementia risk prediction in the population: Are screening models accurate? *Nat. Rev. Neurol.* **6**, 318–326 (2010).

98. Arabi, Z., Syed Abdul Rahman, S. A., Hazmi, H. & Hamdin, N. Reliability and construct validity of the Early Dementia Questionnaire (EDQ). *BMC Geriatr.* **16**, 1–10 (2016).

99. Tombaugh, T. N. & McIntyre, N. J. The mini-mental state examination: a comprehensive review. *J. Am. Geriatr. Soc.* **40**, 922–935 (1992).

100. Nasreddine, Z. S. *et al.* The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* **53**, 695–699 (2015).

101. Bozoki, A., Giordani, B., Heidebrink, J. L., Berent, S. & Foster, N. L. Mild cognitive impariments predict dementia in nondemented elderly patients with memory loss. *Arch. Neurol.* **58**, 411–416 (2001).

102. Hogan, D. B. & Ebly, E. M. Predicting who will develop dementia in a cohort of Canadian seniors. *Can. J. Neurol. Sci.* **27**, 18–24 (2000).

103. Goerdten, J., Čukić, I., Danso, S. O., Carrière, I. & Muniz-Terrera, G. Statistical methods for dementia risk prediction and recommendations for future work: A systematic review. *Alzheimer's Dement. Transl. Res. Clin. Interv.* **5**, 563–569 (2019).

104. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, (2020).

105. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017-Decem**, 4766–4775 (2017).

106. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* **21**, 128–138 (2010).

107. Belleville, S., Fouquet, C., Hudon, C., Zomahoun, H. T. V. & Croteau, J. Neuropsychological Measures that Predict Progression from Mild Cognitive Impairment to Alzheimer's type dementia in Older Adults: a Systematic Review and Meta-Analysis. *Neuropsychol. Rev.* **27**, 328–353 (2017).

108. Hensel, A. *et al.* Does a reliable decline in Mini Mental State Examination total score predict dementia? Diagnostic accuracy of two reliable change indices. *Dement. Geriatr. Cogn. Disord.* **27**,

50–58 (2009).

109. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

110. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* (2016).

111. Chollet, F. Keras.

112. Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems. (2015).

113. Xu, Z. E., Kusner, M. J., Weinberger, K. Q., Chen, M. & Chapelle, O. Classifier cascades and trees for minimizing feature evaluation cost. *J. Mach. Learn. Res.* **15**, 2113–2144 (2014).

114. Viola, P. & Jones, M. J. Robust Real-Time Face Detection. *Int. J. Comput. Vis.* **57**, 137–154 (2004).

115. Early, K., Fienberg, S. E. & Mankoff, J. Test time feature ordering with FOCUS: Interactive predictions with minimal user burden. *UbiComp 2016 - Proc. 2016 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput.* 992–1003 (2016). doi:10.1145/2971648.2971748

116. Early, K., Fienberg, S. & Mankoff, J. Cost-effective feature selection and ordering for personalized energy estimates. *AAAI Work. Artif. Intell. Smart Grids Smart Build.* **WS-16-01-**, 226–232 (2016).

117. Kachuee, M., Darabi, S., Moatamed, B. & Sarrafzadeh, M. Dynamic Feature Acquisition Using Denoising Autoencoders. *IEEE Trans. Neural Networks Learn. Syst.* **30**, 2252–2262 (2019).

118. Ma, C. *et al.* EdDI: Efficient dynamic discovery of high-value information with partial VAE. *36th Int. Conf. Mach. Learn. ICML 2019* 7483–7504 (2019).

119. Covert, I. *et al.* Learning to Maximize Mutual Information for Dynamic Feature Selection. (2023).

120. Yoon, J., Jordon, J. & Van Der Schaar, M. GAIN: Missing data imputation using generative adversarial nets. *35th Int. Conf. Mach. Learn. ICML 2018* **13**, 9042–9051 (2018).

121. Kachuee, M., Karkkainen, K., Goldstein, O., Darabi, S. & Sarrafzadeh, M. Generative Imputation and Stochastic Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **8828**, 1–1 (2020).

122. Weinberger, E., Lin, C. & Lee, S.-I. Isolating salient variations of interest in single-cell data with contrastiveVI. *bioRxiv* 2021.12.21.473757 (2022).

123. Cai, Z., Poulos, R. C., Liu, J. & Zhong, Q. Machine learning for multi-omics data integration in cancer. *iScience* **25**, 103798 (2022).

124. Janizek, J. D. *et al.* Principled feature attribution for unsupervised gene expression analysis. *Genome Biol.* 2022.05.03.490535 (2022). doi:10.1186/s13059-023-02901-4

125. Henry, K. E. *et al.* Human–machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. *npj Digit. Med.* **5**, 1–6 (2022).

126. Rivals, I., Personnaz, L., Taing, L. & Potier, M. C. Enrichment or depletion of a GO category within a class of genes: Which test? *Bioinformatics* **23**, 401–407 (2007).

127. Covert, I. C., Lundberg, S. & Lee, S. I. Understanding global feature contributions with additive importance measures. *Adv. Neural Inf. Process. Syst.* **2020-Decem**, (2020).

128. Stekhoven, D. J. & Bühlmann, P. Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).

129. Beebe-Wang, N. *et al.* Unified AI framework to uncover deep interrelationships between gene expression and Alzheimer's disease neuropathologies. *Nat. Commun.* **12**, 1–17 (2021).

130. Beebe-Wang, N., Dincer, A. B. & Lee, S.-I. An automatic integrative method for learning interpretable communities of biological pathways. *NAR Genomics Bioinforma.* **4**, 1–10 (2022).

131. Beebe-Wang, N., Okeson, A., Althoff, T. & Lee, S. I. Efficient and Explainable Risk Assessments for Imminent Dementia in an Aging Cohort Study. *IEEE J. Biomed. Heal. Informatics* **25**, 2409–2420 (2021).

Supplementary Materials

Chapter 2 Supplementary Materials

Supplementary Table 2.1. Descriptions of the six neuropathological phenotypes used for training MD-AD (and baseline models). We provide descriptions for each phenotype directly from documentation from each individual cohort study.

Phenotype	**ACT**	**MSBB**	**ROSMAP**
CERAD Score	Semiquantitative measure of neuritic plaques. A neuropathologic diagnosis was made of no AD, possible AD, probable AD, or definite AD based on semiquantitative estimates of neuritic plaque density as recommended by the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) [24]		
Neuritic Plaques	NA	"Mean neocortical plaque density across 5 regions (# of plaques/mm2)"	"Neuritic plaques (determined by microscopic examination of silver-stained slides from 5 regions: midfrontal cortex, midtemporal cortex, inferior parietal cortex, entorhinal cortex and hippocampus. Each regional count is scaled by dividing by the corresponding standard deviation."
Aβ-IHC	"Beta-amyloid immunohistochemistry (fresh frozen tissue): areal density." Measurements were available for each structure with corresponding expression data.	NA	"Amyloid beta protein identified by molecularly-specific immunohistochemistry and quantified by image analysis. Value is percent area of cortex occupied by amyloid beta." Used measurement from midfrontal cortex, closest available to DLPFC.
Braak Stage	Semiquantitative measure of neurofibrillary tangle spread across the brain. [25]		
Tangles	NA	NA	"Neurofibrillary tangles (nft) by microscopic examination of silver-stained slides from 5 regions midfrontal cortex, midtemporal cortex, inferior parietal cortex, entorhinal cortex and hippocampus. Each regional count is scaled by dividing by the corresponding standard deviation."
τ-IHC	"Phospho-tau (immature and mature) immunohistochemistry (fresh frozen tissue; labeled by AT8): areal density." Measurements were available for each structure with corresponding expression data.	NA	"Neuronal neurofibrillary tangles are identified by molecularly specific immunohistochemistry (antibodies to abnormally phosphorylated Tau protein, AT8). Cortical density (per mm2) is determined using systematic sampling." Used measurement from midfrontal cortex, closest available to DLPFC.

Supplementary Table 2.2. Overview of cohorts used for MD-AD training. For continuous variables, we show mean and standard deviation and make pairwise comparisons using two-sided t-tests. For categorical or binary variables, show percentages and use chi-square tests to compare across cohorts. For MD-AD training phenotypes, we show the mean and standard deviation (along with ranges due to skewed distributions) of the values consistent with units as indicated in Supplementary Table 2.1 (note that units vary across studies). Because units varied across studies, for computing statistical comparisons between MD-AD training phenotypes, we consider the normalized values used in MD-AD model training (values normalized to range from 0 to 1 within each data set). "NA" indicates that the variable was not available for the dataset.

Across all variables, significant differences between groups (calculated via statistical tests described above) are indicated by the first letter of the dataset (e.g., ^{A} p <.05, ^{AA} p <.01, ^{AAA} p<.001).

Variable	**ACT**	**MSBB**	**ROSMAP**
Counts (by region)	N=337 (89 TCx, 84 HIP, 83 FWM, 81 PCx)	N=879 (244 BM10, 227 BM22, 208 BM44, 200 BM36)	N=542 (all DLPFC)
Age (capped at 90)	86.97 +- 4.08 ^{MMM}	83.30 +- 7.52 $^{AAA\ RRR}$	86.55 +- 4.61 ^{MMM}
Sex (% male)	62.31 $^{MMM\ RRR}$	36.63 ^{AAA}	37.08 ^{AAA}
Race (% white)	97.63 ^{MMM}	80.09 $^{AAA\ RRR}$	98.52 ^{MMM}
years of education	14.31 +- 3.13 ^{RRR}	NA	16.51 +- 3.49 ^{AAA}
APOE genotype (% ε2ε2/ε2ε3/ε2ε4/ε3ε3/ε3ε4/ε4ε4)	0.0/8.0/1.3/70.5/18.9/1.3 $^{MMM\ R}$	1.2/11.7/0.7/55.0/29.2/2.1 $^{AAA\ RR}$	0.9/13.1/2.2/61.1/21.8/0.9 $^{A\ MM}$
APOE ε4 carrier (% carrying at least 1 ε4 allele)	21.47 ^{MM}	32.03 $^{AA\ R}$	24.91 ^{M}
Cognition status (% dementia)	48.19 ^{MMM}	59.27 $^{AAA\ RRR}$	42.80 ^{MMM}
Cognition status including mild impairment* (% None/Mild Impairment/Dementia)	NA	29.5/11.3/59.3 ^{RRR}	31.7/25.5/42.8 ^{MMM}
Diagnosis based on NINCDS/ADRDA criteria** (% No dementia/Probable AD/Possible AD/Dementia-type unknown)	51.3/18.7/20.5/9.5 ^{RRR}	NA	57.2/35.6/5.0/2.2 ^{AAA}
AD to death time (% no dementia/<=2/2-5/> 5 years)	60.8/12.8/16.0/10.4	NA	56.1/16.1/17.2/10.7
RIN	6.35 +- 1.07 ^{MMM}	6.84 +- 1.47 ^{AAA}	NA
PMI (hours)	NA	7.10 +- 5.33	7.16 +- 4.84
MD-AD Training Phenotypes:			
CERAD Score	1.49 +- 1.08 (0.00-3.00) ^{M}	1.66 +- 1.28 (0.00-3.00) ^{A}	1.61 +- 1.16 (0.00-3.00)

Neuritic Plaques	NA	8.05 +- 8.79 (0.00-42.00)	0.73 +- 0.79 (0.00-4.96)
Aβ-IHC	0.02 +- 0.02 (0.00-0.09)	NA	4.72 +- 5.21 (0.00-26.31)
Braak Stage	3.45 +- 1.67 (0.00-6.00)	3.68 +- 1.86 (0.00-6.00)	3.37 +- 1.28 (0.00-6.00)
Tangles	NA	NA	0.55 +- 0.73 (0.00-6.17)
τ-IHC	0.02 +- 0.03 (0.00-0.11) [RRR]	NA	1.44 +- 5.68 (0.00-89.87) [AAA]

*For MSBB, cognition status was based on the clinical dementia rating (CDR). For ROSMAP, this was based on consensus between a neuropsychologist and clinician based on a battery of tests.

**For NINCDS/ADRDA criteria, Possible and Probable AD both indicates that the individual has dementia, but with differing certainty about the cause while they were alive.

Supplementary Table 2.3. Overview of MD-AD consensus nodes. Each node is annotated with -log10(p-values) for phenotypes and covariates of interest. We also show or enrichment scores for KEGG and REACTOME pathways. Enrichment scores were masked to 0 when the enrichment *p*-value were not significant (i.e., $p>0.05$ after Bonferroni correction across nodes).

The table is available for download here: https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-021-25680-7/MediaObjects/41467_2021_25680_MOESM4_ESM.xlsx

Supplementary Table 2.4. MD-AD's consensus gene scores, with high scores being most positively related to neuropathology and low scores being most negatively related to neuropathology. "all-related" is averaged over the six neuropathological phenotypes, while "abeta-related" is averaged over CERAD, PLAQUES, and ABETA_IHC, and "tau-related" is averaged over BRAAK, TANGLES, and TAU_IHC.

The table is available for download here: https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-021-25680-7/MediaObjects/41467_2021_25680_MOESM5_ESM.xlsx

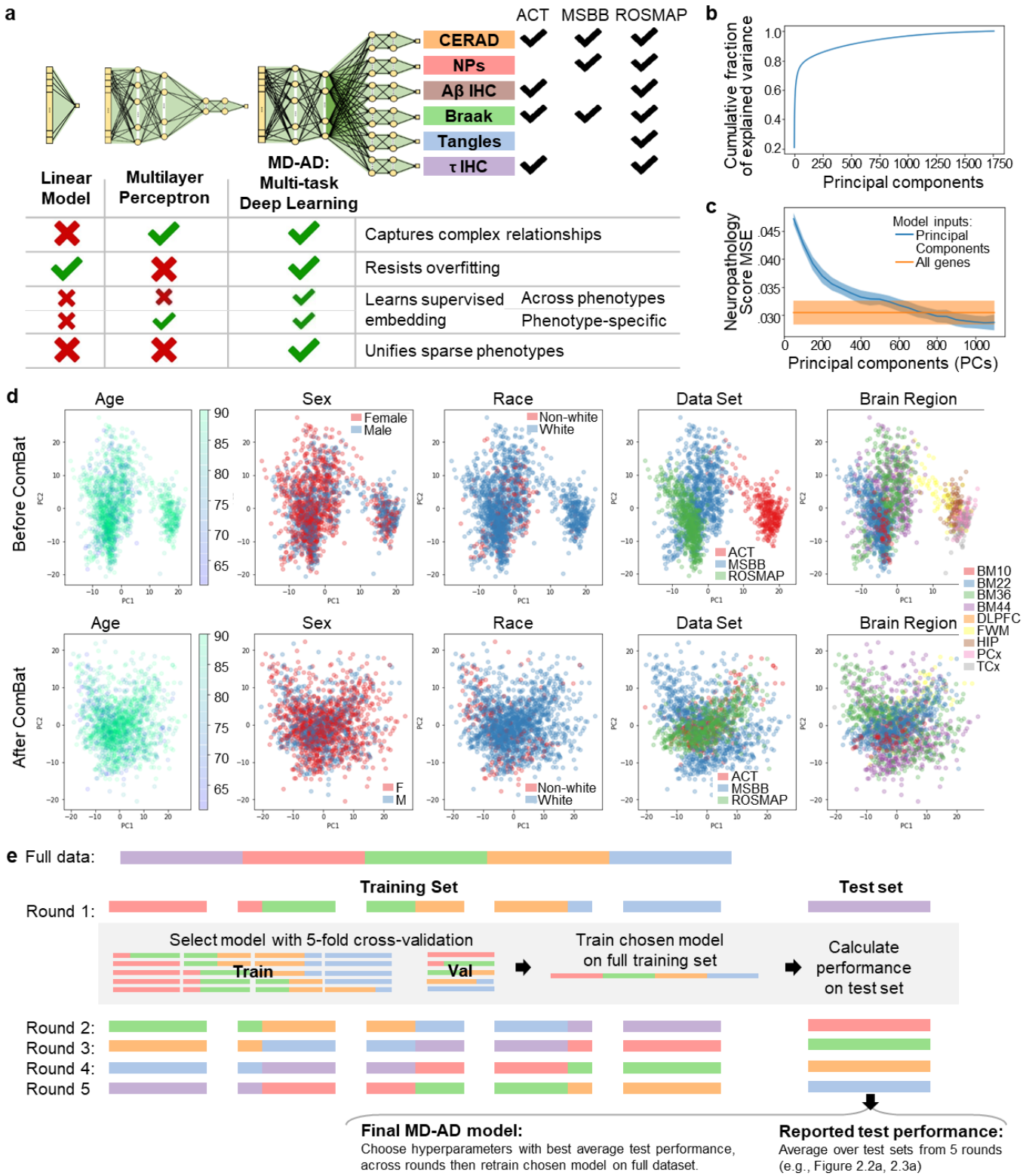
Supplementary Table 2.5. Overview of AddNeuroMed blood samples. For each cell, we show summary statistics (and any group with which this group has a statistically significant difference, as measured by independent t-tests or chi-square tests). For each group, we highlight all groups that are significantly different (*p<.05, **p<.01, ***p<.001).

