High-Dimensional Sensitivity Analysis for Genomic Studies: An Adversarial Framework for Learning Worst-Case Latent Confounders

Anonymous Authors[1]

Abstract

High-dimensional genomics studies are frequently confounded by unmeasured biological processes that obscure disease-specific signals. While existing workflows can estimate these latent confounders, they fail to quantify how robust a discovery is to varying levels of hypothetical confounding. We introduce sensGAN, a deep-learning adversarial framework that systematically explores the confounding spectrum by learning "worst-case" latent variables that nullify the most gene associations under novel predictive-gain constraints. By identifying the minimum confounding strength required to explain away an observed effect, our method shifts the paradigm toward a formal, quantitative sensitivity analysis. In diverse simulations, sensGAN accurately recovers latent structures and outperforms existing methods in identifying confounder-sensitive genes. Applied to human Alzheimer's disease microglia, our framework prioritizes robust disease pathways while successfully isolating signals driven by unmeasured co-occurring neurodegenerative pathologies.

1. Introduction

A fundamental challenge in modern genomics is the presence of unmeasured confounding variables that can obscure true biological signals. In high-dimensional single-cell studies, transcriptional changes attributed to a primary disease state may instead reflect unmeasured comorbid processes, environmental exposures, or systemic pathologies that are incompletely recorded. While conventional differential expression analyses prioritize genes based on statistical significance (Squair et al., 2021), they often fail to quantify the sensitivity of these findings to latent confounders: the more sensitive the findings are to latent confounders, the less robust the discoveries are. Computational frameworks, such as surrogate variable analysis (Leek & Storey, 2007; 2008), provide a single point estimate of latent factors to adjust for residual variation but do not explore the spectrum of hypothetical confounding. This leaves researchers without a metric to determine how "strong" a latent factor would need to be to nullify a discovery.

We argue that high-dimensional genomics requires a shift toward formal sensitivity analysis, mirroring classic statistical approaches like Cornfield's analysis of smoking and lung cancer (Cornfield et al., 1959). Conceptually, we ask: what level of confounding strength, relative to observed covariates, is required to explain away a gene's association with a treatment or disease label? This is a novel machine learning problem involving the optimization of "worst-case" latent variables within high-dimensional outcome spaces. By treating confounding as a spectrum rather than a fixed correction, we can distinguish between transcripts that are robust to latent biological processes and those whose significance is likely an artifact of omitted variable bias.

To address this, we introduce **sensGAN** (Sensitivity Generative Adversarial Network), a generative adversarial network for sensitivity analysis that leverages deep learning. SensGAN learns worst-case latent confounders (also called "unmeasured confounders" in other literature) by constraining their predictive gains on both the treatment and outcome, identifying the weakest confounding strength required to nullify each gene's association with the disease.

1.1. Scientific relevance

The necessity of sensitivity analysis is most evident in the study of complex disorders like Alzheimer's disease (AD). A major computational obstacle in uncovering "direct" disease signals is that donors frequently harbor co-occurring neurodegenerative diseases (co-pathologies). For instance, in late-onset AD, as few as 31% of cases are described by purely AD-specific signatures, with the remainder displaying complex comorbidity patterns that potentially confound existing single-cell analyses (Robinson et al., 2023). Preliminary analyses of large-scale cohorts demonstrate that as primary disease severity increases, donors exhibit an in-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

creasing burden of these co-pathologies, which significantly impacts clinical outcomes and donor lifespan (Spina et al., 2021), see Appendix S11.

The challenge for computational biologists is that the clinical definitions for these co-pathologies are not static; they undergo major updates roughly every decade as scientific understanding evolves (Mirra et al., 1991; Montine et al., 2012; Nelson et al., 2019). Given this context, sensGAN provides a "future-looking" framework for differential expression. Rather than relying solely on currently formalized staging, our sensitivity analysis offers two critical advantages:

- **(Goal #1) Make robust discoveries even when important confounders are unmeasured or unknown**: SensGAN enables a sensitivity analysis with respect to latent confounders by explicitly constructing plausible worst-case latent confounders under controlled confounding strength to identify the one that nullifies the largest number of differentially expressed genes. Genes that remain significant under these worst-case scenarios are therefore robust to a broad class of latent confounding effects.
- **(Goal #2) Use the learned confounders to suggest what hidden biology might exist**: SensGAN learns explicit latent confounder representations that may reflect underlying biological processes not yet measured or formally defined. Although these latent variables are not assumed to correspond to any known pathology, their values can suggest the presence and functional impact of latent biological drivers, providing hypotheses for future experimental or clinical investigation.

