Quantifying Confounding Effect of Co-pathologies in Neurodegenerative Diseases with sensGAN

Yifan Lin[1], Taek Son[1], Kevin Z Lin[1]

University of Washington
Seattle, WA 98107, USA
{yifanlin, sst91, kzlin}@uw.edu

Abstract. Single-cell transcriptomics of Alzheimer's disease (AD) is confounded by unmeasured co-pathologies that can obscure AD-specific signals. We introduce sensGAN, a sensitivity analysis based on generative adversarial networks that learns the worst-case unmeasured confounders and quantifies their effects on differential expression. A generator proposes the unmeasured confounder, while two predictor nets for case-control status and negative-binomial gene counts adversarially constrain McFadden R^2 predictive gains while minimizing the genes that remain significant. [[KZL: Mention we did simulations to compare against other methods, and benchmarked our method against a lupus dataset – this one is the proof of principle]] We then apply sensGAN to microglia profiles from 78 SEA-AD donors to uncover differential genes based on Alzheimer disease burden. Our analysis distinguishes genes that are more specifically responding to AD pathology from genes that are likely responding to unmeasured co-pathologies. By ranking each gene's relevance to unmeasured co-pathology, sensGAN sharpens therapeutic-target discovery and generalizes to other high-dimensional omics studies. [[KZL: Replace this with a finding about – we find genes relevant to cell motility are robustly differentially expressed, even in the presence of unmeausured confounders. (We need another "result" sentence about the genes that cross the threshold – is about what kind of pathway is technically significant without any confounders, but our analysis reveals that these genes for this pathway could become insignificant with some unmeasured confounders)]] Bridging causal sensitivity analysis with deep learning, sensGAN offers a principled tool for disentangling disease-specific biology amid pervasive confounding.

Keywords: asdf

1 Introduction

Alzheimer's disease has no cure, and despite intense investment, clinical advances remain modest. Microglia are a central focus of therapeutic efforts: they are repeatedly implicated by human genetics and, as the resident immune cells of the brain, are primary actors in sensing injury and clearing misfolded proteins such as plaques and tangles. Many investigational strategies have therefore targeted microglial inflammation and clearance pathways with the aim of reducing neurotoxic burden. Yet these approaches have not delivered durable benefit. A leading explanation is the high prevalence of co-pathologies—additional neurodegenerative processes that co-occur with Alzheimer's and modulate microglial responses, altering the apparent aggressiveness and toxicity of proteinopathies. This landscape motivates a different question: how can we reason about microglial function when some disease drivers are unknown or unmeasured? In other words, can we characterize microglial programs in a way that is robust to co-pathologies that studies did not (or could not) quantify?

Our work is primarily motivated by single-cell RNA-seq analyses aimed to delineate how microglial states change across disease. A key challenge is confounding: the studies we integrate differ in clinical assessments and tissue workflows, and many relevant co-pathologies are not yet systematically measured across cohorts. For example, community efforts only recently began tracking certain proteinopathies in a harmonized way, even though these processes can strongly influence the spread of hallmark Alzheimer's lesions. This motivates a shift in reporting. In addition to listing differentially expressed genes, we propose that studies should quantify how much an unmeasured co-pathology would need to explain the observed signal for each gene. Concretely, a robust panel that guides downstream biology should pair (1) a standard measure of differential expression significance with (2) a sensitivity metric that states the burden of evidence an unmeasured factor would require to equally account for that gene's association. Such "how strong would the confounder have to be?" summaries turn single-cell findings into testable, co-pathology-aware hypotheses for future experiments and therapeutic design.

We note that while our work focuses on microglial functions in Alzheimer's disease, our method will also applicable to any complex human diseases where co-occurring diseases or unmeasured environmental effects can hinder the biological progress.

1.1 Scientific relevance

[[KZL: First paragraph: About prevalance of co-occurring diseases. I'll rework this]] However, a major challenge lies in unmeasured co-pathologies – other neurodegenerative diseases (ND) that commonly co-occur with AD [17].For instance, the aged brains frequently exhibit multiple pathologies, rather than a single hallmark pathology. It could range from low/intermediate levels of additional pathologies to mixed severe pathologies [12]. In late-onset AD, only 31% of cases are described with AD-specific signatures [20]. The difficulty in measuring biomarkers for different NDs and the evolving definitions of NDs [9] pose substantial challenges to collecting co-pathology information. The diagnosis of AD and additional lower / intermediate pathologies are associated in a complex way (Fig. 1A). Although some donors with AD diagnoses demonstrate hallmark AD pathologies such as amyloid plaques and neurofibrillary tangles, others show additional pathologies with known associations with other neurodegenerative diseases. xx% of donors exhibit hippocampal sclerosis, among which yy% of donors are diagnosed with FTLD-TDP-43. CVD and LATE-NC, discovered in zz% AD donors, also show up in donors diagnosed with mixed AD-DLB, LBD, FTLD-tau, FTLD-TDP-43, and AGD. The complex association makes discrete diagnoses difficult, even with AD diagnoses, differences in neuropathologic burden complicate analysis to discover biomarkers specific to AD pathologies.

[[KZL: Second paragraph: Short paragraph about TDP-43 as an example of known co-pathology that impacts AD, but only added into standard clinical reporting guidelines recently. Pose the question – how would we know this doesn't happen again? I'll rework this]]

[[KZL: Third paragraph: Talk about the study about Cornfield's smoking on disease, and although there was no causal evidence, the statistical evidence indicated it's likely causal since a covariate would need to explain so much of the disease. We want to emulate this idea. I'll rework this]]

[[KZL: Fourth paragraph: About what "bad things happen" without adjustment. I'll rework this]] These co-pathologies can distort transcriptomic profiles and, when unaccounted for, introduce bias and confounding