Blood1 (GSE63060)	**CTL (N=104)**	**MCI (N=80)**	**AD (N=145)**
age	72.38 +- 6.31 (MCI* AD***)	74.45 +- 5.96 (CTL*)	75.40 +- 6.56 (CTL***)
gender (% female/male)	59.6/40.4	48.8/51.2 (AD**)	68.3/31.7 (MCI**)
ethnicity (% Western European / Other Caucasian / Other or Unknown)	95.2/4.8/0.0 (MCI*** AD*)	73.8/22.5/3.8 (CTL***)	85.5/13.1/1.4 (CTL*)
Blood2 (GSE63061)	**CTL (N=134)**	**MCI (N=109)**	**AD (N=139)**
age	75.29 +- 6.00 (MCI*** AD***)	78.16 +- 7.33 (CTL***)	77.89 +- 6.64 (CTL***)
gender (% female/male)	60.4/39.6	59.6/40.4	61.2/38.8
ethnicity (% Western European / Other Caucasian / Other or Unknown)	91.0/6.7/2.2 (MCI*)	78.9/12.8/8.3 (CTL* AD**)	87.8/11.5/0.7 (MCI**)

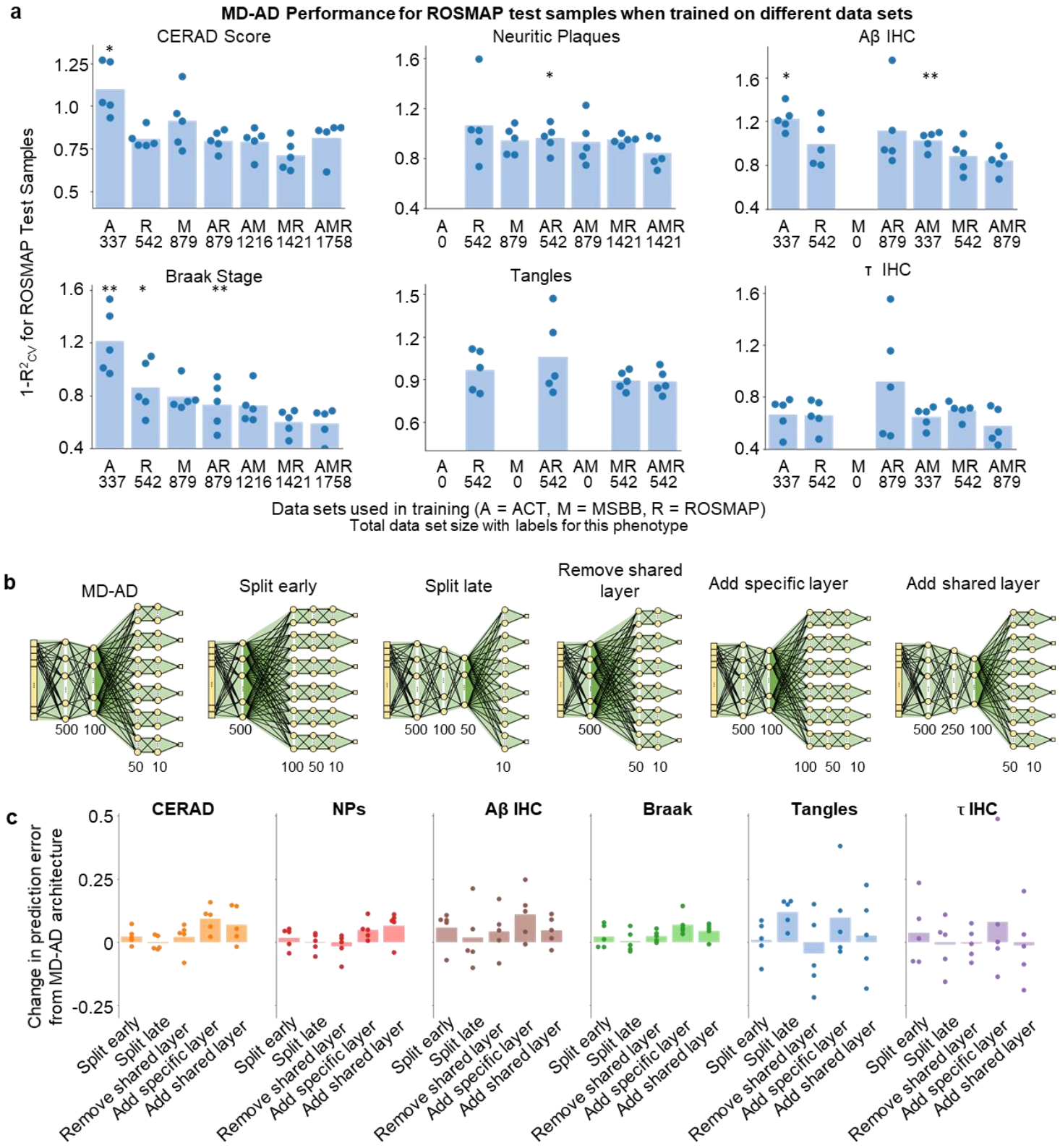
Supplementary Table 2.6. Final selected models for MD-AD and baseline methods, after aggregating performance across all 5 test folds. All networks were trained for 200 epochs with batch sizes of 20, ReLU activations, and dropout units with 0.1 dropout rate, using an Adam optimizer.

(View online materials to see an additional table showing the best hyperparameter setting for each phenotype within each model class, separately computed for each training/test split: https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-021-25680-7/MediaObjects/41467_2021_25680_MOESM7_ESM.xlsx)

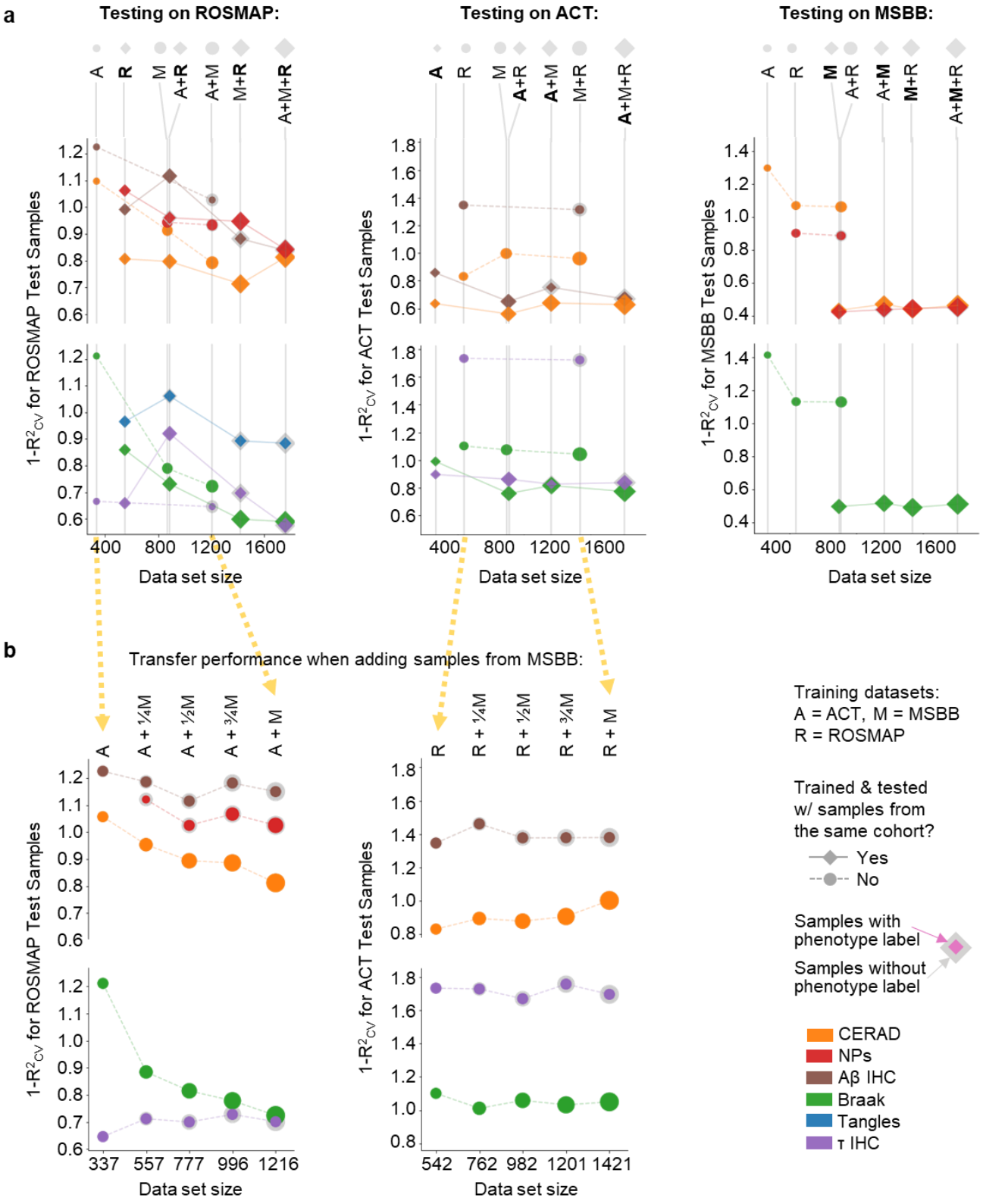
Final selected hyperparameters for each method			
model	**kernel regularization**	**learning rate**	**gradient clipping norm**
MD-AD	0.001	0.001	0.1
MLP_baselines for ABETA_IHC	0.00001	0.001	0.01
MLP_baselines for TAU_IHC	0.00001	0.001	0.01
MLP_baselines for PLAQUES	0.00001	0.001	0.1
MLP_baselines for TANGLES	0.00001	0.001	0.01
MLP_baselines for BRAAK	0.00001	0.001	0.1
MLP_baselines for CERAD	0.00001	0.001	0.01
Linear_baselines for ABETA_IHC	0.00001	0.001	0.01
Linear_baselines for TAU_IHC	0.001	0.001	0.01
Linear_baselines for PLAQUES	0.00001	0.001	0.01
Linear_baselines for TANGLES	0.001	0.0001	0.01
Linear_baselines for BRAAK	0.001	0.001	0.01
Linear_baselines for CERAD	0.00001	0.001	0.1



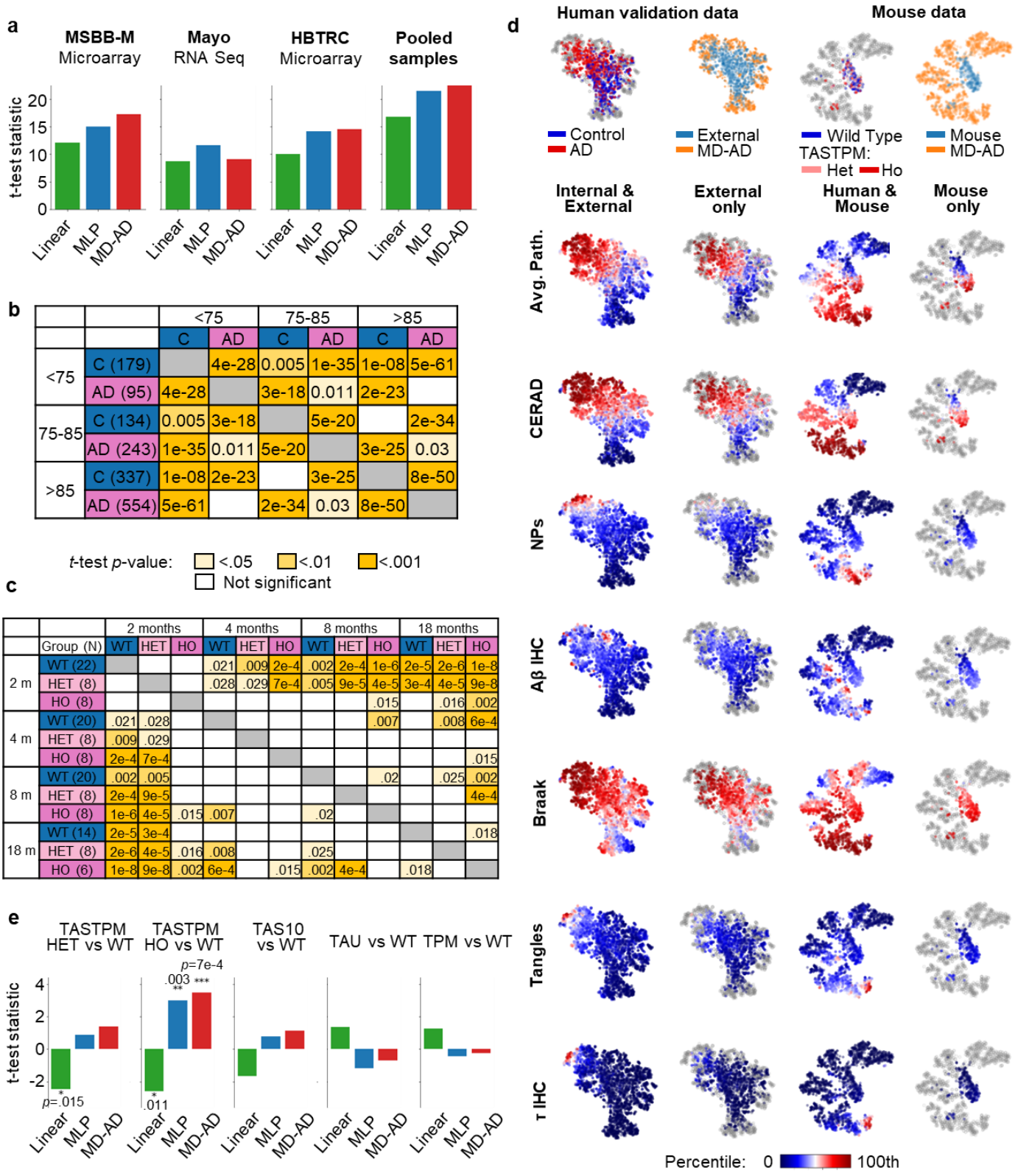
Supplementary Figure 2.1. (a) Overview of MD-AD and its advantages over traditional approaches. **(b)** Cumulative variance explained for principal components (computed from the full dataset after all pre-processing). **(c)** Average test MSE for predicting average neuropathology score from linear model trained on PC-transformed inputs vs all genes with standard error bands (n=5 test runs). **(d)** First two principal components of gene expression data before and after ComBat batch effect correction. Brain regions shown: Brodmann areas 10, 22, 36, 44 (BM10, BM22, BM36, BM44), dorsolateral prefrontal cortex (DLPFC), hippocampus (HIP), frontal white matter (FWM), parietal cortex (PCx), temporal cortex (TCx). **(e)** Overview of our cross-validation (CV) and testing scheme. We generate five separate training and test splits, and then in each round, we perform cross-validation to choose hyperparameters, retrain the model on the full training set, and then report test performance.



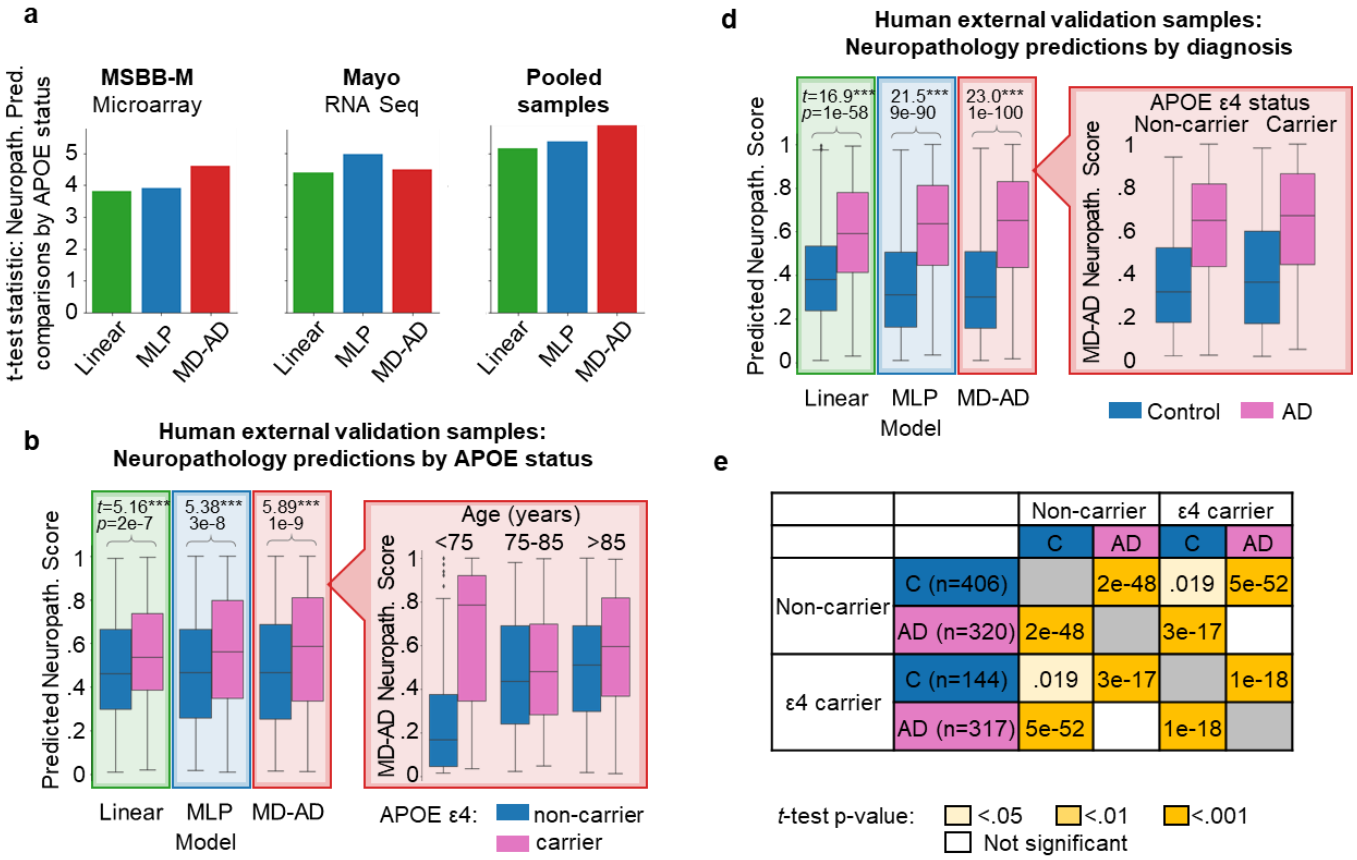
Supplementary Figure 2.2. (a) We evaluate test set performance for ROSMAP using the same training and test splits, but training restricted to different subsets of available data sets, averaged over test folds (n=5 test runs per data set). **(b)** We experimented with several architectures for the MD-AD model. They are depicted with associated dense layer sizes. **(c)** We plot the difference in test set prediction error between five alternative architectures evaluated and the final selected MD-AD architecture, averaged over five test folds. Positive values indicate higher error relative to the MD-AD architecture (n=5 test runs per architecture).