1.2. Existing computational methods and their limitations

We model log-normalized, pseudo-bulked gene expression vector $Y \in \mathbb{N}^{n\times p}$ (i.e., the "outcome" across n donors and p genes), treatment status $D \in \{0,1\}^{n\times 1}$ (i.e., the case-control status), and the d measured covariates $X \in \mathbb{R}^{n\times d}$. Following standard Generalized Linear Model (GLM) frameworks commonly used to model gene expression vector (Sarkar & Stephens, 2021), we assume:

$$\begin{aligned} Y &\sim \mathrm{NB}(\mu, \alpha), \quad \mu = DT_Y + XB_Y + Z\Gamma_Y \\ D &\sim \text{Bernoulli}(\pi), \quad \pi = XB_D + Z\Gamma_D \end{aligned} \tag{1}$$

Our primary objectives are to estimate the direct effect $T_Y \in \mathbb{R}^{1\times p}$ of D on Y (Goal #1) and the k-dimensional latent factor matrix $Z \in \mathbb{R}^{n\times k}$ (Goal #2). Here, $\alpha \in \mathbb{R}^p_+$ denotes gene-specific overdispersions, and $\{B, \Gamma\}$ represent effect matrices for observed and latent variables, respectively, for either the outcome or treatment, depending on the subscript. In practice, Y is often derived from single-cell RNA-seq data, where we aggregate the expression profile across all cells of a particular cell type of interest from each donor, and then log-normalize the expression profile across p highly variable genes of interest.

However, since Z is unobserved, typical DE methods are forced to model the data Y using an "unadjusted" model,

$$\begin{aligned} \widehat{Y}_{\text{un}} &\sim \mathrm{NB}(\widehat{\mu}_{\text{un}}, \widehat{\alpha}_{\text{un}}), \quad \widehat{\mu}_{\text{un}} = D\widehat{T}_{Y,\text{un}} + X\widehat{B}_{Y,\text{un}} \\ \widehat{D}_{\text{un}} &\sim \text{Bernoulli}(\widehat{\pi}_{\text{un}}), \quad \widehat{\pi}_{\text{un}} = X\widehat{B}_{D,\text{un}} \end{aligned} \tag{2}$$

We highlight existing computational frameworks below for handling latent confounders Z in this setting, which motivate our proposed method.

Existing frameworks for latent confounding generally fall into two categories. Omitted Variable Bias (OVB) methods quantify the strength required for a latent factor to nullify a treatment-outcome relation (Cinelli & Hazlett, 2020; Veitch & Zaveri, 2020). However, these are largely restricted to linear models and do not scale to the non-linear, high-dimensional outcome spaces (e.g., thousands of genes modeled via Negative Binomial (NB) distributions) typical of single-cell genomics. Conversely, surrogate-variable approaches, such as causarray (Du et al., 2025) and RUVr (Risso et al., 2014), estimate latent factors to adjust for residual variation. Yet, these provide only point estimates rather than exploring the spectrum of hypothetical confounding. Furthermore, they often assume that latent factors are independent of the treatment, making them poorly suited to capture neurodegenerative diseases that co-occur with primary disease states like AD. See Appendix S2 for additional discussion.

2. Overview of sensGAN

Our method, sensGAN, estimates the worst-case latent confounder under constrained confounding strengths using an adversarial framework and provides a sensitivity contour diagnostic plot to suggest more reliable gene targets based on measured variables. In the following, we motivate the critical components of our framework one at a time: (1) quantifying the two "knobs" that define confounding strength, (2) defining the generative adversarial network (GAN) architecture that enables learning the confounder, and (3) training the GAN architecture.

2.1. Predictive gain: defining the confounding strength

We start with our formalization of quantifying the strength of a hypothetical latent confounder, which we denote as $\widetilde{Z}$. Intuitively, the strongest latent confounder is the one that improves predictive performance the most for either the genes (Y; "outcomes") or the case-control status (D; treatment). Given $\widetilde{Z}$, we can learn $\widetilde{\mu}$ and $\widetilde{\pi}$ following the generative model in Eq. (1). The predictive gain of $\widetilde{Z}$ is defined as the normalized deviance-based partial R^2 (DBPR) defined

below, which we denote as R^2.