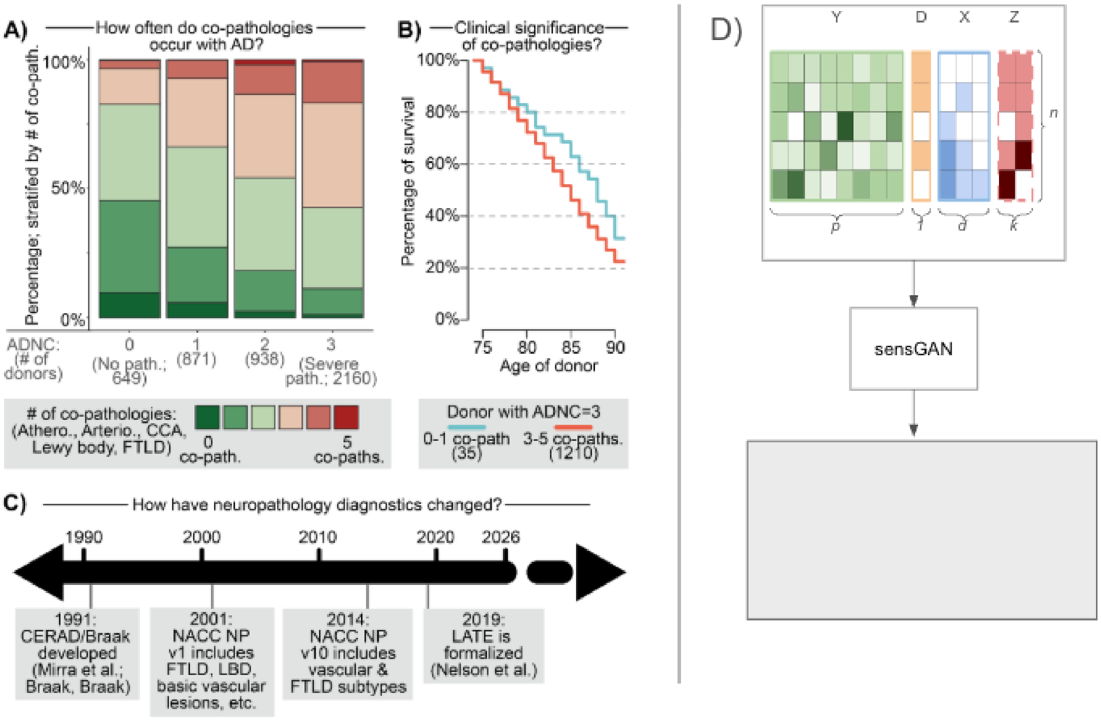


Fig. 1. asdf.

in biomarker discovery efforts [21]. [[KZL: I'll find some papers to cite here about confounding effects. I'll put in references starting from SVA and batch-effects]]

Co-pathology references: [13,1,15,16]

1.2 Existing computational methods and their limitations

Confounder adjustment with high-dimensional outcomes has been an important topic in statistics and genomics in recent years. To characterize the confounding effects, the pioneering work in this field assumes a linear model

$$Y = DT + XB + Z\Gamma + E,$$

where $Y \in \mathbb{R}^{n\times p}$ is the gene expression matrix, $D \in \mathbb{R}^{n\times 1}$ is the case-control vector, $T \in \mathbb{R}^{1\times p}$ is the direct effect to be estimated, $X \in \mathbb{R}^{n\times d}$ is the measured covariate matrix, $B \in \mathbb{R}^{d\times p}$ is the additional effects to be estimated, $Z \in \mathbb{R}^{n\times k}$ is the latent factor matrix, $\Gamma \in \mathbb{R}^{k\times p}$ is the latent factor loading, and $E \in \mathbb{R}^{n\times p}$ is the additive noise (Fig. 1A). However, since Z is unobserved, we are forced to estimate a restricted model:

$$Y = D\widehat{T}_{\text{res}} + X\widehat{B}_{\text{res}} + E_{\text{res}}$$

We highlight two computational frameworks that address unmeasured confounders that motivate our proposed method, sensGAN.

Parameterization of the omitted variable bias. One framework investigates how a hypothetical unmeasured confounder impacts the estimated coefficients. This is called *omitted variable bias* (OVB), where the sensitivity analysis addresses how "extreme" an unmeasured confounder would need to be to nullify the significance of a treatment-outcome relation. For instance, sensmaker [2] defines the bias as $\max_Z(\widehat{T}_{\text{res}} - T)$, OVB for linear models is reparameterized with partial R^2 for the case-control and outcome variable. The appeal of this framework is that these partial R^2's uniquely determine the worse-case change in the estimated coefficients $\widehat{B}$, so practitioners can sweep across partial R^2 values they are interested in to assess if covariates originally deemed significant remain significant. 11/2: Describe Austen plot, it works for general modeling assumption I remembered. [22] However, the applicability of sensmaker is not easily generalizable to a non-linear model and does not handle multiple outcomes. In our modeling of scRNA-seq data, we are interested in how the AD pathology impacts the thousands of genes (i.e., outcomes) via a negative-binomial generalized linear model (GLM) framework. (KZL: 10/25: We'll need to cite the other works in this section. I think one was Austin plots?)

Parameterization of residual variations in outcome not explained by treatment. The other framework assumes that residual variation can be effectively captured by low-dimensional surrogate variables [10,**?**]. This framework have been extended to methods such as CATE [23] and GCATE [6,**?**] to model modern genomic data in the presence unmeasured confounders, i.e., thousands of genes modeled via the negative binomial GLM. While these methods provide an explicit estimate of the unmeasured confounders that can be used in any downstream analyses, they do not provide insight on the spectrum of confounders at varying levels of hypothetical confounding. Additionally, the estimated confounders are generally only reflective of residual variability in the genes (i.e., outcome) not reflected by the treatment and are independent of the treatment itself. In our context, this means these confounders are unlikely to reflect an unmeasured co-pathology that often co-occurs with AD.

Overview of paper. In response to these limitations with existing methods, we introduce a new framework for applying generative adversarial network (GAN) in genomic studies with unmeasured confounding. Our proposed method, sensGAN, (1) provides a diagnostic tool for the robustness (not sure if this is the right wording) of differentially expressed gene biomarkers, and (2) adjusts the DE p-values for unmeasured copathologies to identify gene biomarkers responding directly to AD-specific pathologies. We demonstrate the effectiveness of sensGAN through benchmarking on the simulated dataset, comparing its performance with existing relevant methods (causarray). Next, we apply the method to pseudobulk single-cell genomic studies: AAA and BBB.

2 Overview of sensGAN

The purpose of this paper is to develop a valid framework for estimating the worst-case confounder while constraining predictive gains for the treatment (case-control) variable and high-dimensional results (gene expression). Our work is interested in understanding how AD pathology ("treatment") impacts a cell's function, measured through gene expression ("outcome"). We use the term "confounder" to refer to a broad category of latent variables, including both mediators and confounders, as defined in the context of causal inference literature. In our AD context, we interpret the confounder to be co-occurring neurodegenerative diseases since studies have demonstrate its profound impact on the cellular responses to AD pathology. The "worst-case confounder" in our paper refers to the confounder that inflates the largest number of gene p-values above the significance threshold, i.e., nullifies the largest number of significant genes.

Our method, **sensGAN** (Sensitivity Generative Adversarial Network), is an estimation framework based on GAN that (1) accommodates general relationships among observed covariates, multiple outcomes, and unmeasured confounders, (2) provides the worst-case confounders and quantifies their confounding effects, and (3) leveraging a deep-learning framework to efficiently search for the worst-case confounders for a spectrum of confounding effects. We explain each aspect separately below.

2.1 Quantifying impact of unmeasured confounder with predictive gain.

One of the most important ingredients of our framework is a precise quantification of how an unmeasured confounder Z relates to the treatment D or outcome Y. Our method is capable of identifying a categorical or numerical confounder $Z \in \mathbb{R}^{n\times k}$, which improves prediction of the treatment D or outcome Y. In linear models as analyzed in sensmarker [2], partial R^2 measures the predictive power of a variable, controlling for the other covariates in the model. Partial R^2 ranges from 0 to 1: a partial R^2 of 1 means that $\boldsymbol{Z}$ explains 100% of the remaining variance in other covariates. To emulate similar interpretations and properties of partial R^2 in linear models, we adopted the normalized predictive gains of predicting D and Y, notated as κ and η. They are based on deviance-based partial R^2 (DBPR) of predicting D and Y ($R^2_{D,\mathrm{Dev}}$, $R^2_{Y,\mathrm{Dev}}$) [14,**?**].