Supplementary Figure 2.3. **(a)** Test prediction performance using 5-fold cross-validation when training on different subsets of available datasets. We display performance for each dataset's test samples separately. Circle markers show performance when transferring a trained model to a new dataset; diamond markers show changes in performance when augmenting the training set with samples from other datasets. **(b)** Same analysis as part (**a**), but highlighting how transfer performance changes when adding additional MSBB samples during training.



Supplementary Figure 2.4. **(a)** *t*-test statistics measuring differences between each model's predicted neuropathology scores for AD-diagnosed vs. control individuals. **(b)** Significance of 2-sided *t*-tests measuring between-group differences as shown in boxplots in Figure 2.2c. **(c)** Significance of 2-sided *t*-tests measuring between-group differences as shown in boxplots in Figure 2.2d. **(d)** t-SNE plots of embedded samples for external and mouse data sets. **(e)** *t*-test statistics measuring differences between each model's predicted neuropathology scores for AD model strains of mice vs. wild type (WT) mice.



e

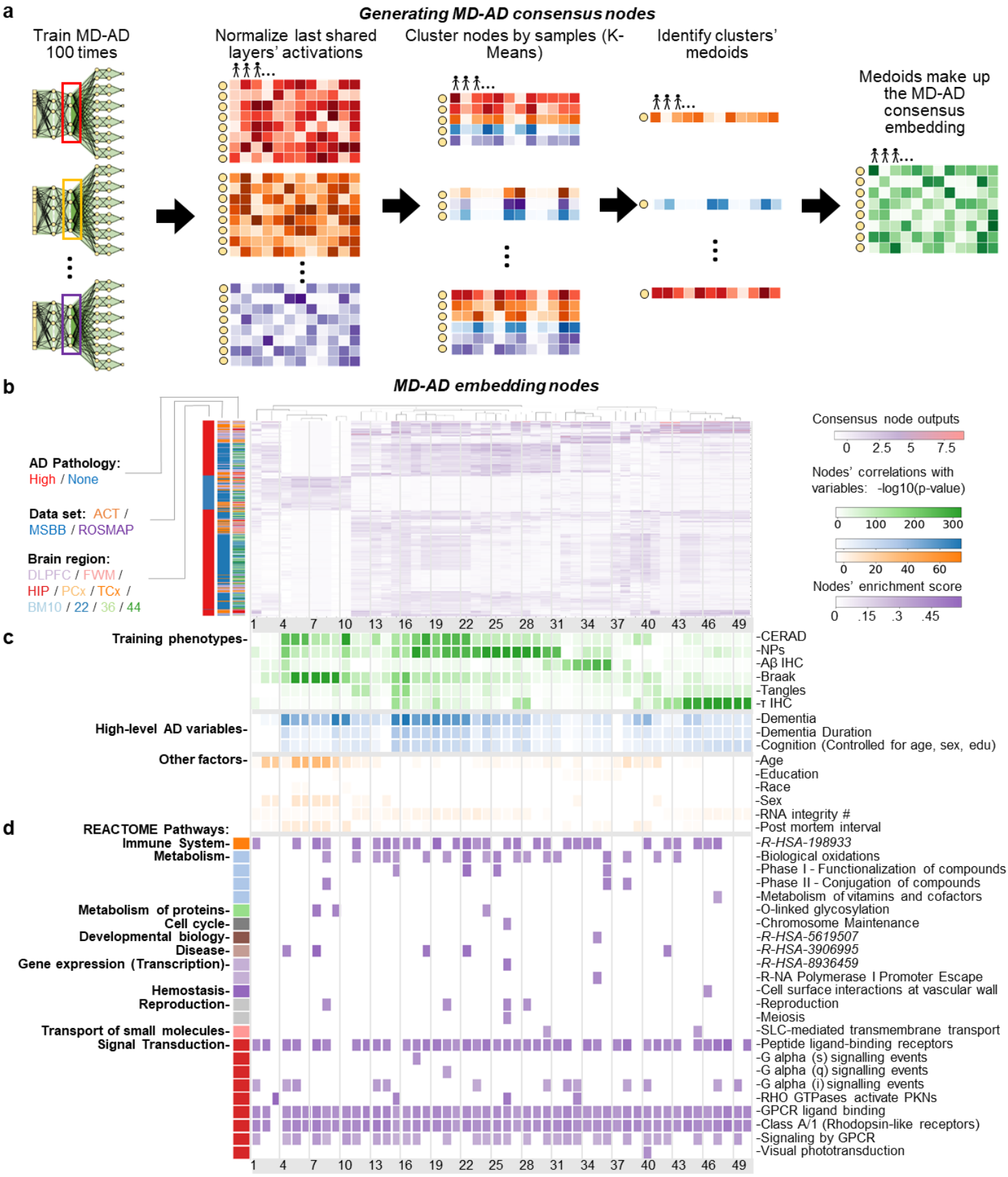
		Non-carrier		ε4 carrier	
		C	AD	C	AD
Non-carrier	C (n=406)		2e-48	.019	5e-52
	AD (n=320)	2e-48		3e-17	
ε4 carrier	C (n=144)	.019	3e-17		1e-18
	AD (n=317)	5e-52		1e-18	

t-test p-value: <.05 <.01 <.001 Not significant

c

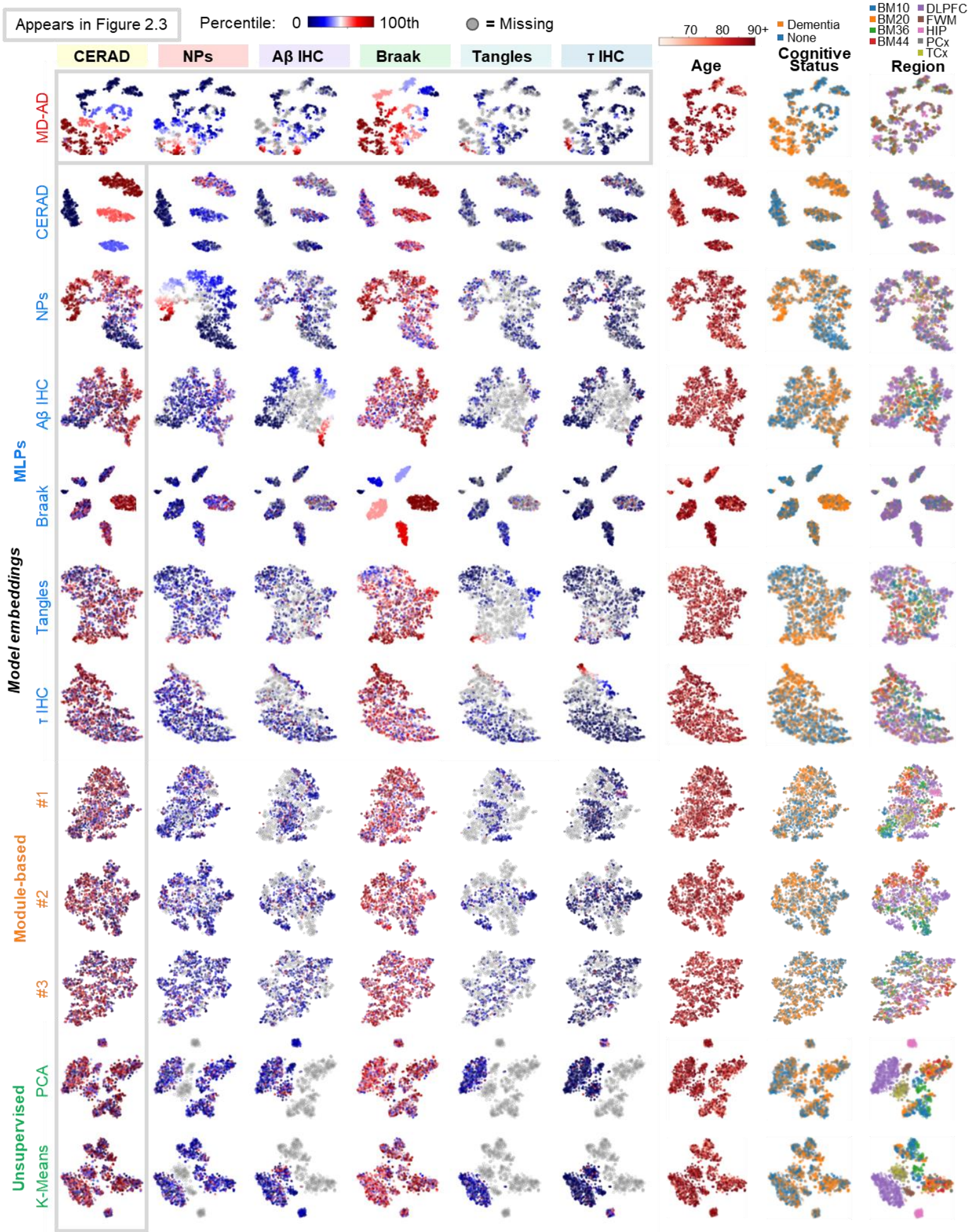
		<75		75-85		>85	
		NC	ε4	NC	ε4	NC	ε4
<75	Non-carrier (n=71)		1E-12	5E-7	3E-9	1E-13	8E-18
	ε4 carrier (n=72)	1E-12		2E-5	4E-4	7E-6	
75-85	Non-carrier (n=125)	5E-7	2E-5				3E-5
	ε4 carrier (n=143)	3E-9	4E-4				.002
>85	Non-carrier (n=530)	1E-13	7E-6				2E-5
	ε4 carrier (n=246)	8E-18		3E-5	.002	2E-5	

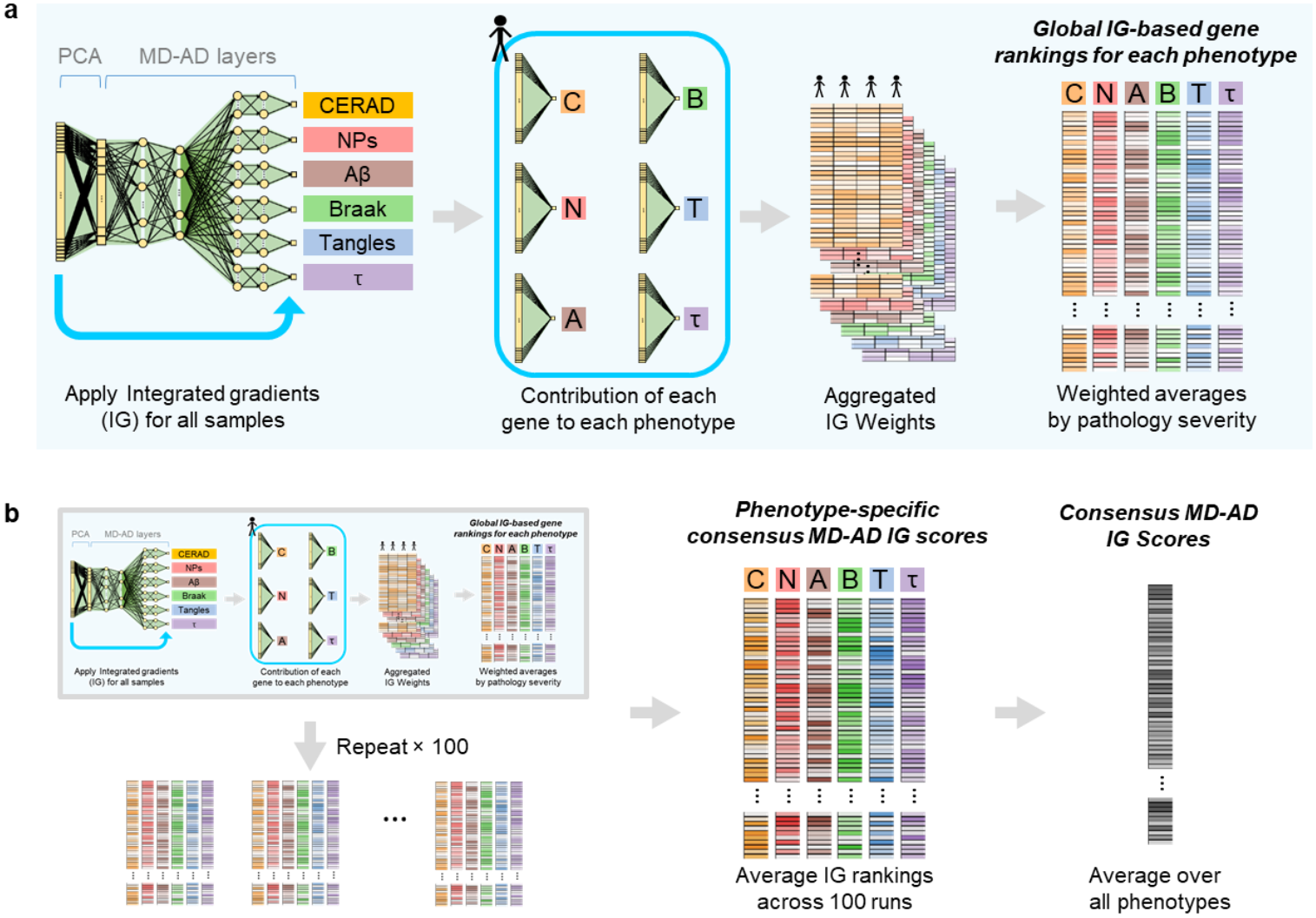
Supplementary Figure 2.5. External validation results considering APOE status. **(a-c)** show how neuropathology predictions compare for carriers vs. non-carriers of the APOE ε4 allele. **(a)** t-test statistics measuring differences between each model's predicted neuropathology scores for carriers vs. non-carriers of APOE ε4. **(b)** For samples from external validation data sets, we obtain neuropathology scores for each sample from each model. *Left*: Box plots displaying the distribution of predicted neuropathology scores from each method for APOE ε4 carriers vs. non-carriers. *T*-tests highlight between-group differences for each method (two-sided *t*-test, ***p<.001; see sample sizes in part c.) *Right*: Box plots displaying the distribution of MD-AD's predicted neuropathology scores split by age group and APOE status. All box plots in this figure indicate median (center line), upper and lower quartiles (box limits), 1.5x interquartile range from quartiles (whiskers), and outliers (points). **(c)** sample sizes and *p*-values from *t*-tests comparing pairs of groups shown in part b.). **(d)** *Left*: Predicted neuropathology scores for each method split by diagnosis (replicated from Figure 2.2c; see sample sizes in part e.). *Right*: Box plots displaying the distribution of MD-AD's predicted neuropathology scores split by both diagnosis and APOE status (same box plot elements as described in part b). **(e)** sample sizes and *p*-values from two-sided t-tests comparing pairs of groups shown in part d.


a
Generating MD-AD consensus nodes
Train MD-AD 100 times
Normalize last shared layers' activations
Cluster nodes by samples (K-Means)
Identify clusters' medoids
Medoids make up the MD-AD consensus embedding
b
MD-AD embedding nodes
AD Pathology: High / None
Data set: ACT / MSBB / ROSMAP
Brain region: DLPFC / FWM / HIP / PCx / TCx / BM10 / 22 / 36 / 44
Consensus node outputs
0 2.5 5 7.5
Nodes' correlations with variables: -log10(p-value)
0 100 200 300
0 20 40 60
Nodes' enrichment score
0 .15 .3 .45
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49
c
Training phenotypes-
-CERAD
-NPs
-Aβ IHC
-Braak
-Tangles
-τ IHC
High-level AD variables-
-Dementia
-Dementia Duration
-Cognition (Controlled for age, sex, edu)
Other factors-
-Age
-Education
-Race
-Sex
-RNA integrity #
-Post mortem interval
d
REACTOME Pathways:
Immune System-
-R-HSA-198933
Metabolism-
-Biological oxidations
-Phase I - Functionalization of compounds
-Phase II - Conjugation of compounds
-Metabolism of vitamins and cofactors
Metabolism of proteins-
-O-linked glycosylation
Cell cycle-
-Chromosome Maintenance
Developmental biology-
-R-HSA-5619507
Disease-
-R-HSA-3906995
Gene expression (Transcription)-
-R-HSA-8936459
-R-NA Polymerase I Promoter Escape
Hemostasis-
-Cell surface interactions at vascular wall
Reproduction-
-Reproduction
-Meiosis
Transport of small molecules-
-SLC-mediated transmembrane transport
Signal Transduction-
-Peptide ligand-binding receptors
-G alpha (s) signalling events
-G alpha (q) signalling events
-G alpha (i) signalling events
-RHO GTPases activate PKNs
-GPCR ligand binding
-Class A/1 (Rhodopsin-like receptors)
-Signaling by GPCR
-Visual phototransduction
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49

Previous page: **Supplementary Figure 2.6.** Generating and annotating MD-AD "consensus" nodes. **(a)** Illustration of how we generate MD-AD consensus nodes. **(b)** Bi-clustered last shared layer consensus node embeddings. **(c)** Correlations between consensus nodes and phenotypes of interest. Cells show the –log10(*p*-value) of the correlation after FDR correction across nodes. **(d)** Nodes' GSEA enrichment score (ES) for REACTOME pathways: For each node, we obtain integrated gradients weights from each gene. We display only cells with |ES|>.2 and $p<.05$ after FDR correction (across nodes). REACTOME pathways with long names are indicated by their REACTOME stable IDs.