For each prediction task (treatment D and outcomes Y), we compare three fitted models: (i) an unadjusted model that does not consider any latent confounders ("un", see (2), (ii) an adjusted model using the *strongest* unobserved confounder $\widehat{Z}$ (the k-dimensional confounder that maximizes the likelihood), and (iii) an adjusted model using a particular hypothetical confounder $\widetilde{Z}$. Let $\ell_D(\pi)$ and $\ell_Y(\mu, \alpha)$ denote the Bernoulli and negative-binomial (NB) log-likelihoods under (1), and let $\ell_{D,\text{sat}}$ and $\ell_{Y,\text{sat}}$ denote the corresponding *saturated* (perfect-fit) log-likelihoods, i.e., the maximum achievable log-likelihood when each observation has its own free parameter. The deviance of a fitted model is $\text{Dev} = 2(\ell_{\text{sat}} - \ell)$, with smaller deviance meaning a better fit. Dev_D is defined as the sum of Bernoulli deviances across all donors, and Dev_{Y_i} is defined as the sum of NB deviances across all genes. Details and derivations provided in Appendix S3. The deviance-based partial R^2 for including Z into the model is the fraction of unadjusted deviance removed:

$$R^2_D = \frac{\text{Dev}_{\widehat{D}_{\text{un}}} - \text{Dev}_D}{\text{Dev}_{\widehat{D}_{\text{un}}}}, \quad R^2_{Y_i} = \frac{\text{Dev}_{\widehat{Y}_{\text{un},i}} - \text{Dev}_{Y_i}}{\text{Dev}_{\widehat{Y}_{\text{un},i}}}.$$

Finally, we normalize by the maximal gain achieved by $\widehat{Z}$ to obtain bounded predictive-gain "knobs" $\kappa, \eta \in [0, 1]$, defined as

$$\kappa = \frac{R^2_{\widetilde{D}}}{R^2_{\widehat{D}}}, \quad \eta = \frac{\frac{1}{n}\sum_{i=1}^n R^2_{\widetilde{Y}_i}}{\frac{1}{n}\sum_{i=1}^n R^2_{\widehat{Y}_i}}, \tag{3}$$

which quantify how predictive $\widetilde{Z}$ is relative to the strongest confounder in our model class. These emulate similar properties and interpretations of partial R^2 for linear models as discussed in sensmarker (Cinelli & Hazlett, 2020). 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 $\widehat{Z}$. A normalized predictive gain of 1 means that $\widetilde{Z}$ achieves 100% of the maximal predictive gain. Importantly, it is also differentiable, allowing gradient calculation in a deep learning model, as we describe later in this section.

2.2. Architecture and loss

Now that we have a quantification of the impact of a latent confounder, we next need to describe sensGAN's deep-learning architecture, which allows us to learn the confounder Z that meets our specifications. Broadly speaking, sensGAN's architecture has two fundamental components: the predictor and the generator. Here, we first describe each component of the architecture as shown in Fig. 2.

Predictors. The predictors $f_D(\cdot)$ and $f_Y(\cdot)$ are 1-layer (i.e., shallow) fully connected networks that estimate model parameters by maximizing the likelihoods. To D_i, the treatment predictor $f_D(\{X_i, Z_i\})$ uses a binary cross-entropy (BCE) loss. To predict Y_i, the outcome predictor $f_Y(\{D_i, X_i, Z_i\})$ uses a NB log-likelihood loss. Both predictors are designed to emulate logistic regression and NB GLM regression, respectively. There are no activation functions involved, so the weight corresponding to variable D in $f_Y(\cdot)$ can be interpreted as the "coefficient" T_Y in our generative model; see Eq. (1). The overdispersion parameters for modeling Y are estimated from the unadjusted NB model fitting.

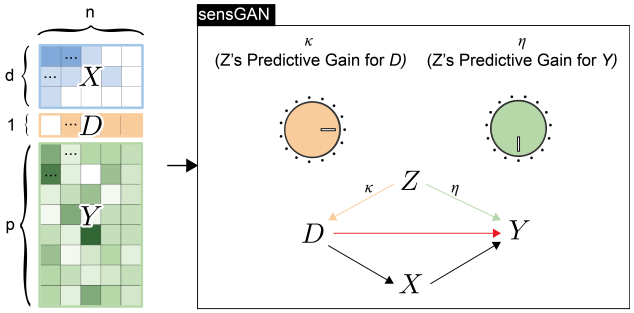

Figure 1. **sensGAN core idea.** As illustrated, unobserved confounders may simultaneously influence both the treatment D and the outcome Y, leading to spurious associations. sensGAN parameterizes this confounding strength using two normalized predictive-gain knobs and systematically explores the resulting confounding scenarios.