The normalized predictive gain of predicting D is defined as the relations between three models: (1) the full model $\widehat{f}_D$ that leverages an unmeasured confounder $\widehat{Z}$ that is most predictive of D (i.e., the "most powerful" confounder that explains the most variability of the residuals), (2) the baseline model f_D that predicts D without considering any confounding variable, and (3) an arbitrary model $\widetilde{f}_D$ that predicts D using any proposed confounder $\widetilde{Z}$. We quantify the normalized predictive gain of $\widetilde{f}_D$ based on its performance relative to $\widehat{f}_D$ and f_D:

$$\mathrm{Dev}_f = 2\left[\ell(\text{saturated}) - \ell(f)\right] \tag{1}$$

$$\widehat{R}^2_{D,\mathrm{Dev}} = \frac{\mathrm{Dev}_{f_D} - \mathrm{Dev}_{\widehat{f}_D}}{\mathrm{Dev}_{f_D} + \varepsilon}, \quad \widetilde{R}^2_{D,\mathrm{Dev}} = \frac{\mathrm{Dev}_{f_D} - \mathrm{Dev}_{\widetilde{f}_D}}{\mathrm{Dev}_{f_D} + \varepsilon}, \quad \kappa = \frac{\widetilde{R}^2_{D,\mathrm{Dev}}}{\widehat{R}^2_{D,\mathrm{Dev}}}$$

Similarly, the normalized predictive gain of predicting Y is defined as the following: (1) the full model $\widehat{f}_Y$ using the "most powerful" confounder $\widehat{Z}$ that predicts Y, the baseline model f_Y, and an arbitrary model $\widetilde{f}_Y$ using any proposed confounder $\widetilde{Z}$:

$$\widehat{R}^2_{Y,\mathrm{Dev}} = \frac{\mathrm{Dev}_{f_Y} - \mathrm{Dev}_{\widehat{f}_Y}}{\mathrm{Dev}_{f_Y} + \varepsilon}, \quad \widetilde{R}^2_{Y,\mathrm{Dev}} = \frac{\mathrm{Dev}_{f_Y} - \mathrm{Dev}_{\widetilde{f}_Y}}{\mathrm{Dev}_{f_Y} + \varepsilon}, \quad \eta = \frac{\widetilde{R}^2_{Y,\mathrm{Dev}}}{\widehat{R}^2_{Y,\mathrm{Dev}}}$$

In this context, the normalized predictive gain denotes the proportional predictive gain introduced by an arbitrary $\widetilde{Z}$, normalized by the maximal predictive gain introduced by the most powerful $\widehat{Z}$. The normalized predictive gain is bounded by 0 and 1: a normalized predictive gain of 1 means that $\widetilde{Z}$ achieves 100% of the maximal predictive gain. It is also differentiable, allowing for gradient calculation in the ML framework.

2.2 Quantifying "worst-case" p-value for a desired amount of predictive gain

After quantifying confounding effects with certain normalized predictive gains, sensGAN explores the "worst-case" confounder Z under predictive gain constraints. Operationally, this is the confounder that nullifies the largest number of significantly differentiated genes. Given a desired DBPR values κ^* and η^*, we formulate estimating such a worst-case confounder as an optimization under constraints:

$$\max_{\widetilde{Z}} \left\{ \sum_{j=1}^{p} \mathbb{I}\left(\frac{\widetilde{B}_j}{\mathrm{SE}(B_j)} - z_{1-\alpha} > 0 \right) \right\} \quad \text{subject to} \quad \kappa = \kappa^*, \quad \eta = \eta^*, \tag{2}$$

where $z_{1-\alpha}$ is the $(1-\alpha) \times 100$ quantile of a Gaussian distribution, and $\widetilde{B}_j$ and B_j defined as the coefficient for the treatment D when adjusting or not adjusting for the unmeasured confounder Z, respectively. Specifically, in our work, we use negative binomial (NB) GLMs where

$$\begin{aligned} &(\text{Baseline model, } f_Y): \quad Y_j \sim NB(\mu = \exp(XT + DB)), \quad \text{and} \\ &(\text{Adjusted model, } \widehat{f}_Y): \quad Y_j \sim NB(\mu = \exp(X\widetilde{T} + D\widetilde{B} + \widetilde{Z}\widetilde{G})). \end{aligned}$$

To compute the DE p-value of gene j adjusting for $\boldsymbol{Z}$, we assume that $\mathrm{SE}(\widehat{B}_{j,\mathrm{res}}) \approx \mathrm{SE}(\widehat{B}_j)$, so we can model the DE test statistic $|\widetilde{B}_j / \mathrm{SE}(B_j)|$ as a z-score.

2.3 Deploying a deep-learning framework for efficient training of a spectrum of unmeasured confounders

sensGAN overview. To tie all the different components of sensGAN together, we adopt a neural network (NN) systems, sensGAN, to solve the optimization to identify an unmeasured confounder. Motivated by Equation (2), sensGAN poses an optimization problem that eliminates the significance of the largest number of features (genes) associated with the case-control variable while maintaining plausibility by limiting the predictive gains that accounted for the confounding. Here, sensGAN has 3 steps: (1) train a NN to model the baseline predictors; (2) train a NN to estimate the most powerful confounder $\widehat{Z}$; (3) train a the multitask GAN to generate the worst-case $\widetilde{Z}$ given normalized predictive gain constraints, yielding the worst-case significance profile under constraints. We describe each step below:

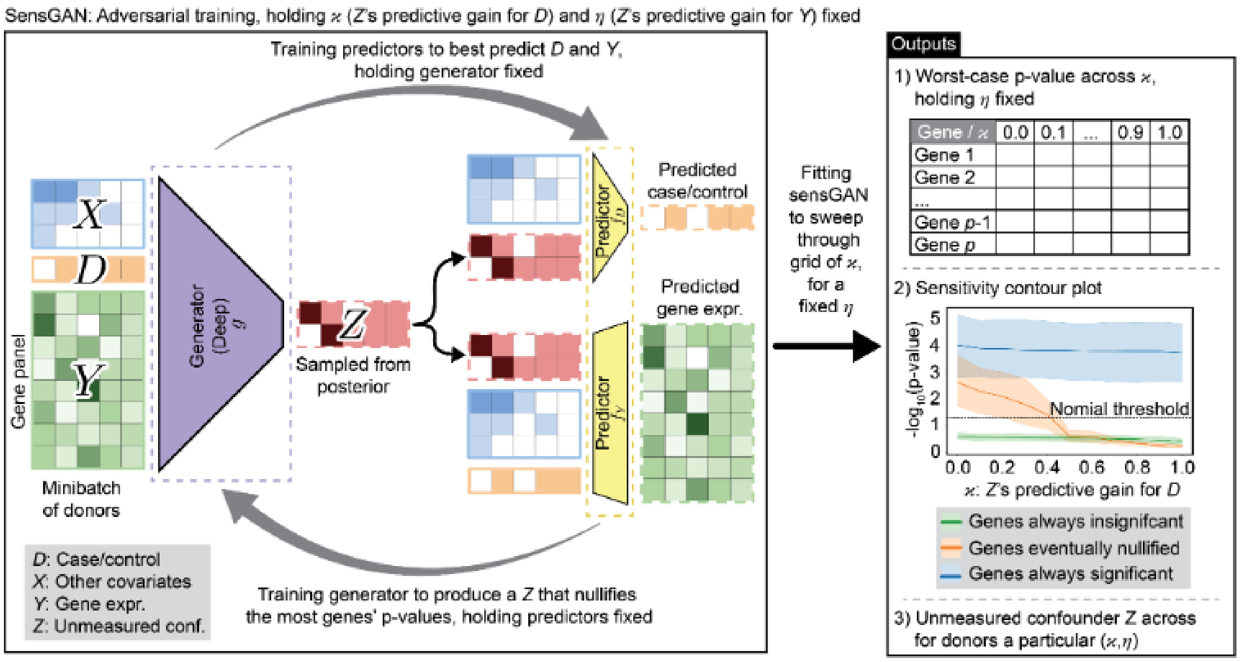

SensGAN: Adversarial training, holding ϰ (Z's predictive gain for D) and η (Z's predictive gain for Y) fixed
Training predictors to best predict D and Y, holding generator fixed
X
D
Y
Gene panel
Minibatch of donors
Generator (Deep) η
Z
Sampled from posterior
Predictor f_D
Predictor f_Y
Predicted case/control
Predicted gene expr.
Training generator to produce a Z that nullifies the most genes' p-values, holding predictors fixed
D: Case/control
X: Other covariates
Y: Gene expr.
Z: Unmeasured conf.
Fitting sensGAN to sweep through grid of ϰ, for a fixed η
Outputs
1) Worst-case p-value across ϰ, holding η fixed
Gene / ϰ
0.0
0.1
...
0.9
1.0
Gene 1
Gene 2
...
Gene p-1
Gene p
2) Sensitivity contour plot
$-\log_{10}$(p-value)
5
4
3
2
1
0
Nomial threshold
0.0
0.2
0.4
0.6
0.8
1.0
ϰ: Z's predictive gain for D
Genes always insignifcant
Genes eventually nullified
Genes always significant
3) Unmeasured confounder Z across for donors a particular (ϰ,η)

Fig. 2. asdf.

Baseline predictors. In the first step, we trained baseline predictor NN system with the objective to maximizing likelihood, i.e. minimizing the likelihood loss, notated as $\mathcal{L}_L$. This step behaves like solving for the MLEs of the logistic and negative binomial (NB) models. This system takes in X, D, Y to train 1-layer FC NNs, outputting the baseline prediction models. Models are later used for calculating maximum predictive gains in Step 2 and applying the DBPR constraints in Step 3.

Most powerful confounder. The most powerful confounder explained away the most residuals in D and Y unexplained by the baseline predictors. In the second step, we trained an NN system to identify the most powerful confounder, $\widehat{Z}$, and the corresponding predictors that took $\widehat{Z}$ as one of the inputs. The NN system consists of 2 components: (1) The predictor networks for D and Y minimized likelihood losses ($\mathcal{L}_L$). They were similar to the predictors in step 1, but they additionally took $\widehat{Z}$ as input; (2) The generator network minimized likelihood losses ($\mathcal{L}_L$) and the regularization loss ($\mathcal{L}_Z$). It took the data and learned predictors as inputs, and outputted the confounder $\widehat{Z}$ that explained away the most residual. $\mathcal{L}_Z$ was the regularization term with respect to $\widehat{Z}$ that considered the prevalence, entropy, and correlation with D to prevent label leakage. This step started with a random initialization and ultimately identified (1) the most powerful confounder, $\widehat{Z}$; (2) the generator that found such a confounder; and (3) the most powerful predictors that took the most powerful confounder as one of the inputs to yield the maximum likelihood.

Then, we quantified the confounding effect of the most powerful confounder by normalized predictive gains. Specifically, we calculated the DBPRs with the baseline predictors from step 1 and the most powerful predictors just learned in step 2, yielding the maximal predictive gains.

Plausible confounders constrained by predictive gains. The most powerful confounders explain away too much residual in D and Y, which is typically unlikely to be a real measure like co-pathology. Therefore, we propose the plausible confounder: it has a non-negative predictive gain, which is lower than that of the most powerful confounder. To identify the plausible unmeasured confounder Z with a specific predictive gain, we train the GAN system to nullify the largest number of significant genes associated with D. The architecture of the GAN system is similar to the NN system in step 2, but the predictor and the generator now have adversarial objectives. The predictor maximized likelihood, while the generator constrained predictive gains, which are proportional to likelihood. Among confounders with specific predictive gains, we were interested in the confounder, $\widetilde{Z}$, that resulted in the worst-case gene significance profile. To achieve these objective, the generator optimizes for $\mathcal{L}_{\text{DBPR}} + \mathcal{L}_{\text{pval}}$. The GAN system answered such a question: what is the worst-case gene significance profile while introducing a confounder with certain predictive gains of predicting D and Y? At the end of training, the generator finds the confounder, $\widetilde{Z}$, which resulted in the worst-case gene significance profile while maintaining certain predictive gains.

[[KZL: Things we're missing: We never defined the generator as g. Also double check: 1) we state how exactly the most-powerful Z is used during training g, and 2) Z is resampled]]

To compute the DE p-value of gene j adjusting for $\widetilde{Z}$, we assume that $\text{SE}(\widehat{B}_{j,\text{res}}) \approx \text{SE}(\widehat{B}_j)$, so we can model the DE test statistic $|\widehat{B}_j/\text{SE}(\widehat{B}_{j,\text{res}})|$ as a z-score. Subsequently, we identify the worst-case p-values for all features, a p-value vector $\widetilde{P} \in \mathbb{R}^{p\times 1}$.

$$\text{Generator objective: } \min \left\{ \lambda_{\text{pval}} \cdot \frac{1}{p}\sum_{j=1}^{p} \sigma_\lambda\left(\left|\frac{\widehat{B}_j}{\text{SE}_j}\right| - z_{1-\alpha}\right) + \lambda_Y\left[\left((\bar{\eta} - \eta^*)\, s_Y\right)^2\right] + \lambda_D\left[\left((\kappa - \kappa^*)\, s_D\right)^2\right]\right\}$$

Sequential training of plausible confounders constrained by a spectrum of predictive gains. To span over all plausible confounders with non-negative predictive gains, we trained the model sequentially with a series of κ's and η's from 0 to 1, constraining predictive gains at varying levels from 0 to the most powerful predictive gains calculated in step 2. At the end of the sequential training, the generator found a series of worst-case confounders, which resulted in the worst-case gene significance profile while maintaining varying levels of predictive gains. We believe the impact of confounders would not exceed that of the most impactful measured covariate associated with the case-control variable. Therefore, we fix η for each gene based on a GLM-NB model with this known covariate. We then allow κ to vary from 0 to 1, by any number of levels. For a certain level, $\kappa = \kappa^*$, predictive gain constraint parameters (η, κ^*) are provided to step 3 of the method. As result, we have a series of plausible confounder $\widehat{Z}$ and worst-case p-values $\widetilde{P}$, ready for downstream analyses.