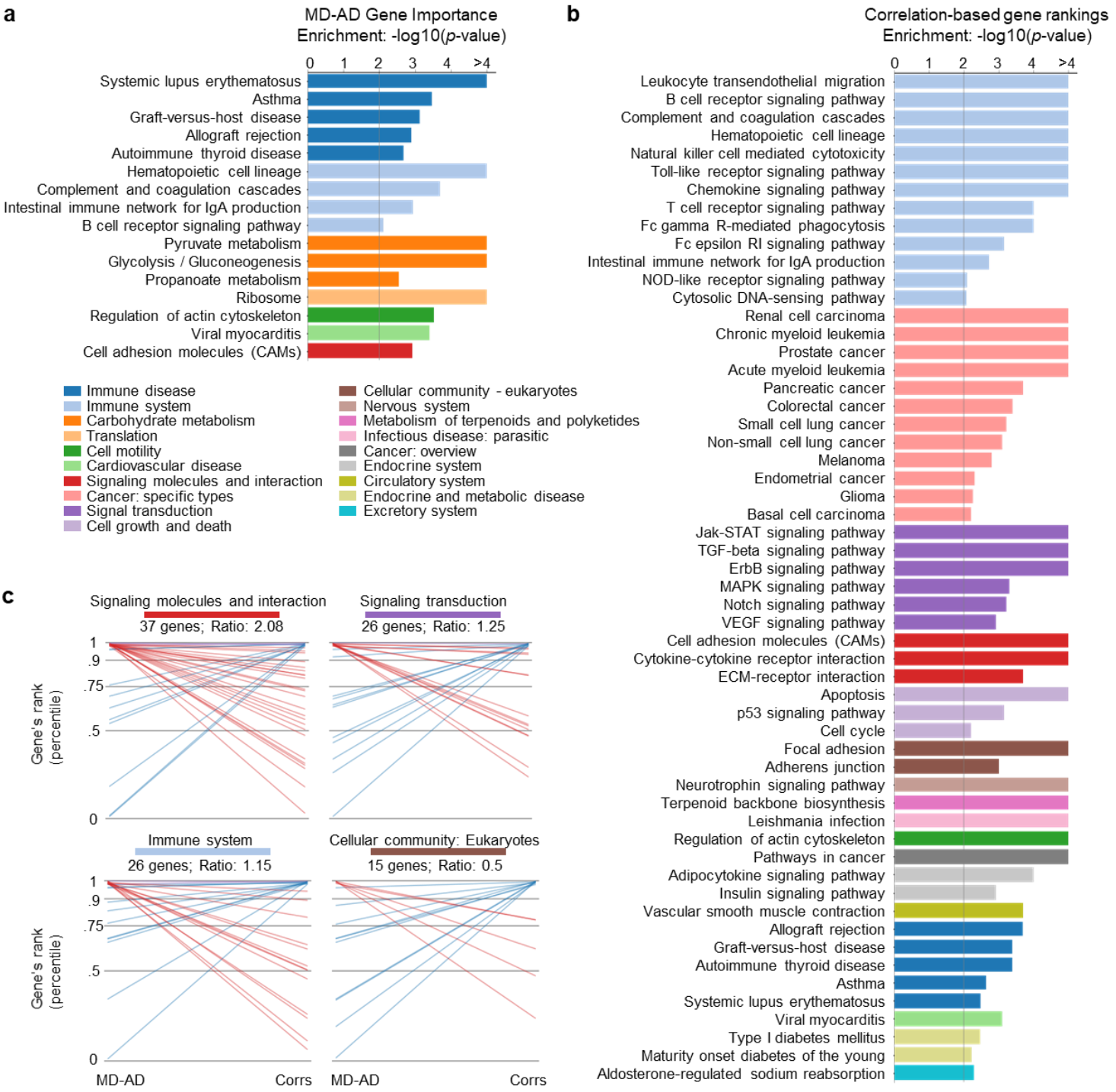
Next page: **Supplementary Figure 2.7.** *t*-SNE representation of all methods' embeddings colored by pathology values.


Appears in Figure 2.3
Percentile: 0 100th
= Missing
70 80 90+
Age
Dementia
None
Cognitive Status
BM10 DLPFC
BM20 FWM
BM36 HIP
BM44 PCx
TCx
Region
CERAD
NPs
Aβ IHC
Braak
Tangles
τ IHC
MD-AD
CERAD
NPs
Aβ IHC
Braak
Tangles
τ IHC
MLPs
Model embeddings
#1
#2
#3
Module-based
PCA
K-Means
Unsupervised

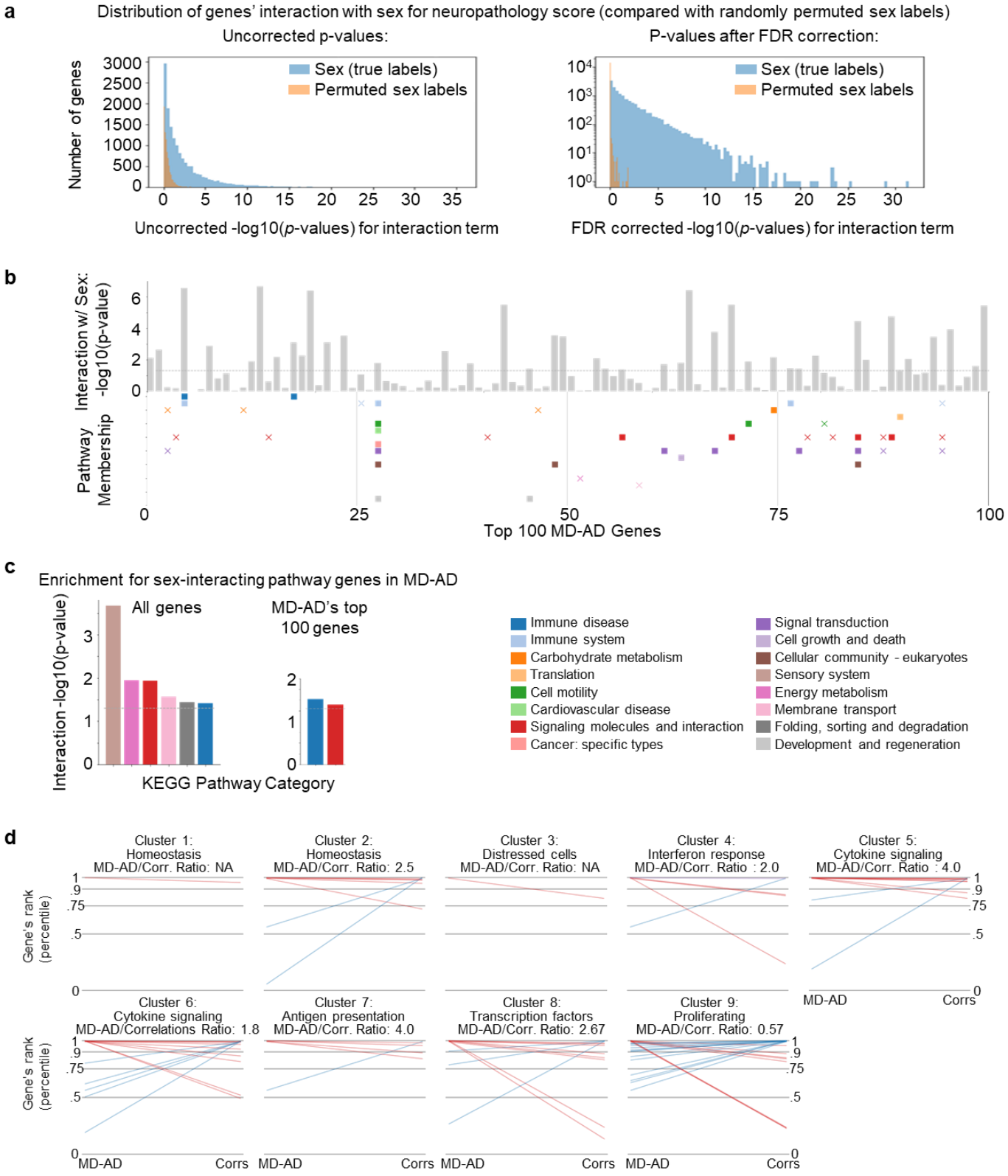


Supplementary Figure 2.8. Illustration of how MD-AD "consensus" gene scores are generated. **(a)** For a single training run, we aggregate IG scores across samples to obtain a ranking over genes for each phenotype. **(b)** We aggregate IG gene scores across 100 re-trainings of MD-AD. **(c)** Consistency of top genes from aggregating IG gene scores across multiple training runs.

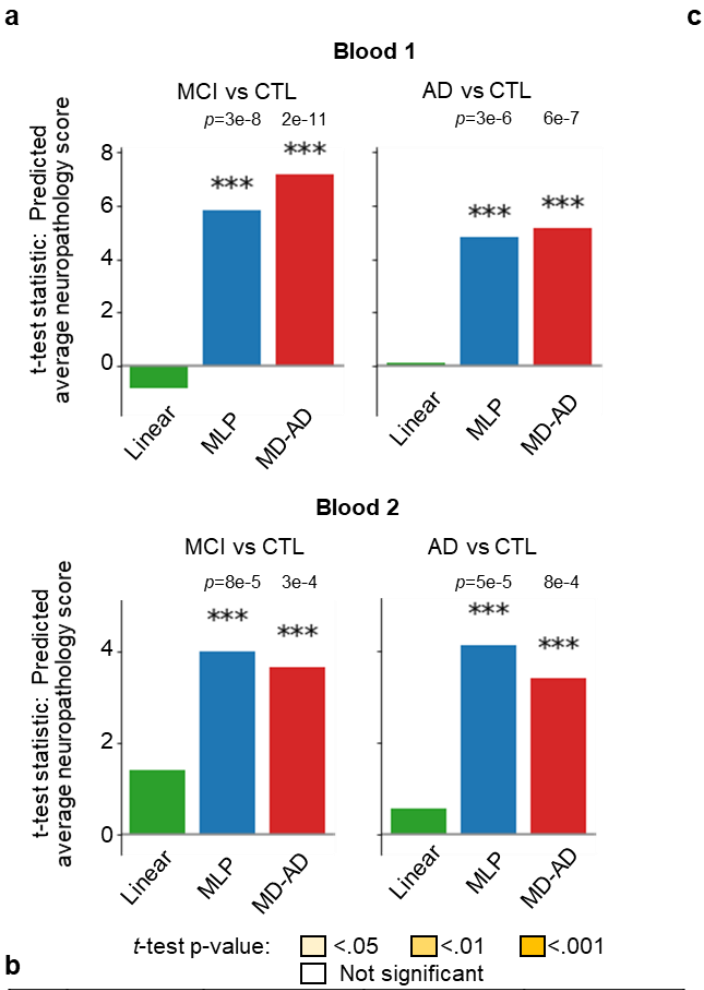
Supplementary Figure 2.9. Top genes and REACTOME pathway enrichment for correlation-based ranking: (a) Top 50 genes ranked by correlations between expression and pathology. **(b)** GSEA enrichment *p*-values for pathways enriched in correlation-based ranking.



Supplementary Figure 2.10. (a) GSEA enrichment *p*-values for KEGG pathways enriched in MD-AD gene ranking, **(b)** GSEA enrichment *p*-values for pathways enriched in correlation-based ranking, **(c)** Comparison of top genes for MD-AD vs correlations. For MD-AD and correlation-based rankings, we identify all genes in the top 2% of the ranking, and then check their membership in KEGG categories.

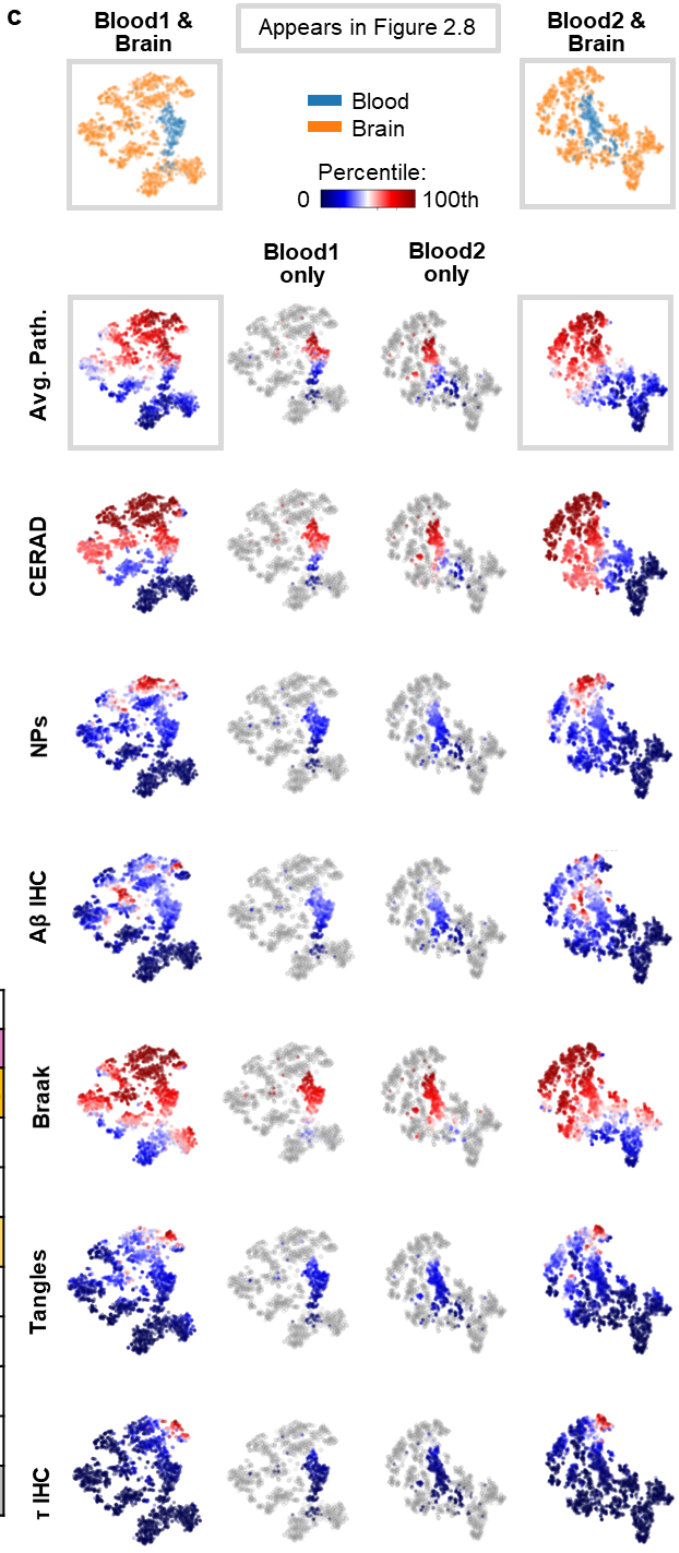


Supplementary Figure 2.11. Additional details for sex interaction results. **(a)** Distribution of genes' interactions with sex for MD-AD scores. For each gene, we compute the –log10(p-value) for the interaction term between the gene's expression and sex. For comparison, we show the distribution from an experiment with all sex labels shuffled across samples. **(b)** Replicated results from Figure 2.5a with KEGG categories. **(c)** Replicated results from Figure 2.5b with KEGG categories. **(d)** Comparison of top 1% ranked genes from MD-AD vs a correlation-based approach for microglial cluster members.

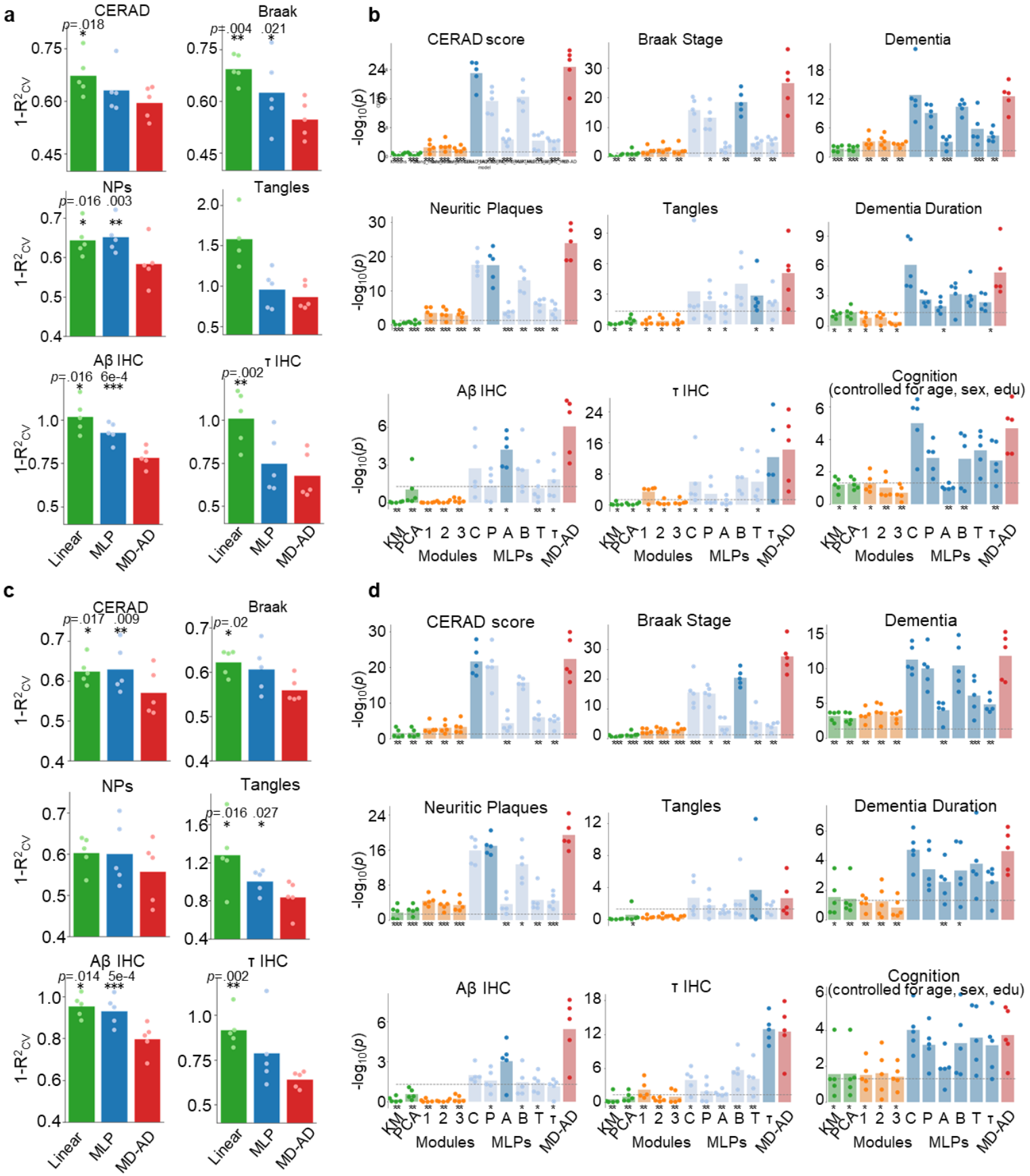


b

		<74			[74-80)			80+		
	Group (N)	CTL	MCI	AD	CTL	MCI	AD	CTL	MCI	AD
<74	CTL (n=113)		1E-8	5E-7		8E-8	7E-5		3E-5	1E-5
	MCI (n=64)	1E-8			7E-6		.042	.003		
	AD (n=89)	5E-7			3E-4			.029		
[74-80)	CTL (n=82)		7E-6	3E-4		3E-5	.01		.003	.004
	MCI (n=63)	8E-8			3E-5			.007		
	AD (n=90)	7E-5	.042		.01					
80+	CTL (n=43)		.003	.029		.007				
	MCI (n=62)	3E-5			.003					
	AD (n=105)	1E-5			.004					