Generator. The generator $g(\cdot)$ deterministically maps donor-level information to k latent confounders $\widetilde{Z}_i \in [0, 1]^k$ for donor i, based on covariates X_i, treatment status D_i, and outcome (gene expression) vector Y_i. By default, the generator is a two-layer fully connected neural network with ReLU activation that produces $\widetilde{Z}_i$ via a sigmoid output layer.

2.3. Training

Equipped with the predicator and generator architectures, we are now ready to explain how we train sensGAN. Drawing inspiration from the OVB literature (Cinelli & Hazlett, 2020; Veitch & Zaveri, 2020), we strive to learn latent confounders $\widetilde{Z}$ that nullifies as many genes (i.e., eliminates the significance of the largest number of genes associated with D), such that $\widetilde{Z}$ has a particular pre-specified predictive gain $\{\kappa, \eta\}$. The overall procedure has two main steps:

- **Step 1**: Learn the unadjusted model and the most powerful confounder, both of which are needed to define the normalized DBPRs; see Eq. (3),
- **Step 2**: Train the predictors $f_D(\cdot)$ and $f_Y(\cdot)$ and generator $g(\cdot)$ in an adversarial manner.

We highlight each step below. The full pseudocode is described in Appendix S7.

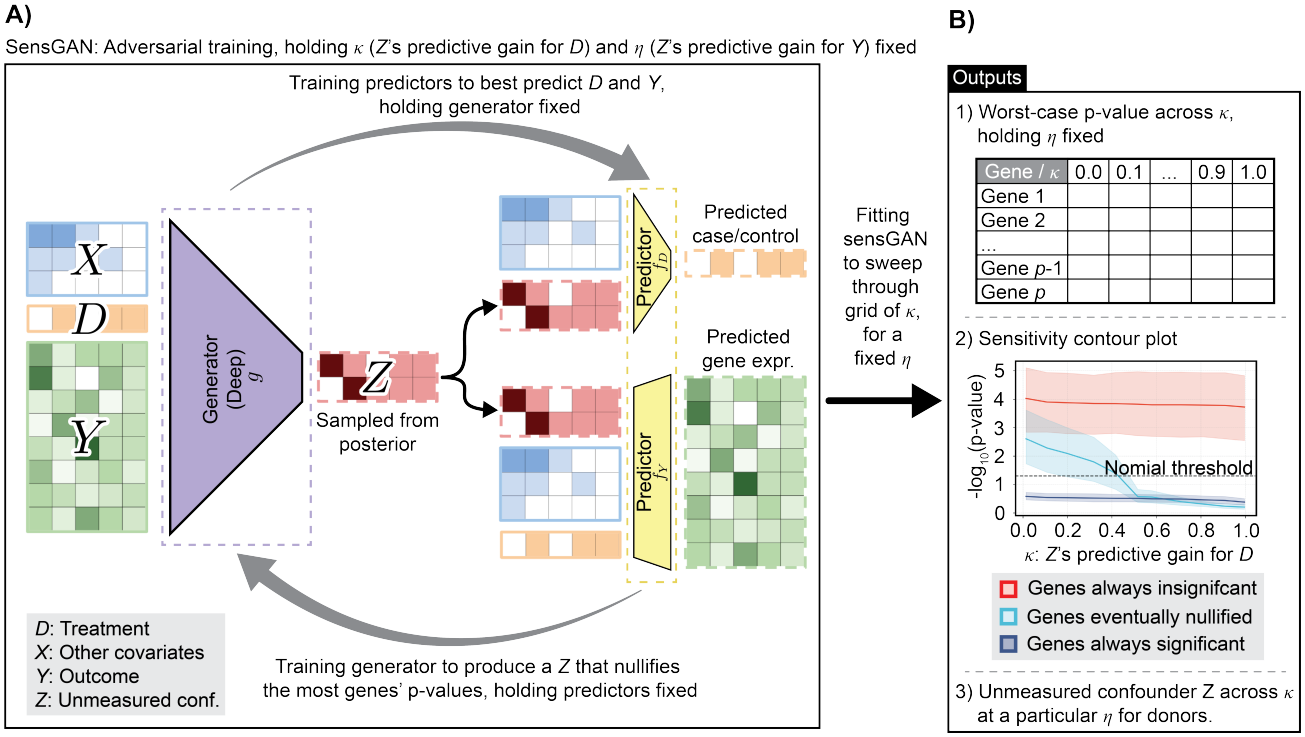

Figure 2. **sensGAN method overview. A)** The GAN system that learns the confounder Z with target predictive gains (κ, η), achieving predictive gains (κ', η'). **B)** Method outputs while sweeping through a grid of κ and holding η fixed.