2.4 Downstream analysis: Calibration with measured covariates

[[KZL: I'll talk to you about this on Monday (10/27)]]

2.5 Downstream analysis: Plotting of sensitivity curves

Assuming that any confounder is no stronger than the strongest known covariate [[KZL: We'll rework this once we have the calibration section written]], we fix $\eta = \eta^*$ by the GLM-NB model with these known covariates in application and let κ vary. With $((\kappa_1, \eta^*), (\kappa_2, \eta^*), \ldots, (\kappa_r, \eta^*))$, we performed r rounds of sensGAN trainings. Then, we have a list of p-value vectors $(p_{\kappa_1,\eta^*}, p_{\kappa_2,\eta^*}, \ldots, p_{\kappa_r,\eta^*})$ for all genes, where $(\kappa_1, \kappa_2, \ldots, \kappa_r)$ are bounded by 0 and 1. After that, we plotted the sensitivity contour plots, each curve corresponding to a gene. Figure 1D shows the sensitivity contour plot, which mimicks that proposed by sensmaker[2]. If a gene loses significance at a small κ, it may be more likely to be associated with a confounder. If it remains significant until a large κ, it is more likely to respond directly to the case-control variable.

[[KZL: Make sure we mention the isotonic smoothing]]

3 Experimental Setup

3.1 Simulation Specification

We evaluated the performance of our approach using a simulated dataset. First, we simulated the measured covariate matrix $X \in \mathbb{R}^{n\times d}$ and the unmeasured confounder $Z \in \mathbb{R}^{n\times k}$ independently. Then, we simulated the case-control variable $D \in \mathbb{R}^{n\times 1}$ with a Bernoulli distribution with a logit link while calibrating strengths of X and Z. We first drew $B_1 \in \mathbb{R}^{d\times 1}$ from N(0,1), then rescaled it so that the variance of the XB_1 hit the target. We then chose $G_1 \in \mathbb{R}^{k\times 1}$ so that ZG_1 contributed signals strong enough to be reflected in D. With the Negative Binomial GLM link function, we simulated the log mean of the RNA-seq pseudocount matrix $Y \in \mathbb{R}^{n\times p}$.

$$D \sim \text{Bernoulli}(\sigma(XB_1 + ZG_1))\,,$$

$$\log(\mu) = (X, D)\,B_2 + ZG_2$$

With different combinations of $B_2 \in \mathbb{R}^{(d+1)\times 1}$ and $G_2 \in \mathbb{R}^{k\times 1}$, we simulated 3 categories of genes with varied significance profiles. Genes in category (1) were significantly associated with D and not with Z. Their expressions were significantly associated with D with or without adjusting for Z. Genes in category (2) were significantly associated with Z and not with D. In other words, they were significant without adjustment for Z and became insignificant with adjustment for Z. Genes in category (3) were not significantly associated with D or Z. With and without adjustment for Z, they were insignificant.

Specifically, the simulated dimensions are: $n = 100, p = 100, d = 4, k = 1$.

Three methods are applied to the simulation datasets: (1) GCATE, the unified statistical estimation and inference framework with the Poisson likelihood. (2) RUVr, the statistical method that removes unwanted residuals with high-dimensional data. (3) sensGAN, our proposed method that exploits the theory of sensitivity analysis and produces a spectrum of plausible Z with varying predictive gains.

To evaluate different methods, we summarized (1) the correlations between the true confounder and the learned confounder, from causarray, RUVr, and sensGAN with varying predictive gains, (2) the contingency table of genes' simulated significances and estimated significances, from baseline GLM model, causarray, RUVr, and sensGAN. sensGAN yields a list of plausible confounders based on predictive gain constraints; therefore, the estimated significances had 3 categories, the same as those simulated.

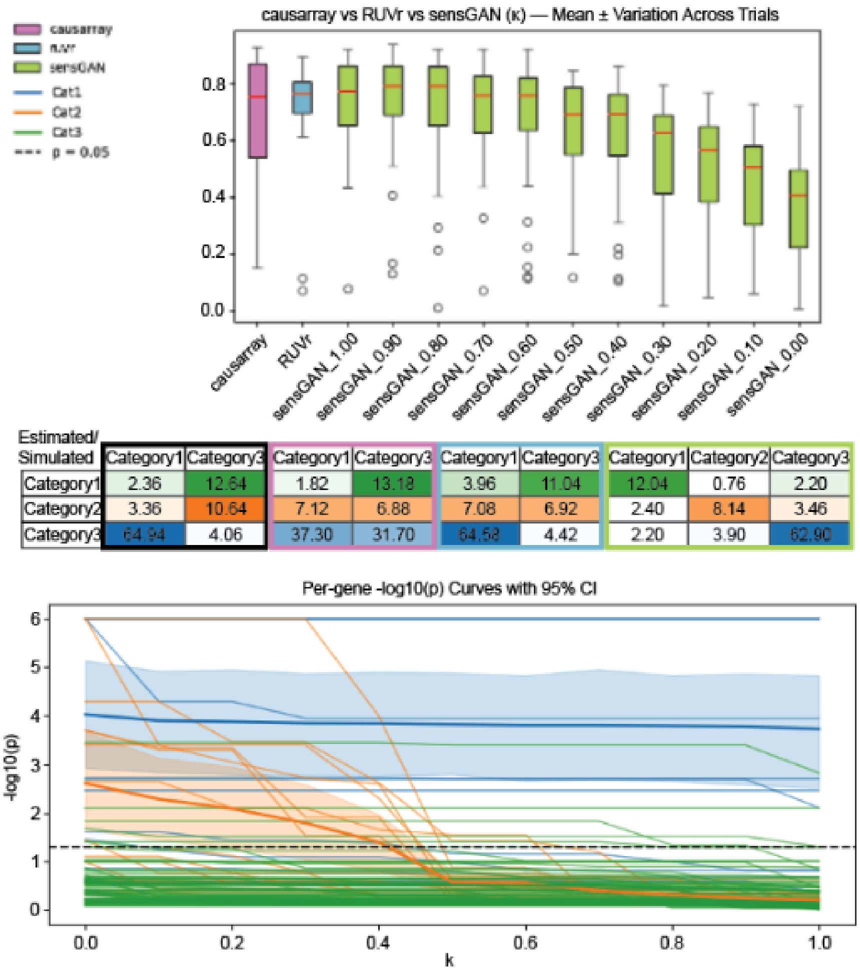
In Figure 3A, we compare the correlations between the learned confounder and the true confounder. The most predictive confounder recovered by sensGAN achieves a mean correlation comparable to those obtained by causarray and RUVr. As the predictive gain of the confounder is gradually reduced from 1 to 0, the mean correlation between the learned and true confounders decreases smoothly from approximately 0.8 to 0.6. These results indicate that sensGAN not only identifies the strongest latent confounding structure—on par with competing approaches—but also generates a continuum of plausible worst-case confounders constrained by user-specified predictive gains. This additional capability enables researchers to calibrate confounder identification and incorporate domain knowledge to focus on the most biologically or scientifically sensible confounders.

With Figure 3B, 3C, 3E, and 3F, we compare the contingency tables of gene significance across methods. The baseline GLM method and RUVr are both capable of identifying genes in category 1 and category 3, a.k.a. the genes that are truly significantly associated with D and those that are not significantly associated with either D or Z. causarray is good at identifying the genes in category 1, but considers over half of the genes in category 3 as significantly associated with D. This compromise in causarray performance is likely because the method is designed for a single-cell resolution instead of a donor-level pseudobulk resolution.