Supplementary Figure 2.12. (a) *T*-test statistics for comparisons among MD-AD predicted neuropathology and cognitive states, separately for blood datasets (see sample sizes in part b). **(b)** Pair-wise *t*-test *p*-values from Figure 2.7b. **(c)** Embeddings from MD-AD's last shared layer for brain and blood data. Each plot is colored by dataset or predicted pathology severity for various phenotypes. Pathology severity plots are produced with and without MD-AD training (brain) samples for clarity.


a
CERAD
p=.018
Braak
p=.004 .021
NPs
p=.016 .003
Tangles
Aβ IHC
p=.016 6e-4
τ IHC
p=.002
1-R2CV
Linear
MLP
MD-AD
b
CERAD score
Braak Stage
Dementia
Neuritic Plaques
Tangles
Dementia Duration
Aβ IHC
τ IHC
Cognition
(controlled for age, sex, edu)
-log10(p)
KM PCA 1 2 3 C P A B T τ MD-AD
Modules
MLPs
c
CERAD
p=.017 .009
Braak
p=.02
NPs
Tangles
p=.016 .027
Aβ IHC
p=.014 5e-4
τ IHC
p=.002
d
CERAD score
Braak Stage
Dementia
Neuritic Plaques
Tangles
Dementia Duration
Aβ IHC
τ IHC
Cognition
(controlled for age, sex, edu)

Supplementary Figure 2.13. Cross-validation performance metrics for experiments with alternative methods (all bars indicate an average over five test runs, and each run's performance is overlaid as a dot). For each subplot, we highlight differences in performance between alternative methods and MD-AD via 2-sided paired *t*-tests using n=5 test runs (*p*-values: *<.05, **<.01, ***<.001). **(a)** Figure 2.2a replicated with results from training and evaluation using GE normalized to account for post-mortem interval (PMI) and RNA integrity number (RIN). **(b)** Figure 2.3a replicated with results from training and evaluation using GE normalized to account for PMI and RIN. Each bar represents the –log10(p-value) after FDR correction over nodes for the correlation between the best node and phenotype listed, averaged over 5 test run. **(c)** Figure 2.2a replicated with new cross-validation and test splits generated by pseudorandomly assigning samples by individual (rather than completely randomly). **(d)** Figure 2.3a replicated with the new cross-validation and test splits.

Chapter 3 Supplementary Materials

SUPPLEMENTARY METHODS

Pathway graph construction

In the PAC method, we create a pathway network in which each node represents a pathway, and each edge represents how closely two pathways are related in terms of their shared genes. In order to represent relatedness of pathways, we use Fisher's exact test (FET), which is a commonly used statistical test for measuring gene set enrichment and relatedness between pathways[1]. To that end, we first generated a reference set of 17,640 genes (G) which were present in at least one pathway in any of the four pathway databases. For each pathway i we have P_i, the set of all genes in the pathway, and P_i^c, the set of all genes not in the pathway (i.e., in the reference set G but not P_i). For a pair of pathways Pi and P_j, we therefore use Fisher's exact test based on the contingency table of these two pathway gene sets, as shown in Supplementary Figure 3.1a. Finally, we use the –log10(p-values) from each pairwise Fishers exact test (after Bonferroni correction across all pairs of pathways) as edges in the graph (note that edge weights are set to 0 for pairs of pathways that were not significantly overlapping at the p<.01 level). As described by Rivals et al.[126], several formulations of enrichment tests all rely on an underlying hypergeometric null distribution, and all of these may equally be called Fisher's exact tests or hypergeometric tests. For our purposes, we used the python scikit-learn implementation of FET. In particular, we use two-sided tests to be consistent with the fact that they are most commonly appropriate in practice for evaluating the enrichment or depletion of pathways in differential gene expression analyses (e.g., our example using breast cancer data).

Preprocessing of pathways' curated hierarchies

While KEGG assigns each of its pathways to a high-level category that we use directly, REACTOME, GO Biological Processes (BP) and GO Molecular Function (MF) provide a tree hierarchy for their pathways that we trace to define the associated highest-level categories (see Figure 3.2 for the number of pathways and categories from each database). The processing steps below yield a set of "curated" category labels for pathways in each of the four databases, listed in Supplementary Table 3.1, which we use to evaluate our community detection methods:

KEGG: KEGG consists of 186 pathways, which are divided across 35 higher-level categories provided by the resource, and thus needed no further pre-processing

REACTOME: REACTOME consists of 1,499 pathways which are related to each other as a hierarchy of pathways ranging from highly specific to 25 general pathways (as shown on their interactive webtool: https://reactome.org/PathwayBrowser/). For each pathway, we trace the hierarchy tree and assign the pathway's category as the highest-level pathway to which we can trace the original pathway. A small number of pathways have multiple parents, and in these cases, we count all possible paths to higher-level categories, and select the category label for which there are the most paths.

Gene Ontology (GO): GO consists of three subcategories, which we treat as separate sources: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) containing 7,350, 1,001, and 1,645

gene sets, respectively. For this analysis, we exclude the CC gene sets because unlike the other databases and GO categories which relate more closely to biological processes, GO CC relates more closely to cellular structures. The GO resource contains detailed mappings (of several types) among gene sets which form a hierarchy, and only considered the 'is_a' relations between gene sets which were the most frequent relation type and we discarded the 'obsolete' GO terms that are not connected to any other terms. For BP and MF gene sets separately, we traced the entire hierarchy of gene sets mapping each gene set to its direct parents. Since GO terms often have multiple parents, we traversed the hierarchy multiple times for all possible traversals from a leaf node to the high-level node and recorded the most frequent curated category for each level of the tree for each gene set. This pipeline returned us a multiple layer of hierarchy and to define a universal set of curated labels, we selected the highest possible level in these hierarchies with a reasonable coverage, obtaining 64 and 69 high level categories for GO BP and MF, respectively.

Community detection: Stability and Robustness

The Louvain algorithm employs a greedy approach, and outcomes are dependent on an initial ordering of nodes. Thus, we repeatedly ran the algorithm with different initializations to explore the stability of the resulting communities in the full graph. Using a resolution of 0.4 as described above, we find slight variations in learned communities, however find overwhelmingly similar results, as shown in Supplementary Figure 3.5. When running the algorithm 100 times, we found that the median pair-wise NMI of learned communities between runs was 0.84, and all pairs had NMIs above 0.77.

We further examined alternative approaches for computing edges in our community graph, including using the Jaccard similarity coefficient, and overlap coefficient. These approaches tended to produce similar or slightly worse downstream community detection results compared with FET-based edges; however, the Louvain algorithm continued to be the best community detection approach even with alternative edge construction methods (Supplementary Figure 3.3).

Although the resolution of 0.4 was selected because it tended to yield the highest agreement with curated categories (Supplementary Figure 3.3), and produced a relatively easy-to-interpret 35 communities on the full pathway set, we also explored alternative resolutions for our final community detection analysis and found that the communities tended to agree with each other (although lower resolution-based communities tended to be subsets of larger communities learned in for higher resolutions; Supplementary Figure 3.6).

Identifying significantly overrepresented genes in communities

For each community, we identified genes which we believe are disproportionately represented across member communities. For each gene-community pair, we use one-way chi square tests to calculate whether the number of pathways in the community that contain the gene appears at a different rate than if the gene were randomly distributed across all pathways. We then consider genes to be significantly overrepresented genes to be those with a positive chi-square statistic and $p<.01$ after Bonferroni correction over communities and genes. Users may query genes to see whether they are significantly overrepresented in any communities on the *gene-level* view of our webpage at https://nicasia.github.io/PAC, and all results are available in Supplementary Table 3.3.

Data processing supplementary info for breast cancer data

We used data provided by the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database, which we downloaded from the cBioPortal. In total, the dataset consists of gene expression profiles for primary tumors from 2,509 patients. Gene expression levels are reported for 24,368 genes, and phenotypic or treatment labels were also provided with each sample. To demonstrate the use of

our tool with minimal overhead, we used the gene expression data directly as it was provided cBioPortal with no further pre-processing.

For our analyses, we restrict our analysis to 2,469 sampled profiles for which estrogen receptor status was reported, of which 74% of samples were positive for estrogen receptors. For each of 24,368 genes (labeled by HUGO gene symbols) measured, we compare expression levels for the 1,825 ER+ and 655 ER- samples using two-sided independent *t*-tests. After Bonferroni correction accounting for testing all 24,368 genes, we found that 8,984 genes were significantly differently expressed between groups. Of these genes, we use the 243 (top 1%) of genes with the lowest significant p-values (Supplementary 3.8) for our enrichment analyses.

SUPPLEMENTARY TABLES

All supplementary tables are too large to display here; they are available online at: https://academic.oup.com/nargab/article/4/2/lqac044/6617323#supplementary-data

Supplementary Table 3.1. List of all pathways included in our analysis and their community assignment from PAC.

Supplementary Table 3.2. Top 10 automatically generated descriptions for each pathway community.

Supplementary Table 3.3. Full list of all genes with significant overrepresentation in any community.

a

		Pathway *j* members		
		Yes	No	Total:
Pathway *i* members	Yes	$\lvert P_i \cap P_j \rvert$	$\lvert P_i \cap P_j^C \rvert$	$\lvert P_i \rvert$
	No	$\lvert P_j \cap P_i^C \rvert$	$\lvert G - (P_i \vee P_j) \rvert$	$\lvert P_i^C \rvert = \lvert G \rvert - \lvert P_i \rvert$
	Total:	$\lvert P_j \rvert$	$\lvert P_j^C \rvert = \lvert G \rvert - \lvert P_j \rvert$	$\lvert G \rvert$

b

		Differentially expressed genes		
		Yes	No	Total:
Pathway *i* members	Yes	$\lvert P_i \cap DE \rvert$	$\lvert P_i \cap DE^C \rvert$	$\lvert P_i \rvert$
	No	$\lvert DE \cap P_i^C \rvert$	$\lvert G - (P_i \vee DE) \rvert$	$\lvert P_i^C \rvert = \lvert G \rvert - \lvert P_i \rvert$
	Total:	$\lvert DE \rvert$	$DE^C = \lvert G \rvert - \lvert DE \rvert$	$\lvert G \rvert$

Supplementary Figure 3.1. Contingency tables used for Fisher's exact tests. (a) Contingency tables used to calculate pairwise pathway similarity among all pathways. Each pathway P_i is a set of genes, and we consider P_i^C to be the set difference between G (our reference gene set) and P_i. (b) Contingency tables used to query new gene sets (such as differentially expressed genes in our Breast Cancer example).

Supplementary Figure 3.2. While our main analyses use normalized mutual information (NMI) to compare methods, other metrics are available for evaluating clusters. We show two alternative evaluation metrics: adjusted Rand index (ARI) and variation of information (VI).

a

NMI results using Fisher's Exact Test Edges

Methods	Kegg	Reactome	GO BP	GO MF
Agglomerative	0.65	0.1866	0.1923	0.2396
Spectral	0.6330	0.3265	0.2984	0.3239
CNM	0.6099	0.2496	0.2865	0.3661
Louvain	**0.6911**	**0.4484**	**0.3890**	**0.5106**

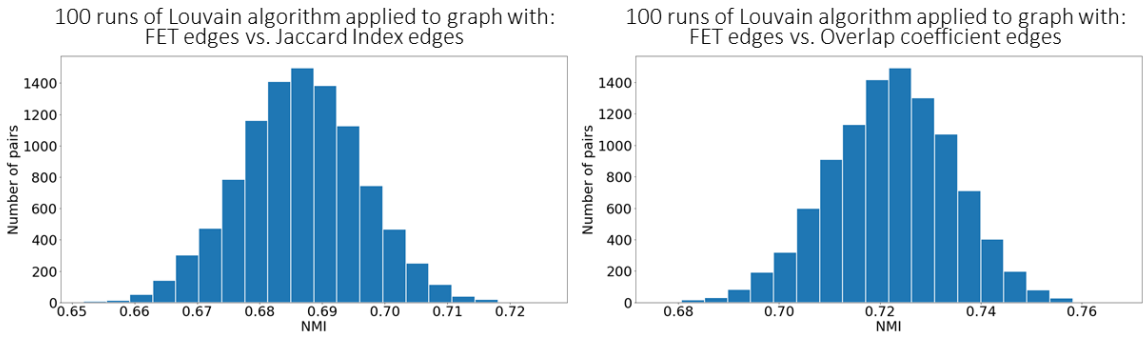
NMI results using Jaccard Index Edges

Methods	Kegg	Reactome	GO BP	GO MF
Agglomerative	0.6544	0.1875	0.1924	0.2398
Spectral	0.6334	0.2464	0.2985	0.3344
CNM	0.4008	0.0668	0.1249	0.1220
Louvain	**0.6841**	**0.3495**	**0.3714**	**0.4398**

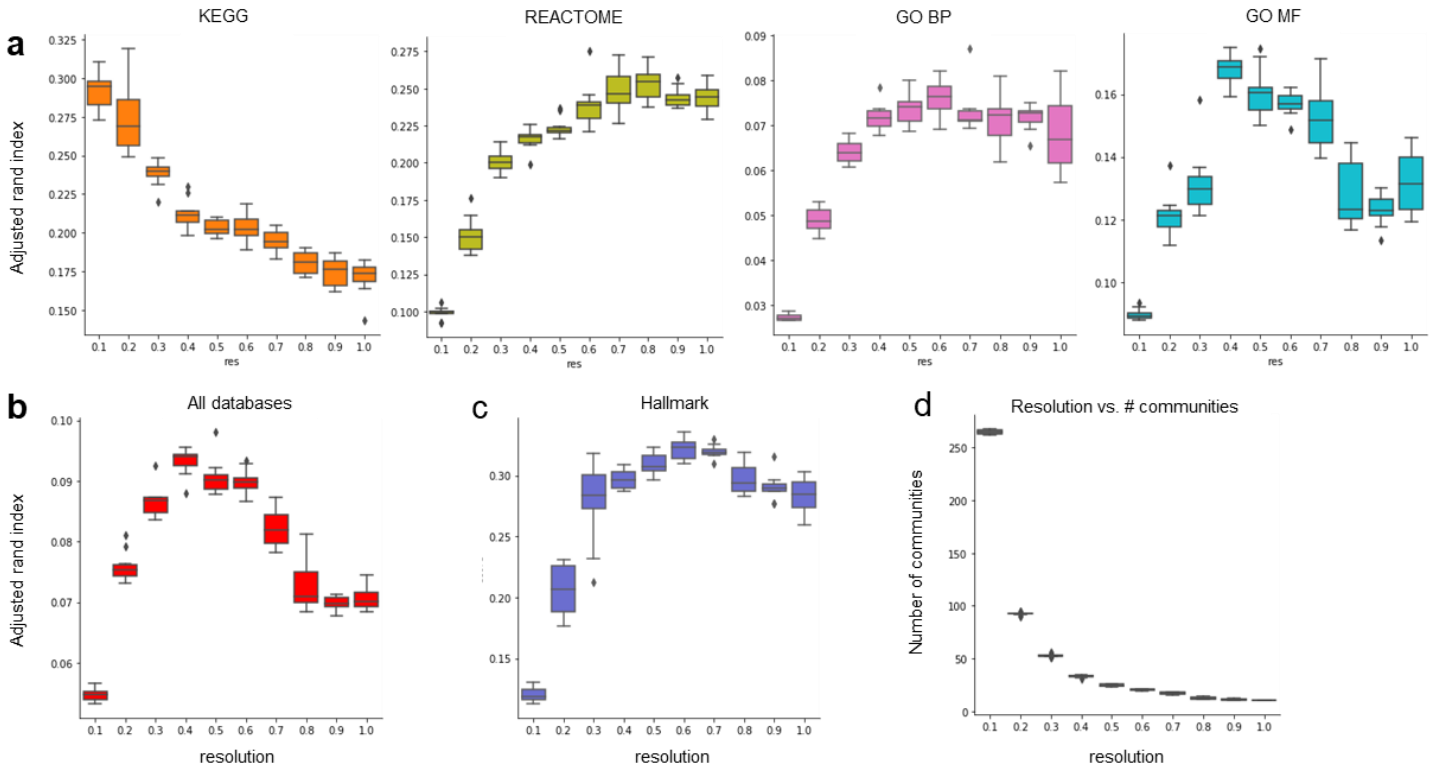
NMI results using Overlap Coefficient Edges

Methods	Kegg	Reactome	GO BP	GO MF
Agglomerative	0.661	0.1837	0.2011	0.2310
Spectral	0.6420	0.2785	0.3132	0.1186
CNM	0.4164	0.0880	0.1208	0.35
Louvain	**0.6705**	**0.4573**	**0.3711**	**0.5050**