Step 1: Unadjusted predictors and most powerful confounder. To operationalize our definition of the predictive gains in Eq. (3), we need to first compute the likelihood of the unadjusted model and the adjusted model using the most predictive confounder $\widehat{Z}$.

To learn the unadjusted models, we fit a logistic regression to predict D from X, and fit p NB GLMs to predict each gene in Y from X and D. From these fitted models, we can compute $\text{Dev}_{\widehat{D}_{\text{un}}}$ and $\text{Dev}_{\widehat{Y}_{\text{un}}}$, respectively. The gene-specific overdispersion parameters α estimated at this stage are held fixed and used throughout all subsequent sensGAN calculations.

We then train the predictors and the generator to minimize the joint negative log-likelihood, yielding the most powerful confounder $\widehat{Z}$ and the corresponding models $f_D(\cdot)$, $f_Y(\cdot)$, and $g(\cdot)$. Specifically, we define the BCE loss based on predicted logits as $\mathcal{L}_{\text{BCE}}$ and the NB loss based on the predicted mean as $\mathcal{L}_{\text{NB}}$. Predictors are trained respectively by minimizing $\mathcal{L}_{\text{BCE}}$ and $\mathcal{L}_{\text{NB}}$. Generators are trained by minimizing

$$\mathcal{L}_L = \lambda_{\text{BCE}} \cdot \mathcal{L}_{\text{BCE}} + \lambda_{\text{NL}} \cdot \mathcal{L}_{\text{NB}},$$

where $\lambda_{\text{BCE}}, \lambda_{\text{NL}} \geq 0$ are hyperparameters controlling the weight of their respective terms. From this fitted model, we can compute $\text{Dev}_{\widehat{D}}$ and $\text{Dev}_{\widehat{Y}}$, respectively. More details are discussed in Appendix S5.

Step 2: Adversarial training We now describe how we train predictors $f_D(\cdot)$ and $f_Y(\cdot)$ and generator $g(\cdot)$. After initializing the GAN architecture (see Appendix S6), we then update the predictors and generator alternatively and adversarially – the predictors maximize log-likelihoods while the generator controls for predictive gains. Specifically, in one iteration, the generator is frozen and the predictors $f_D(\cdot)$ and $f_Y(\cdot)$ are updated; in the other, the predictors are frozen and the generator $g(\cdot)$ is updated.

When updating the predictors, we minimize $\mathcal{L}_{\text{BCE}}$ to update the weights in $f_D(\cdot)$ and minimize $\mathcal{L}_{\text{NB}}$ to update the weights in $f_Y(\cdot)$. This is relatively straightforward, as both predictors are designed to emulate logistic and NB GLMs, respectively, where the optimization is convex.

Generator optimization is more involved and represents a key component of the sensGAN framework. The loss function when optimizing $g(\cdot)$ involves three main terms: the plausibility objective $\mathcal{L}_{\text{P}}$, the significance objective $\mathcal{L}_{\text{S}}$, and a regularizer $\mathcal{L}_{\text{reg}}$. We explain each component below.

- To motivate the plausibility objective, we note that the strongest confounder $\widehat{Z}$ explains away all unexplained residuals in D and Y, by construction. This is unlikely to reflect any practical latent confounder. We thus define the *plausibility objective* ($\mathcal{L}_{\text{P}}$): the predictive gains of plausible confounders $\widetilde{Z}$ with respect to D and Y are constrained by

prespecified targets, $(\kappa, \eta) \in [0, 1]$,

$$\mathcal{L}_P = \lambda_Y \cdot \left[\left((\eta' - \eta)\, s_Y\right)^2\right] + \lambda_D \cdot \left[\left((\kappa' - \kappa)\, s_D\right)^2\right],$$

where κ' and η' denote the achieved predictive gains with respect to D and Y, respectively; s_Y and s_D are scaling factors, and $\lambda_Y, \lambda_D \geq 0$ are hyperparameters.