For genes in **Category 2**, which are truly associated with the confounder Z, the baseline GLM approach identifies the majority of these genes as significant (xx%), whereas both causarray and RUVr classify many of them as insignificant (xx%). In contrast, `sensGAN` successfully recovers a substantial fraction of Category 2 genes (xx%) and appropriately nullifies those whose effects diminish as the confounder becomes more powerful. At the same time, `sensGAN` maintains high accuracy in identifying the correct signal patterns for **Category 1** and **Category 3** genes. The eight genes that `sensGAN` nullifies are more directly associated with Z rather than D, which is consistent with the predictive-gain constraint that sets the normalized predictive gain of predicting Y to 0.5.

Based on the full sensGAN output, a list of potential confounders constrained by predictive gains, we further present the sensitivity diagnostic contour plot (Fig 3H). The figure tells us the most inflated p-values (y-axis) of DEGs given certain predictive gain constraints. Here, we fix the predictive gains for outcome (Y, legend) and allow those for the case-control variable (D, x-axis) to vary. Every gene has a contour curve that depicts its sensitivity to the estimated unmeasured confounder. Smaller predictive gains (for the case-control variable) that allow the gene to cross the significance threshold indicate that the gene is more sensitive to the unmeasured confounder and that the conclusion of differential expression is less robust.

As mentioned previously, we simulated 3 categories of genes with different significance profiles. We summarized their significances associated with the case-control variable and the confounder in Table/Figure x.x. xxx genes in category 1 are DEGs with respect to the case-control variable, while xxx genes in category 2 are DEGs with respect to the confounder. Without adjusting for the unmeasured confounders, xxx genes



Estimated/ Simulated	Category1	Category3	Category1	Category3	Category1	Category3	Category1	Category2	Category3
Category1	2.36	12.64	1.82	13.18	3.96	11.04	12.04	0.76	2.20
Category2	3.36	10.64	7.12	6.88	7.08	6.92	2.40	8.14	3.46
Category3	64.94	4.06	37.30	31.70	64.58	4.42	2.20	3.90	62.90

Fig. 3. **asdf.**

are DEGs w.r.t the case-control variable, when they are truly associated with the unmeasured confounders. With sensGAN, we were able to recover the true significance profile by estimating the simulated unmeasured confounder (Fig. x.x). In addition, sensGAN was capable of identifying a series of potential confounding structures with the most inflated significance constrained by predictive gains. The sensitivity contours of genes in categories 1 and 3 do not experience much p-value inflation across increasing predictive gains. The contours of genes in category 2, however, exhibit decreasing trends of the negative *log* of the most inflated p-values as the predictive gains increase. It showed that more powerful confounders generally correspond to more inflation in p-values, nullifying more DEGs and revealing information about true significance profiles effectively. The simulation dataset results demonstrated the method's capability of identifying the most powerful confounding structures and potential/plausible confounding structures with the worst-case p-values constrained by predictive gains.

3.2 Analysis of Lupus dataset as proof-of-principle

(from GCATE, need paraphrase) Systemic lupus erythematosus (SLE) is an autoimmune disease predominantly affecting women and individuals of Asian, African, and Hispanic descent. [19] developed multiplexed single-cell RNA sequencing (mux-seq) to capture the complexity of immune cell populations and systematically profile the composition and transcriptional states of immune cells in a large multiethnic cohort. The dataset contains 1.2 million peripheral blood mononuclear cells from 8 major cell types and 261 individuals,

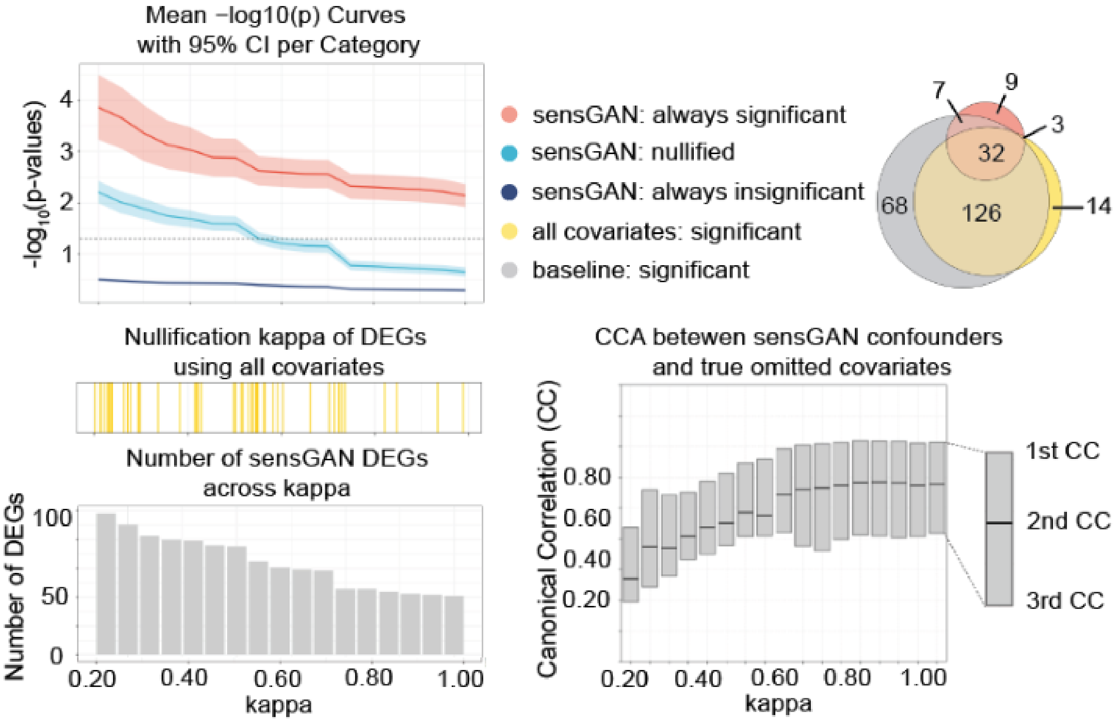


Fig. 4. asdf.

including 162 SLE cases and 99 healthy controls of either Asian or European ancestry. The cell-type-specific DE analysis aims to provide insights into the diagnosis and treatment of SLE.

To remove the genes with small variations, we use the Python package scanpy [?] to pre-process the single-cell data and select the top 2,000 highly variable genes (HVGs) within each cell type (look like it from the lupus ipynb you shared). With a focus on the XX cell type, we aggregated single-cell expression profiles by summing counts across cells from the same subject, yielding a gene-level pseudo-bulk count matrix. We then computed the total library size per subject and normalized counts to counts per ten thousand (CP10K) to mimic Seurat's NormalizeData [?] procedure. Normalized counts were logarithmically transformed and scaled to stabilize the variance. Finally, we removed genes with over 90% zero counts across subjects, retaining those expressed in at least 10% of samples for downstream analysis. In total, the dataset included 256 subjects (158 cases and 98 controls). For each subject, the case–control variable represented systemic lupus erythematosus (SLE) status, the measured covariate was sex, and the final number of retained genes was 815.

3.3 Analysis of microglia studying Alzheimer's disease, adjusting for unmeasured co-pathology

Alzheimer's disease (AD) is the leading cause of dementia in older adults and is characterized by progressive accumulation of amyloid and tau pathologies across cortical regions. The Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) consortium was established to build a high-resolution, multimodal cell atlas of AD and related dementias.