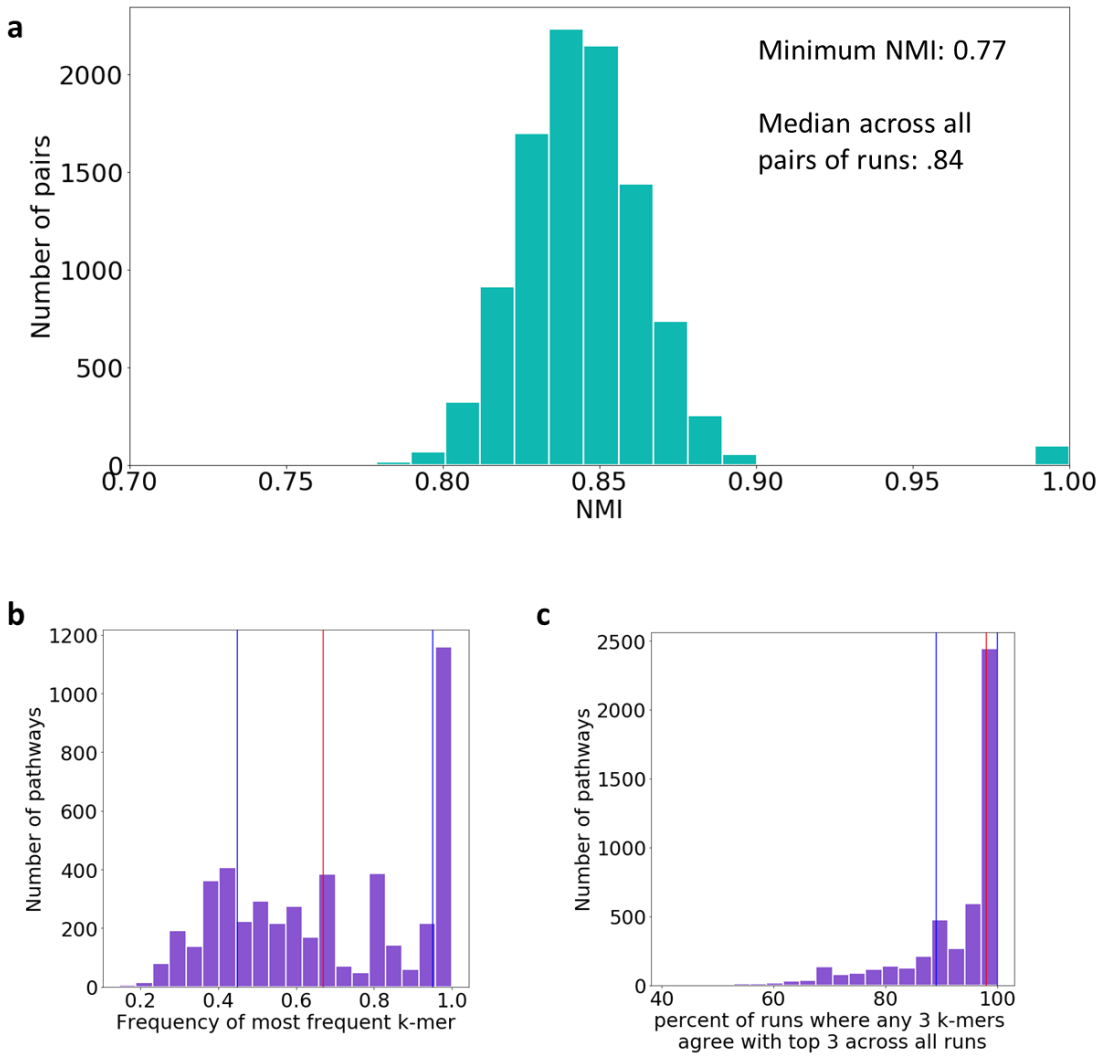
b



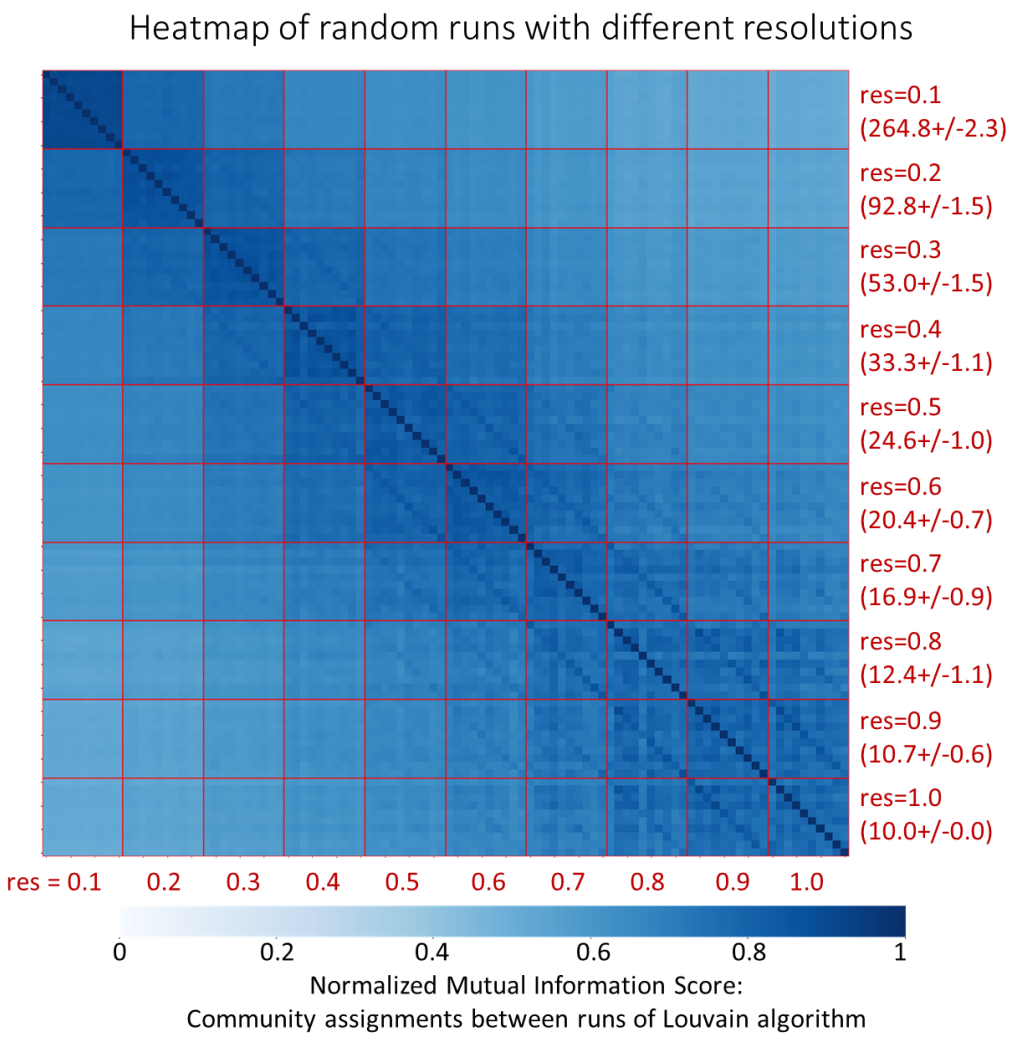
Supplementary Figure 3.3. (a) Comparison of downstream community detection performance for three different edge types in the pathway network. For each edge type, the pathway network is constructed based on the edge type described in the title of the table separately for each pathway database. We next perform the same four community detection methods on each of the four pathway databases and compare assignments to our ground truth labels for each community. **(b)** Comparison of Louvain community detection for alternative edge types using all pathways. While our method uses network edges calculated from Fisher's exact test, we also evaluate the similarity of our results compared with Jaccard index-based edges and overlap-coefficient base edges. For each edge calculation method, we run 100 rounds of the Louvain algorithm and compare pairwise normalized mutual information between all pairs of runs from FET-based edges to the alternative approach.



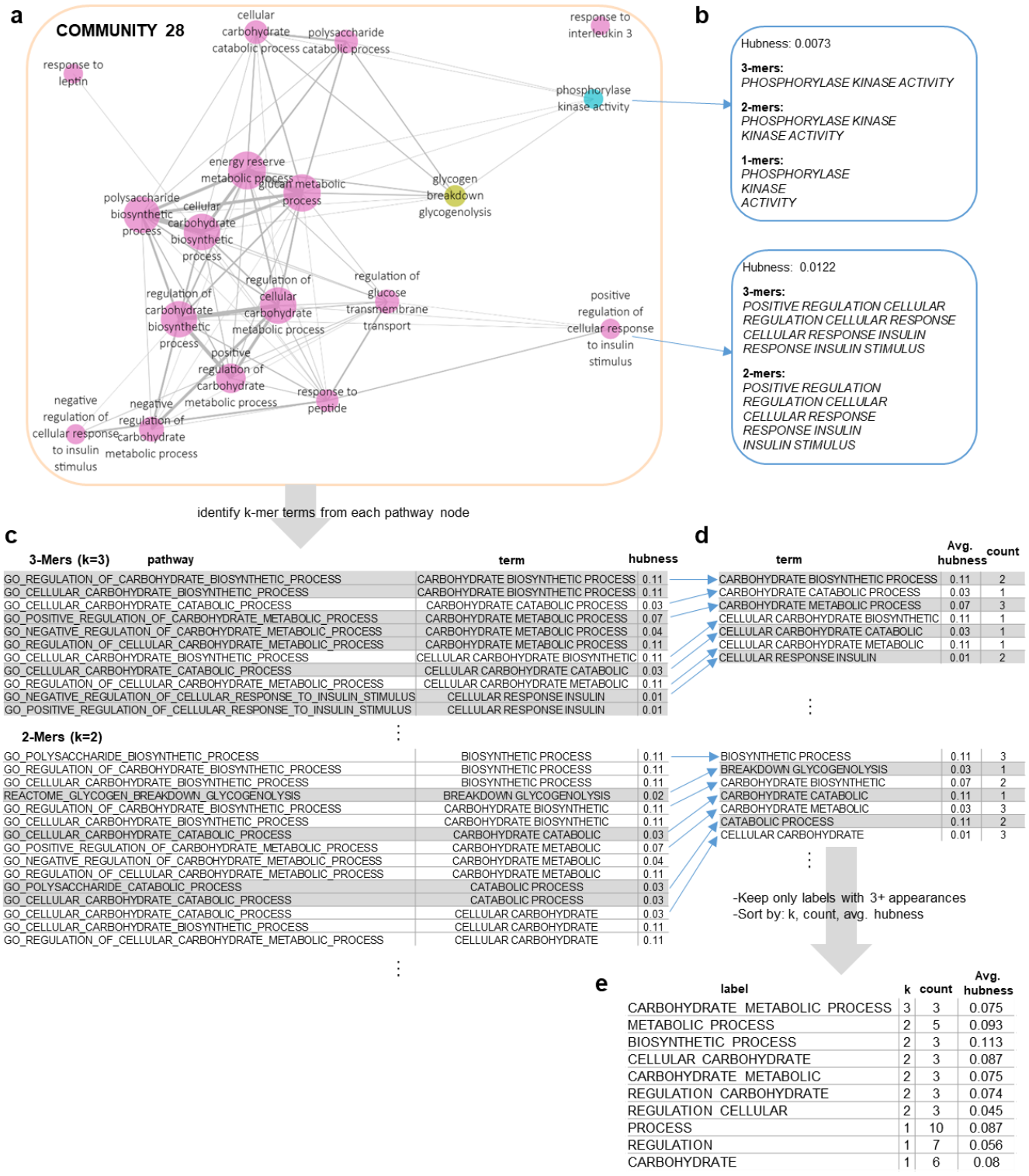
Supplementary Figure 3.4. Comparison of full graph Louvain community detection at different resolutions with curated categories. For the graph of 4,847 pathways from 4 pathway databases, we perform ten random runs of the Louvain community detection algorithm at resolutions ranging from 0.1 to 1, and for each run, obtain a set of learned communities. We then compare the assigned communities with curated categories. **(a)** Separately for each pathway database, we evaluate the adjusted rand index with curated category labels (ignoring all pathways outside of the database). We note that ARI was used because it is sensitive to cluster sizes. **(b)** We combine all 198 curated category labels from each of the four databases and compare our learned communities to these combined labels. **(c)** Among the 550 pathways that are founders of any Hallmark pathway, we compare the learned community label with labeling by Hallmark pathways to which each pathway was a founder. **(d)** Sizes of each community. Each subplot shows box-and-whisker plots based on the 10 repeated trainings for each resolution.



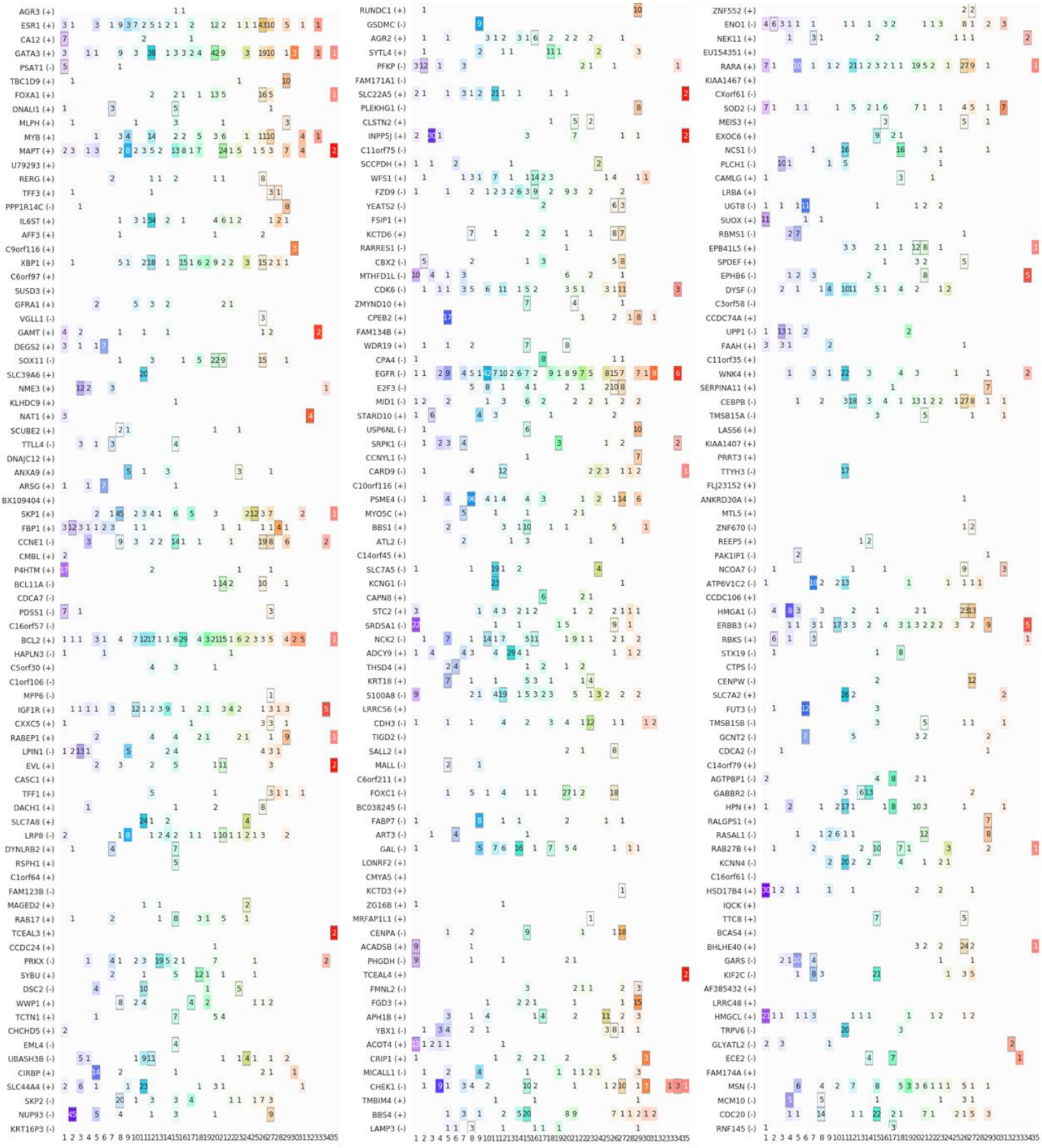
Supplementary Figure 3.5. Consistency of Louvain communities over random initializations. We ran the Louvain algorithm 100 times on the full pathway network using a resolution of 0.4. **(a)** From the 100 sets of Louvain community assignments, the distribution of all pairwise normalized mutual information (NMI) scores between all pairs of runs. **(b)** Frequency of the most frequent k-mer for each pathway across 100 runs. Blue lines indicate the first and third quartile; red line indicates the median frequency of most frequent k-mers across all pathways. **(c)** A more relaxed version of part b. For each pathway, we identify the top three k-mer labels from each run, and then aggregate these to identify the top three k-mers which appeared most commonly in top-three labels across all runs. We then compute, for each pathway, how many runs had any top three k-mer labels from that run overlapping with the top three overall k-mer labels.