With a focus on the microglia in the middle temporal gyrus (MTG), we aggregated single-cell expression profiles by summing counts across cells from the same subject, yielding a gene-level pseudo-bulk count matrix. We performed DE analysis adjusting for measured covariates, including sex, age, APOE4 status, ethnicity, and PMI and kept differentially expressed genes (p-value < 0.2). We then computed the total library size per subject and normalized counts to counts per ten thousand (CP10K) to mimic Seurat's NormalizeData [?] procedure. Normalized counts were logarithmically transformed and scaled to stabilize the variance. In total, the dataset included 80 subjects (59 cases and 21 controls), each with 448 genes. For each subject,

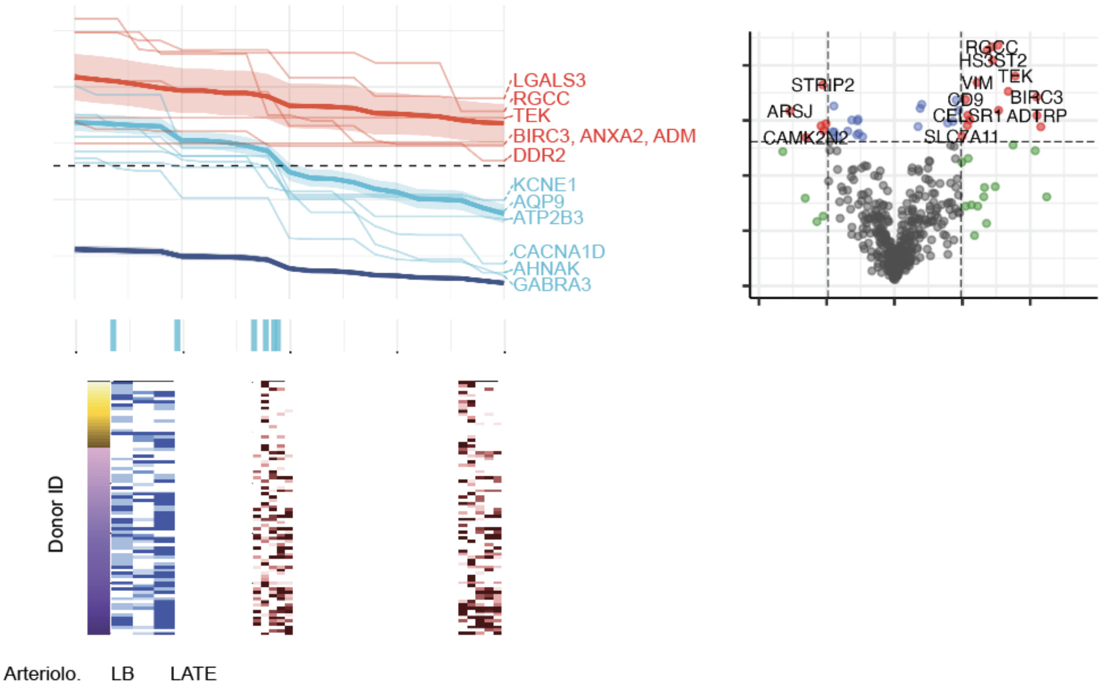


Fig. 5. asdf.

the case–control variable represented ADNC status, the measured covariates are sex, age, APOE4 status, ethnicity, and PMI.

3.4 Discussion

References

1. Braak, H., and Braak, E. Neuropathological stageing of alzheimer-related changes. *Acta neuropathologica 82*, 4 (1991), 239–259.
2. Cinelli, C., and Hazlett, C. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology 82*, 1 (2020), 39–67.
3. Di Mari, R., Ingrassia, S., and Punzo, A. Local and overall deviance r-squared measures for mixtures of generalized linear models. *Journal of classification 40*, 2 (2023), 233–266.
4. Drange, O. K., Smeland, O. B., Shadrin, A. A., Finseth, P. I., Witoelar, A., Frei, O., Group, P. G. C. B. D. W., Wang, Y., Hassani, S., Djurovic, S., et al. Genetic overlap between Alzheimer's disease and bipolar disorder implicates the MARK2 and VAC14 genes. *Frontiers in Neuroscience 13* (2019), 220.
5. Du, J.-H., Shen, M., Mathys, H., and Roeder, K. Causal differential expression analysis under unmeasured confounders with causarray. *bioRxiv* (2025), 2025–01.
6. Du, J.-H., Wasserman, L., and Roeder, K. Simultaneous inference for generalized linear models with unmeasured confounders. *Journal of the American Statistical Association*, just-accepted (2025), 1–24.
7. Filippova, G. N., Casad, M., Groneck, C., Hui, K., Mishra, S., MacDonald, J. W., Bammler, T., Van Dyke, D. L., Skakkebaek, A., Gravholt, C. H., et al. Modeling sex differences in Alzheimer's disease using isogenic hiPSC lines with different sex chromosome complements and APOE alleles. *Alzheimer's & Dementia 20*, Suppl 1 (2025), e093548.
8. Gabitto, M. I., Travaglini, K. J., Rachleff, V. M., Kaplan, E. S., Long, B., Ariza, J., Ding, Y., Mahoney, J. T., Dee, N., Goldy, J., et al. Integrated multimodal cell atlas of Alzheimer's disease. *Nature Neuroscience 27*, 12 (2024), 2366–2383.
9. Jack Jr, C. R., Andrews, J. S., Beach, T. G., Buracchio, T., Dunn, B., Graf, A., Hansson, O., Ho, C., Jagust, W., McDade, E., et al. Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. *Alzheimer's & Dementia 20*, 8 (2024), 5143–5169.