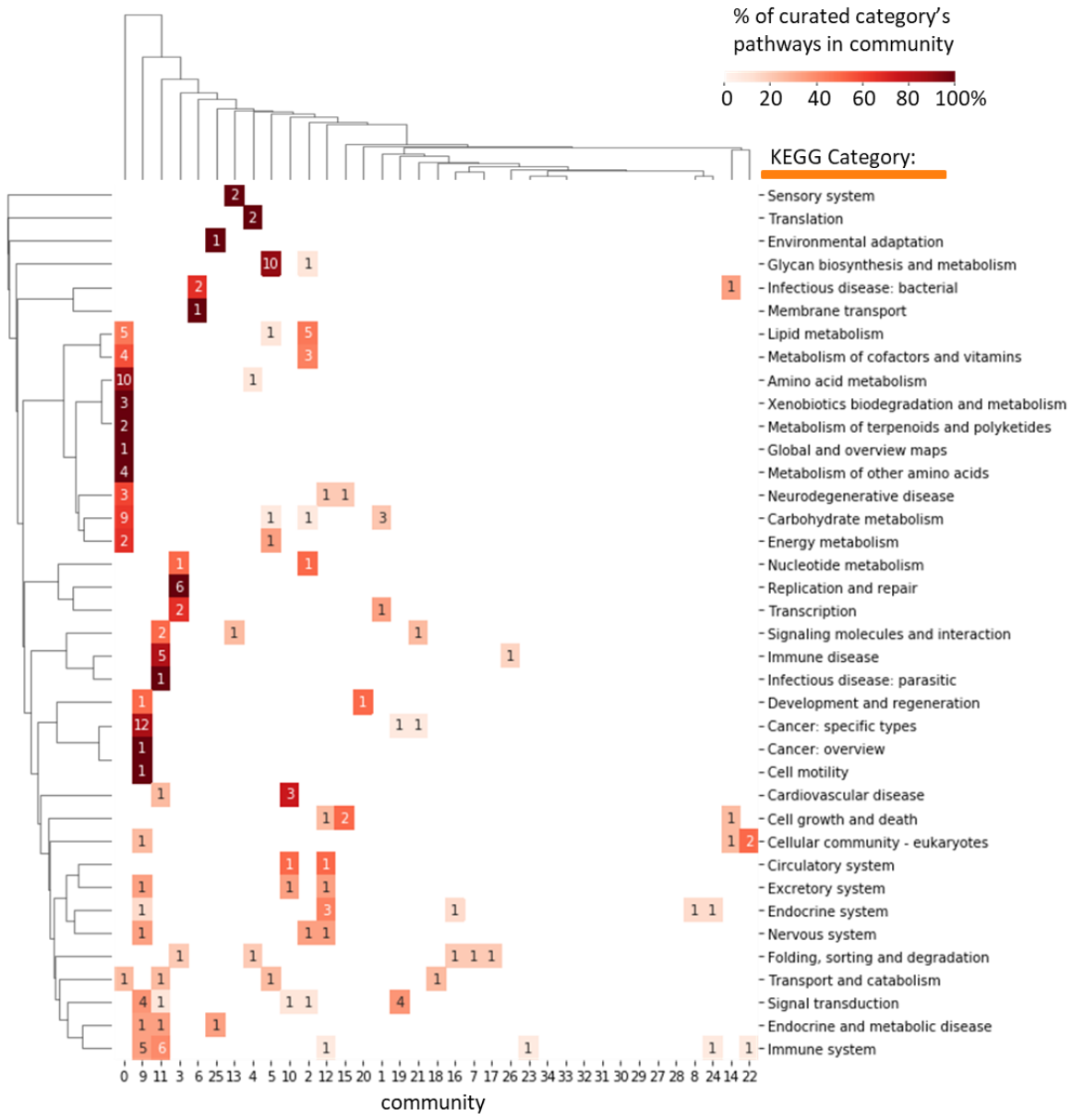
Supplementary Figure 3.6. Comparison of Louvain community detection results across random runs of different resolutions. We re-ran the Louvain algorithm ten times for resolutions ranging from 0.1 to 1. The heatmap indicates pairwise normalized mutual information between community assignments from these different runs. Red lines indicate resolutions for the associated block of random runs. Listed below the resolution, we also indicate the average number of communities (+/- standard deviation) for each resolution parameter evaluated.



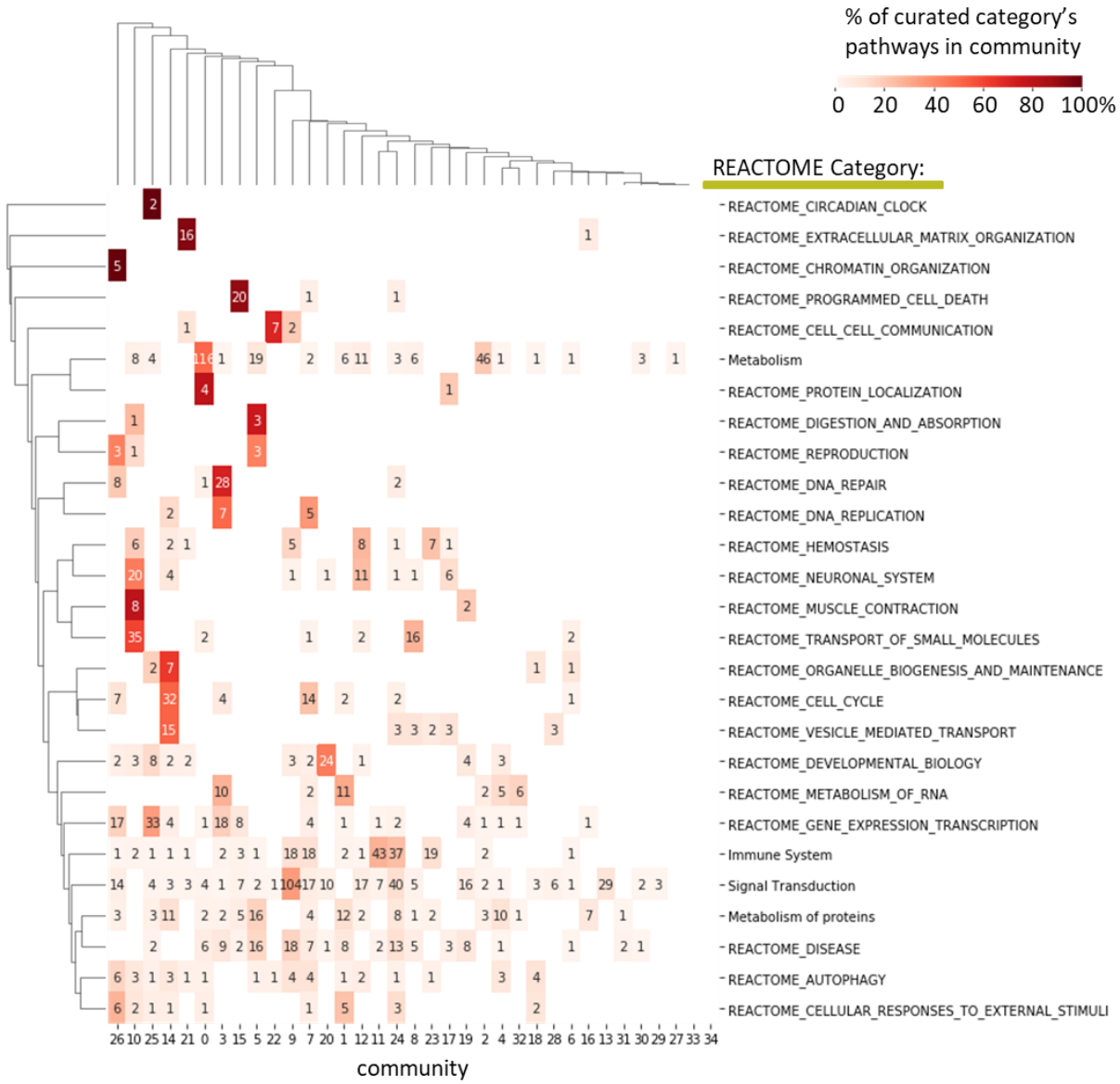
Supplementary Figure 3.7. Example of automatic label generation for Community 28. **(a)** Overview of pathways in Community 28 (as shown in Figure 3.4b). **(b)** Examples of how pathway names are converted to k-mer terms. **(c)** K-mer terms are gathered from each pathway, and within-community hubness is recorded for each pathway. **(d)** We aggregate unique k-mer terms, computing the number of times it appears across the pathway names, and average hubness of those pathways. **(e)** Our rank our final k-mers first by k, then count (filtering out those appearing in fewer than 3 pathways), and finally break ties with average hubness of the pathways containing the label.



Supplementary Figure 3.8. Pathway membership for the top 243 genes with highest differential expression between ER+ and ER- breast cancer samples in the METABRIC dataset. The genes are sorted from lowest to highest p-value (top to bottom, and then left to right). We additionally indicate whether the gene appears in any pathways (split by community). Gray squares on the heat map indicate communities in which the gene is significantly overrepresented, based on a one-way chi square test ($p<.05$).

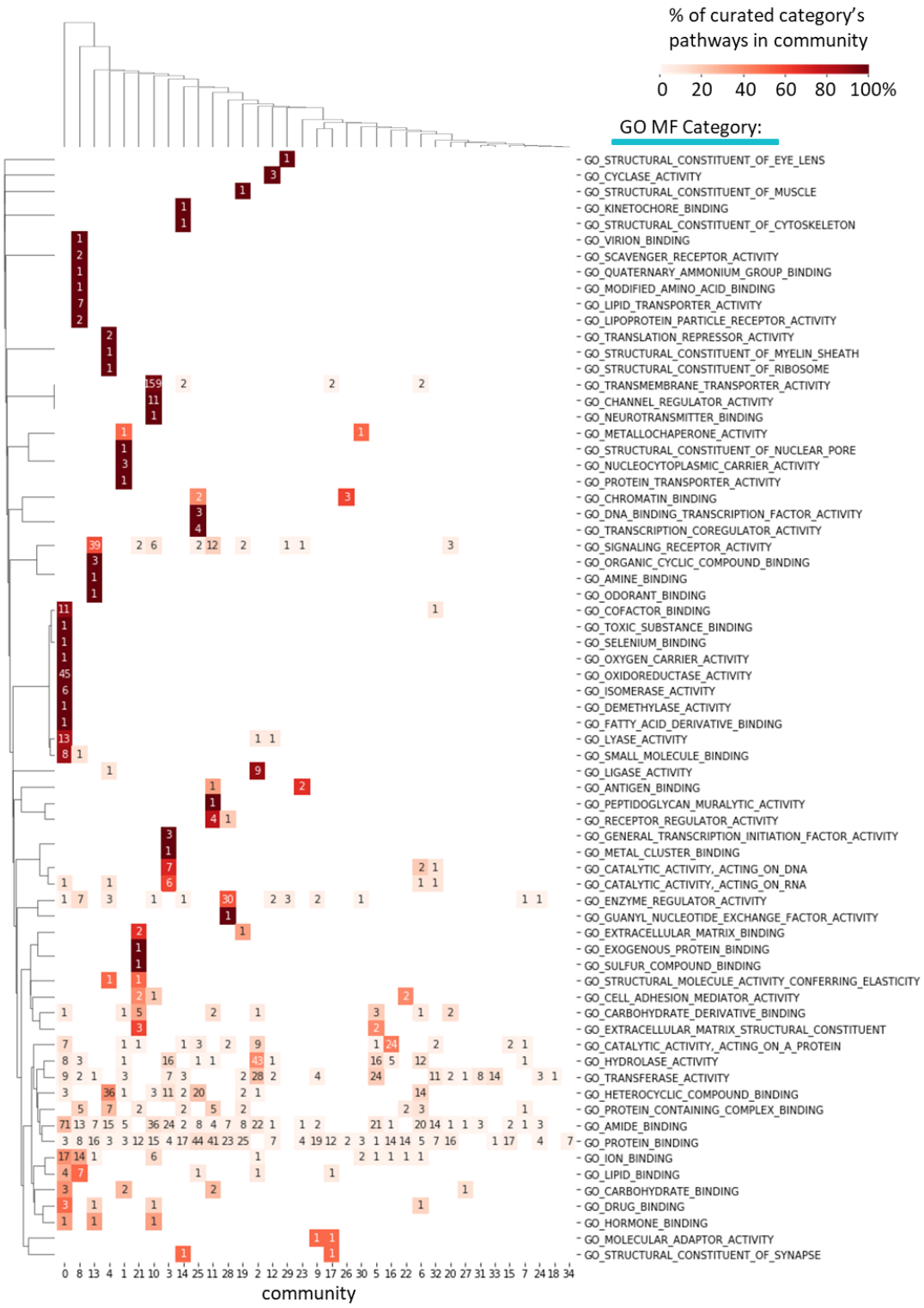


Supplementary Figure 3.9. Clustering consistency between final communities (learned across all 4 pathway databases) and curated KEGG categories. Cells are colored by the percent of curated category's pathway in each community; cell text indicates the exact number of pathways in each community and KEGG category (see Supplementary Figure 3.13 to see these results in the context of all databases).

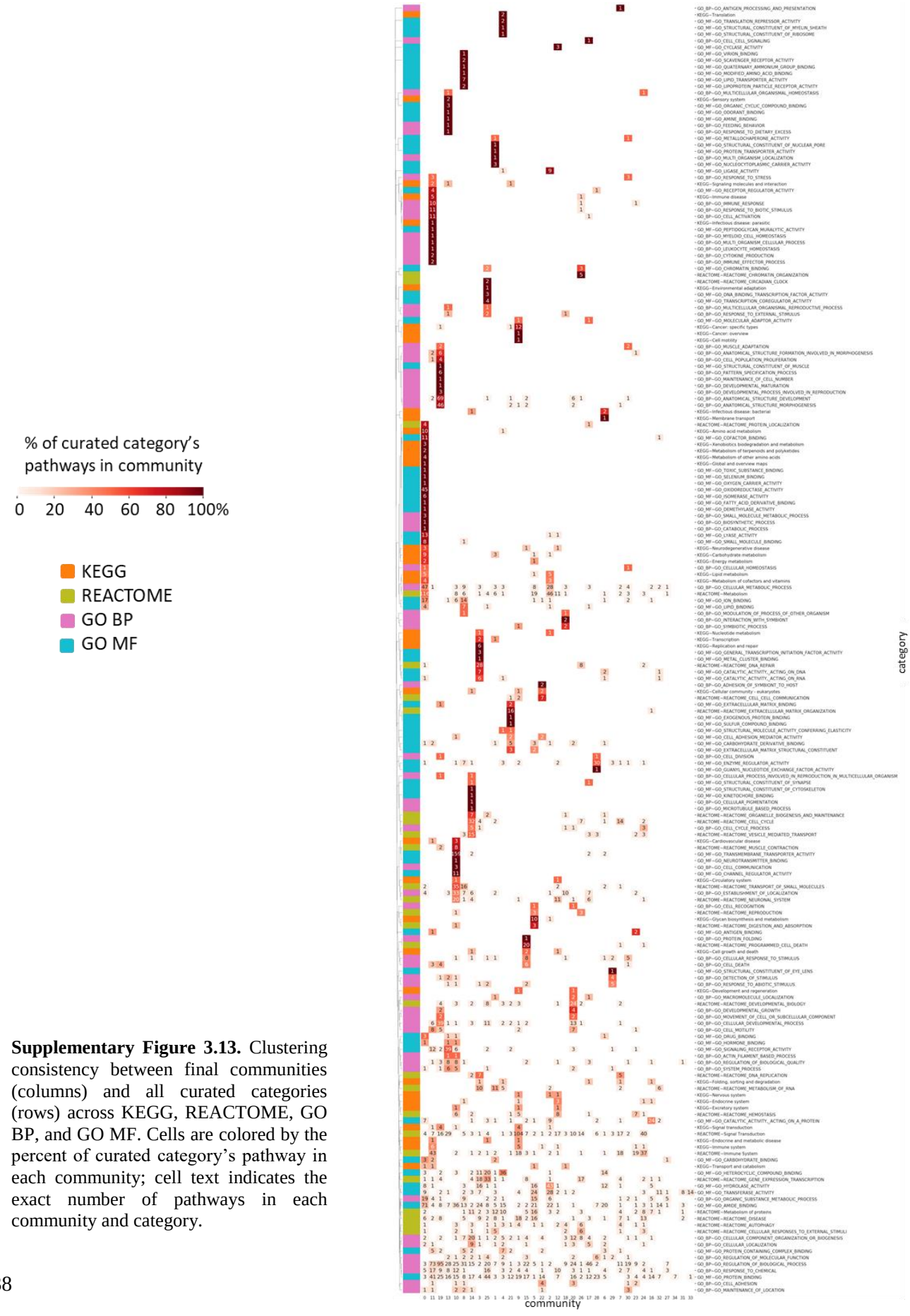


Supplementary Figure 3.10. Clustering consistency between final communities (learned across all 4 pathway databases) and curated REACTOME categories. Cells are colored by the percent of curated category's pathway in each community; cell text indicates the exact number of pathways in each community and REACTOME category (see Supplementary Figure 3.13 to see these results in the context of all databases).

Supplementary Figure 3.11. Clustering consistency between final communities (learned across all 4 pathway databases) and curated GO BP categories. Cells are colored by the percent of curated category's pathway in each community; cell text indicates the exact number of pathways in each community and GO BP category (see Supplementary Figure 3.13 to see these results in the context of all databases).



Supplementary Figure 3.12. Clustering consistency between final communities (learned across all 4 pathway databases) and curated GO MF categories. Cells are colored by the percent of curated category's pathway in each community; cell text indicates the exact number of pathways in each community and GO MF category (see Supplementary Figure 3.13 to see these results in the context of all databases).



Supplementary Figure 3.13. Clustering consistency between final communities (columns) and all curated categories (rows) across KEGG, REACTOME, GO BP, and GO MF. Cells are colored by the percent of curated category's pathway in each community; cell text indicates the exact number of pathways in each community and category.

Chapter 5 Supplementary Materials

SUPPLEMENTARY METHODS

Toy dataset experiment

Data generating process and conditional imputation

We first consider an example with synthetic data where we can stochastically impute samples directly from a known conditional distribution (thereby testing the dynamic feature selection model independently of the performance of an imputer). As illustrated in Supplementary Figure 5.1, the toy dataset contains pairs of correlated features, and a prediction task $y = 4(x^1 + x^2 + \epsilon) + 2 * (x^3 + x^4 + \epsilon) + (x^4 + x^6 + \epsilon)$ where $\epsilon \sim N(0,0.01)$ represents random noise. The six features are normally distributed (mean 0, unit variance), and consist of three pairs of redundant features that are independent of the other pairs ($\rho = .9$ between paired features, as illustrated in Supplementary Figure 5.1a). Thus, we have simulated a case in which we have pairs of redundant features, where within each pair, the features contribute equally to y, and there's a clear ranking of importance between pairs.