10. Leek, J. T., and Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics 3*, 9 (2007), e161.
11. Leek, J. T., and Storey, J. D. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences 105*, 48 (2008), 18718–18723.
12. McAleese, K. E., Colloby, S. J., Thomas, A. J., Al-Sarraj, S., Ansorge, O., Neal, J., Roncaroli, F., Love, S., Francis, P. T., and Attems, J. Concomitant neurodegenerative pathologies contribute to the transition from mild cognitive impairment to dementia. *Alzheimer's & Dementia 17*, 7 (2021), 1121–1133.
13. Mirra, S. S., Heyman, A., McKeel, D., Sumi, S., Crain, B. J., Brownlee, L., Vogel, F., Hughes, J., Belle, G. v., Berg, L., 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*, 4 (1991), 479–479.
14. Mittlböck, T. W. M., Haidinger, G., Zidek, T., and Schober, E. Partial r2-values based on deviance residuals in poisson regression models. *N MEDZN UND BOlOGE*, 341.
15. Montine, T. J., Phelps, C. H., Beach, T. G., Bigio, E. H., Cairns, N. J., Dickson, D. W., Duyckaerts, C., Frosch, M. P., Masliah, E., Mirra, S. S., et al. National institute on aging–alzheimer's association guidelines for the neuropathologic assessment of alzheimer's disease: a practical approach. *Acta neuropathologica 123*, 1 (2012), 1–11.
16. Nelson, P. T., Dickson, D. W., Trojanowski, J. Q., Jack, C. R., Boyle, P. A., Arfanakis, K., Rademakers, R., Alafuzoff, I., Attems, J., Brayne, C., et al. Limbic-predominant age-related tdp-43 encephalopathy (late): consensus working group report. *Brain 142*, 6 (2019), 1503–1527.
17. Nichols, E., Merrick, R., Hay, S. I., Himali, D., Himali, J. J., Hunter, S., Keage, H. A. D., Latimer, C. S., Scott, M. R., Steinmetz, J. D., Walker, J. M., Wharton, S. B., Wiedner, C. D., Crane, P. K., Keene, C. D., Launer, L. J., Matthews, F. E., Schneider, J., Seshadri, S., White, L., Brayne, C., and Vos, T. The prevalence, correlation, and co-occurrence of neuropathology in old age: Harmonisation of 12 measures across six community-based autopsy studies of dementia. *The Lancet Healthy Longevity 4*, 3 (2023), e115–e125.
18. Ovsepian, S. V., and O'Leary, V. B. Can arginase inhibitors be the answer to therapeutic challenges in Alzheimer's disease? *Neurotherapeutics 15*, 4 (2018), 1032–1035.
19. Perez, R. K., Gordon, M. G., Subramaniam, M., Kim, M. C., Hartoularos, G. C., Targ, S., Sun, Y., Ogorodnikov, A., Bueno, R., Lu, A., et al. Single-cell rna-seq reveals cell type–specific molecular and genetic associations to lupus. *Science 376*, 6589 (2022), eabf1970.
20. Robinson, J. L., Xie, S. X., Baer, D. R., Suh, E., Van Deerlin, V. M., Loh, N. J., Irwin, D. J., McMillan, C. T., Wolk, D. A., Chen-Plotkin, A., et al. Pathological combinations in neurodegenerative disease are heterogeneous and disease-associated. *Brain 146*, 6 (2023), 2557–2569.
21. Santiago, J. A., and Potashkin, J. A. The impact of disease comorbidities in Alzheimer's disease. *Frontiers in Aging Neuroscience 13* (2021), 631770.
22. Veitch, V., and Zaveri, A. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *Advances in neural information processing systems 33* (2020), 10999–11009.
23. Wang, J., Zhao, Q., Hastie, T., and Owen, A. B. Confounder adjustment in multiple hypothesis testing. *Annals of statistics 45*, 5 (2017), 1863.
24. Whalley, H. C., Papmeyer, M., Romaniuk, L., Johnstone, E. C., Hall, J., Lawrie, S. M., Sussmann, J. E., and McIntosh, A. M. Effect of variation in diacylglycerol kinase eta (DGKH) gene on brain function in a cohort at familial risk of bipolar disorder. *Neuropsychopharmacology 37*, 4 (2012), 919–928.
25. Yazar, S., Alquicira-Hernandez, J., Wing, K., Senabouth, A., Gordon, M. G., Andersen, S., Lu, Q., Rowson, A., Taylor, T. R., Clarke, L., et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science 376*, 6589 (2022), eabf3041.

A Details about loss terms

The Likelihood loss term $\mathcal{L}_L$ consists of (1) the BCE loss with logits $\mathcal{L}_{\mathrm{BCE}}$, a combination of sigmoid and BCE in one step to improve numerical stability compared with BCE loss; and (2) the negative log likelihood loss of the NB distribution $\mathcal{L}_{\mathrm{NLL}}$.

$$\mathcal{L}_L = \lambda_{\mathrm{BCE}}\mathcal{L}_{\mathrm{BCE}} + \lambda_{\mathrm{NB}}\mathcal{L}_{\mathrm{NB}},$$

where

$$\mathcal{L}_{\mathrm{BCE}} = \frac{1}{n}\sum_{i=1}^{n} \ell_i^{\mathrm{BCE}}, \quad \mathcal{L}_{\mathrm{NB}} = \frac{1}{n}\sum_{i=1}^{n} \ell_i^{\mathrm{NB}}.$$

The per-sample loss terms are given by

$$\ell_i^{\mathrm{BCE}} = \max(\widehat{d}_i, 0) - \widehat{d}_i d_i + \log\left(1 + e^{-|\widehat{d}_i|}\right),$$

and

$$\ell_i^{\mathrm{NB}} = -\Big[\log \mathit{\Gamma}(y_i + r) - \log \mathit{\Gamma}(r) - \log \mathit{\Gamma}(y_i + 1) + y_i \log \frac{\mu_i}{\mu_i + r} + r \log \frac{r}{\mu_i + r}\Big].$$

The Z regularization term $\mathcal{L}_Z$ consists of (1) a prior-matching loss, (2) an entropy regularizer, and (3) a correlation penalty.

$$\mathcal{L}_{\mathrm{reg}} = \lambda_{\mathrm{prior}}\mathcal{L}_{\mathrm{prior}} + \lambda_{\mathrm{ent}}\mathcal{H}(Z) + \lambda_{\mathrm{corr}}\mathcal{L}_{\mathrm{corr}}.$$

$$\mathcal{L}_{\mathrm{prior}} = -\frac{1}{n}\sum_{i=1}^{n}\Big[q \log p_i + (1-q)\log\big(1 - p_i\big)\Big],$$

where $p_i = P(Z_i = 1)$ and $q = q_{\mathrm{target}}$.

$$\mathcal{H}(Z) = -\frac{1}{n}\sum_{i=1}^{n}\Big[p_i \log(p_i + \varepsilon) + (1 - p_i)\log(1 - p_i + \varepsilon)\Big],$$

$$\rho_{Z,D} = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(d_i - \bar{d})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - \bar{d})^2 + \varepsilon}}, \quad \mathcal{L}_{\mathrm{corr}} = \rho_{Z,D}^2.$$

The plausibility constraint term $\mathcal{L}_P$ consists of (1) DBPR constraints and (2) p-value penalties.

$$\mathcal{L}_P = \mathcal{L}_{\mathrm{pval}} + \mathcal{L}_{\mathrm{DBPR}}.$$

$$\begin{aligned} \mathcal{L}_{\mathrm{pval}} &: \text{penalizes excessive gene-level significance (soft threshold on } z\text{-scores)}, \\ \mathcal{L}_{\mathrm{DBPR}} &: \text{enforces predictive gain constraints.} \end{aligned}$$

$$\mathcal{L}_{\mathrm{pval}} = \lambda_{\mathrm{pval}} \cdot \frac{1}{p}\sum_{j=1}^{p}\sigma_\lambda\left(\left|\frac{\widehat{B}_j}{\mathrm{SE}_j}\right| - z_{1-\alpha}\right),$$

where $\sigma_\lambda(x) = \frac{1}{1+e^{-\lambda x}}$ is a smooth sigmoid centered at the significance threshold $z_{1-\alpha}$, which behaves as the following:

$$\text{Below threshold: } \approx 0, \quad \text{at threshold: } \approx 0.5, \quad \text{above threshold: } \approx 1.$$

$$\mathcal{L}_{\mathrm{DBPR}} = \lambda_Y\Big[\big((\bar{\eta} - \eta^*)\, s_Y\big)^2\Big] + \lambda_D\Big[\big((\kappa - \kappa^*)\, s_D\big)^2\Big],$$