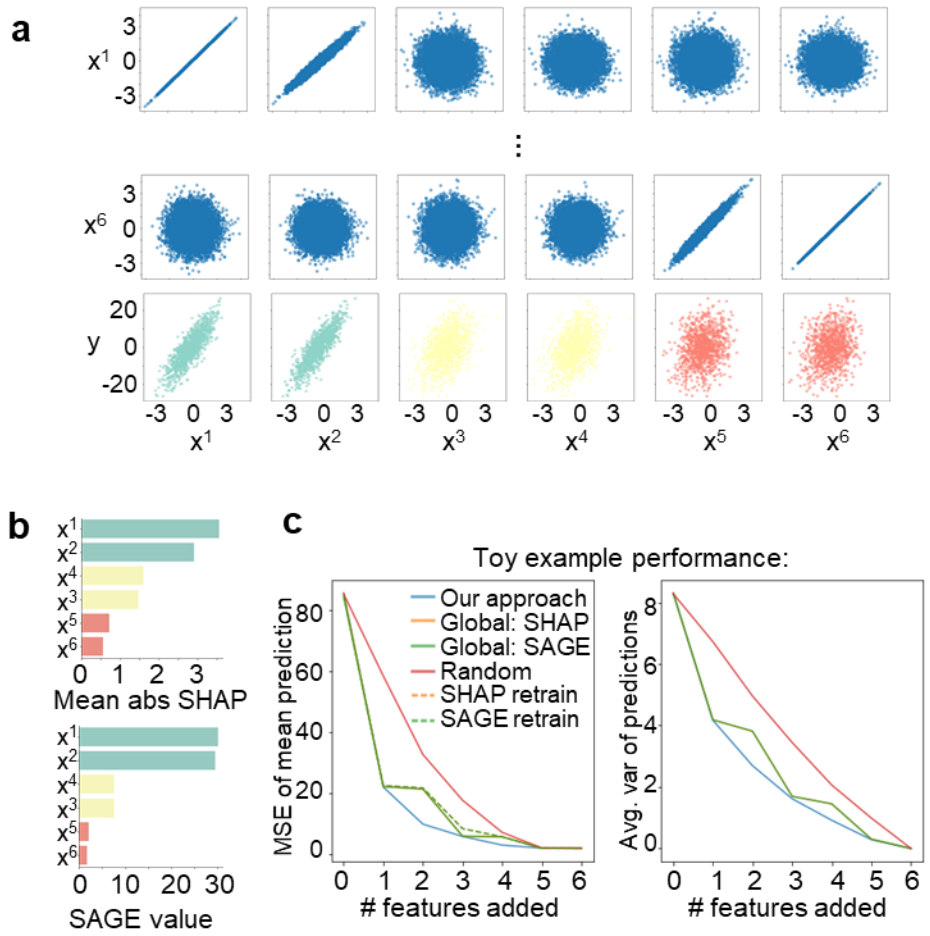
Given this data generating process, it's then straightforward to randomly sample imputations from a partially observed sample: for each correlated pair of features x^a and x^b, if x^a is missing but x^b was observed, we sample from the conditional distribution $x^a | x^b \sim N(.9 * x^b, .19)$ (or vice versa). Otherwise, if both are missing for a given test sample, we randomly sample from the marginal distribution $x^b \sim N(0,1)$, and then sample x^a conditioned on x^b as just described. Given a dataset simulated with this data generating process, we can now mask features and use our approach above to directly sample imputations for the masked features conditioned on the observed ones.

Feature acquisition experiment

For the supervised prediction model, we train an XGBoost model (with default parameters) on a 10,000-sample simulated training set. First, we apply SHAP (a local feature attribution method)[105] and SAGE (a global feature attribution)[127] method to assess how an XGBoost predictor model relies on our simulated features to generate predictions. Consistent with the data generation process, these methods both identify a clear ranking among the pairs of features (Supplementary Figure 5.1b), although the relative importance within pairs slightly differs. In a standard global feature acquisition strategy, features would always be selected based on this global ranking (i.e., $x^1 \rightarrow x^2 \rightarrow x^4 \rightarrow x^3 \rightarrow x^5 \rightarrow x^6$).

We now turn to the evaluation of our approach on this toy example. In Supplementary Figure 5.1c, we show the average performance of our approach when simulating prediction and selection of features from no observed features until all features have been observed. In particular, for each of 1,000 test samples, we repeat the following steps until all features have been observed: (1) we sample 100 imputed samples conditioned on the features observed up until now, (2) we report the mean and variance of the model predictions on these 100 imputations, (3) we then compute SHAP values for each of the imputed examples and select the missing feature with the highest-variance SHAP value, (4) we un-mask the missing feature. In Supplementary Figure 5.1c, we also show how global feature selection policies compare with our context-aware approach (while still using stochastic imputation to generate prediction intervals). We additionally show results for "retrained" global selection approaches because in practice, when using global feature selection, given that the order is fixed, it is reasonable to re-train a prediction model with

the subsets of features that would be used at test time (i.e., a 1-feature model containing the top feature, a 2-feature model containing the top two features, etc.). In this toy example, we find that our dynamic feature selection approach allows our method to more efficiently gather features at test time. By modeling the missing features conditioned on the observed ones, our approach avoids collecting redundant features that would be unlikely to substantially alter the model predictions, unlike global feature selection policies.



Supplementary Figure 5.1. A toy example illustrating the advantage of our method of dynamic feature selection over a fixed ordering. **(a)** Scatter plots illustrating the relationships among features and a target variable in a simulated toy dataset. The feature set contains 3 pairs of correlated features which are independent of other pairs. Our target feature is a function of all features, but relies most heavily on x^1 and x^2 and least on x^5 and x^6. **(b)** Global feature importances for an XGBoost model trained to predict y on the simulated data shown in a. To compute global feature importance, we consider both SHAP (where we compute the mean absolute SHAP value)[105] and SAGE[127]. **(c)** When simulating an interactive risk prediction and dynamic feature selection process on synthetic data illustrated in a, we assess the average performance of our approach on samples (y-axis) when starting at 0 information and then selecting a new feature at each time point (x-axis) with respect to both error (left) and ensembles' prediction variation (right). Baselines: "random" follows our multiple imputation and prediction procedure, but with the next feature selected randomly at each time point. "Global: SHAP" and "Global: SAGE" follow our multiple imputation and prediction procedure, but features are acquired in order of global feature importance as shown in b. "SHAP retrain" and "SAGE retrain" use the same global feature orderings from b, but we retrain models at each feature budget.

UK Biobank (UKB) data overview and preprocessing

UKB participants were enrolled between 2007-2014 from 21 assessment centers across England, Wales, and Scotland. Our study includes all measurements taken during their initial visit, available on December 13th, 2021. During an initial comprehensive visit, hundreds of features were collected, including information about sociodemographic and lifestyle factors, health and medical history, cognitive testing, physical measures (such as composition and hearing tests), and lab tests from biological samples (including blood and urine). We exclude (1) features that are missing in more than 80% of the samples, and (2) highly correlated features with correlations greater than 0.98 (when such correlations existed, we kept just one of the features and removed the others). After excluding features, our UKB dataset has 825 features from numerous categories: demographics, blood assays, health and medical history, lifestyle and environment, physical measures, etc.

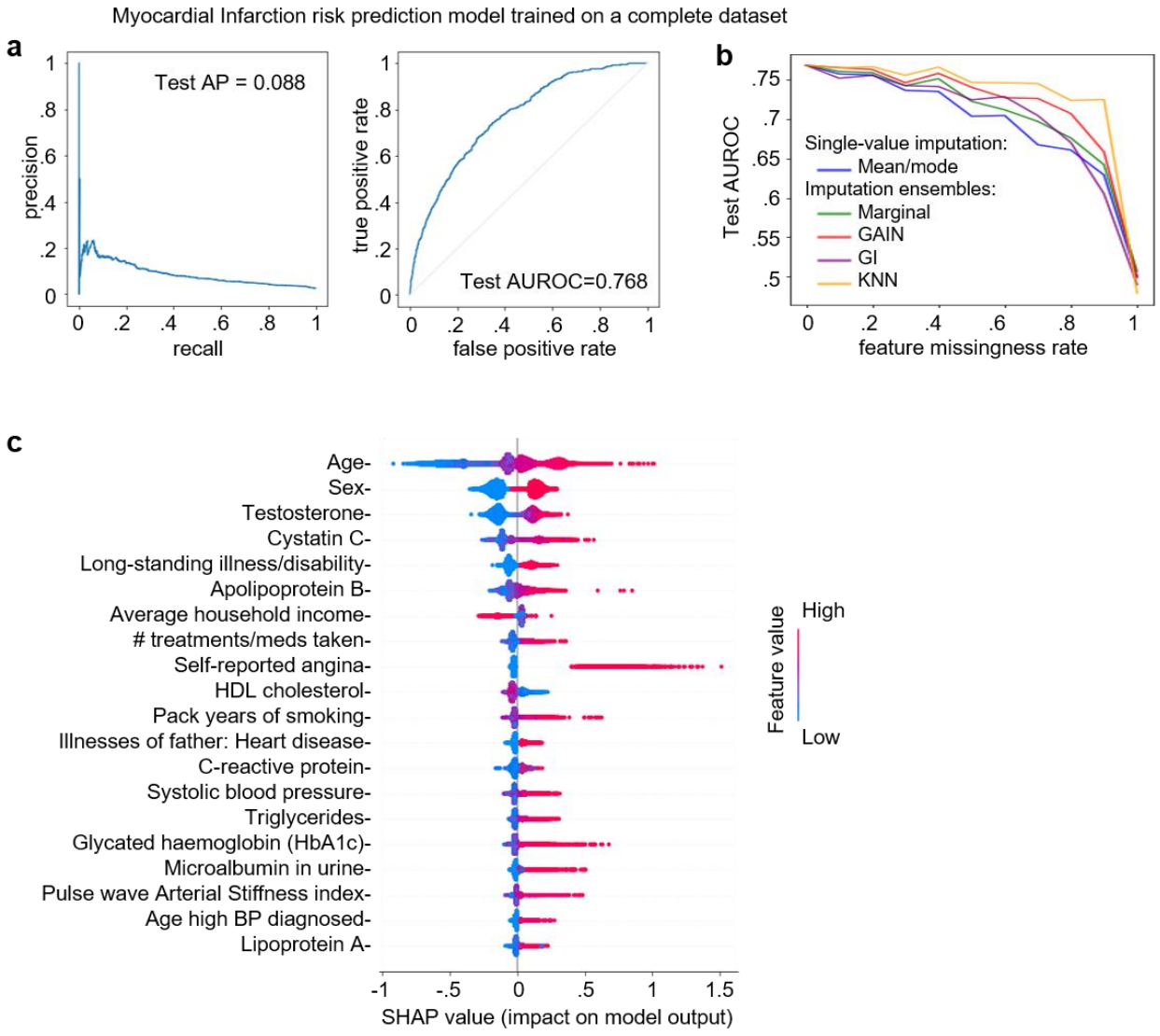
For our analyses, we initially considered seven health outcomes which are provided by the UKB database as "algorithmically defined outcomes," meaning that they are outcomes linked to hospital admissions and death registries (https://biobank. ndph.ox.ac.uk/showcase/label.cgi?id=42): chronic obstructive pulmonary disease, asthma, all-cause dementia, end-stage renal disease, myocardial infarction, all-cause parkinsonism, and stroke. One of the most densely annotated outcomes, and the focus of our analyses, are myocardial infarctions (also known as heart attacks; around 2% incidence over 10 years). For our analyses, we use a ten-year follow-up period as our prediction goal. For each of these conditions, we considered the individual to be a control case if they had no history of the condition during their intake visit as well as no report of the condition within 10 years after the visit, and a positive case if they had no prior history of the condition at their intake and subsequently had a reported incidence of the condition within the next ten years. For a given condition, the label was considered to be unknown if they had a pre-existing report of the condition during their intake, or if they had a report with an unknown time, and such samples were excluded from training our supervised models.

For training and evaluating our models, we used a randomly selected sample of 100,000 individuals, which we divided into training (64,000 individuals), validation (20,000), and test (16,000) splits. In order to train our models, we performed the following additional preprocessing steps: (1) We imputed missing features using MissForest[128], a nonparametric random forest-based multiple imputation method for mixed-type data, (2) we normalized features to have 0 mean and unit-variance (which was necessary to compute distances for KNN-based imputation with similarly scaled dimensions).

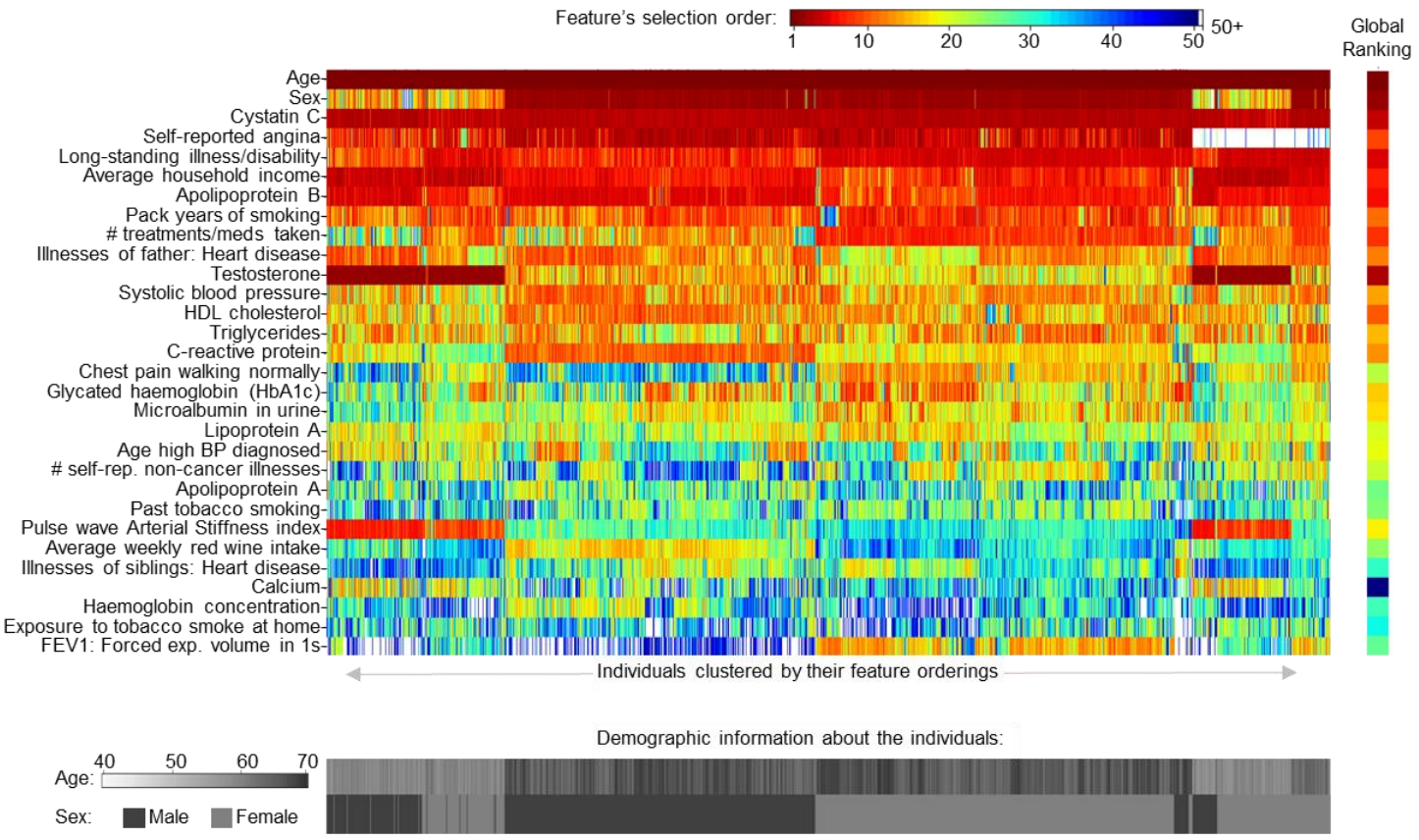
Finally, we generated a reduced feature set for our final models: for each of the seven conditions listed above, we trained a separate XGBoost model to predict whether the condition occurred within 10 years based on the full 825 feature set. We then applied SHAP to each of the seven models and found that many of the features played no significant role among any of the models (SHAP values of 0 across all samples). Thus, to provide a more reasonable starting point for our models, we chose a reduced feature set of 252 features consisting of all features that had a mean absolute SHAP value of at least 0.001 for at least one of seven outcome models. Our final data for the myocardial infarction risk prediction experiments consisted of 62,444 training samples (from the initial 64,000 training samples, we excluded samples with unknown labels) with the 252 features described above, and 15,612 test samples which were used to evaluate the effectiveness of our approach. In our experiments, the 62,444-sample training set is used for both fitting the XGBoost model, as well as the KNN-based imputation approach.

UK Biobank experiment details

For our feature imputer, we use KNN-based imputation with an ensemble size of 100. In particular, we use an imputation ensemble size of 100, meaning that for a given partially observed test sample, we identify the 100 nearest neighbors in the training set (with distances computed in Euclidian space for normalized features based on the observed features only). We then generate an ensemble of imputations where, for each sample in the ensemble, the observed features are kept as is, and the remaining unobserved features are imputed from the record of the selected neighbor. For our XGBoost risk prediction model, we used the implementation from Chen & Guestrin[110] and used default hyperparameter settings. Once fit, this trained model was used as the fixed supervised model across all feature collection strategies.



(Previous page) **Supplementary Figure 5.2**. Summary of our 10-year myocardial infarction XGBoost model. (a) Test set performance when considering the full feature set without missingness. (**b**) Simulating MCAR missingness at different rates, we compare the effectiveness of different imputation strategies. For each strategy (except mean/mode), we generate an imputation ensemble for each sample (100 imputed samples for each real sample) and consider the average of model predictions across the ensemble as our final risk estimate. For mean/mode imputation, we simply impute missing features with the mean (for continuous features) or mode (for binary features). We also evaluate the following strategies: marginal (uniformly sampling the feature's value across the entire training set), GAIN (a neural network-based approach by Yoon et al.[120]), GI (a neural network based approach by Kachuee et al.[121]), and KNN, our final selected approach of identifying nearest neighbors in the training set based on observed features and directly using features observed in those neighbors. (**c**) SHAP summary plot for the trained model: distribution of training set values vs. their impact on the model output (SHAP values). The features are sorted by mean absolute SHAP value, and this ranking is used as our feature ordering for the fixed feature selection strategy.



Supplementary Figure 5.3. Feature orderings across all samples in the test set. In the top heat map, each column represents a sample in the test set, and the cell color indicates the order in which features were collected (dark red for a sample's first selected feature, dark blue for the 50th, and white if the feature was selected after the top 50 features for that sample). Below the main heatmap, we also show each sample's age and sex for to highlight some possible relationships between collected features and the subsequent ordering of later features (e.g., for younger subjects, our approach tends to select testosterone early instead of sex). To the right, we also show the global ranking of features which is used for the fixed global feature ordering baseline.