- To motivate the significance objective, we formalize "worst-case confounder" given κ and η as the confounder $\widetilde{Z}$ that nullifies the largest number of significant genes. Thus, we define the *significance objective* ($\mathcal{L}_S$) that minimizes the average sigmoid-transformed excess of the test statistics over the significance threshold $\delta \in (0, 1)$ (i.e., typically $\delta = 0.05$),

$$\mathcal{L}_S = \lambda_{\text{pval}} \cdot \frac{1}{p} \sum_{j=1}^{p} \sigma_\lambda \left(\left| \frac{[\widetilde{T}_Y]_j}{\text{SE}([\widehat{T}_{Y,\text{un}}]_j)} \right| - z_{1-\delta} \right),$$

where $[\widehat{T}_{Y,\text{un}}]_j$ and $[\widetilde{T}_Y]_j$ denote the estimated coefficients of D for gene j in the unadjusted and adjusted NB models, see Eq. (2) and (1) respectively. $\text{SE}(\widehat{T}_{Y,\text{un},j})$ is the corresponding standard error from the unadjusted model. To compute the test statistic, we approximate $\text{SE}(\widehat{T}_{\text{un},j}) \approx \text{SE}(\widetilde{T}_j)$, assuming comparable uncertainty before and after adjustment. $z_{1-\delta}$ is the $(1 - \delta)$ quantile of the standard normal distribution. $\sigma_\lambda(\cdot)$ denotes a sigmoid function, which smoothly maps test statistics onto the unit interval. $\lambda_{\text{pval}} \geq 0$ is a hyperparameter controlling the overall strength of the significance penalty.

- The last term is the *regularizer* ($\mathcal{L}_{\text{reg}}$). This includes a correlation regularizer, a cosine similarity loss, and a variance-matching penalty, all of which we have empirically found help stabilize the learning of the latent confounder Z and prevent the model from collapsing to a constant value. See more details in Appendix S5.

Combining all the loss terms together, the objective function of the generator $g(\cdot)$ given pre-specified (and fixed) predictive gains κ^* and η^* is to minimize,

$$\mathcal{L}_S + \mathcal{L}_P + \mathcal{L}_{\text{reg}}.$$

To span all plausible confounders, we train the sensGAN model sequentially by sweeping over the predictive gains (κ, η) on the unit interval $[0, 1]$. At the end of the sequential training, the generator finds a series of worst-case confounders with varied predictive gains.

The latent confounders Z can be used in a standard NB GLM analysis where we regress $\{X, D, Z\}$ onto Y, where we record the significance of each gene's coefficient in T_Y, see Eq. (1). Gene-level p-values are computed using Wald statistics, where the test statistic for each gene is formed by dividing the adjusted coefficient by its standard error estimated from the unadjusted model ($[\widetilde{T}_Y]_j / \text{SE}([\widehat{T}_{Y,\text{un}}]_j)$). Mathematically, the p-value of each gene changes in the model that adjusts for our learned latent confounder, compared to the unadjusted model, see Eq. (2)). We are interested in the specific value of κ^* that "nullifies" the gene given a fixed η, if it is nullified in the adjusted model.

Additional details, such as the choice of hyperparameters, are deferred to Appendix S6. We also elaborate further on the modeling strategy, explaining why we chose to model scRNA-seq data as pseudobulked samples rather individual cells, as done in work like causarray (Du et al., 2025) and CoCoA-diff (Park & Kellis, 2021), in Appendix S4.

2.4. Downstream analysis: Calibration of predictive gains with measured covariates

From the sensitivity analysis, sensGAN produces a series of plausible worst-case confounders indexed by normalized predictive gains (κ, η) with respect to the exposure D and outcome Y. To interpret these latent confounders relative to observed covariates, we perform a calibration step that computes predictive gains for measured covariates, placing latent and measured effects on a common scale. If a worst-case confounder with strength x times that of the strongest measured covariate is required to nullify a gene's significance, large values of x indicate that such confounding is potentially implausible, supporting the robustness of the discovery. See more details in Appendix S6.

2.5. Downstream analysis: Diagnostic sensitivity contour plots

To connect sensGAN's core idea to gene-level robustness, we introduce a *sensitivity diagnostic* inspired by (Cinelli & Hazlett, 2020). Fixing the predictive-gain knob for the outcome (η) to reflect a conservative but bounded amount of confounding in the outcome, we vary the treatment-side knob (κ) and record the corresponding worst-case p-values. This assesses how predictive a latent confounder needs to be of D in order to nullify a gene's significance. Genes that lose significance at small κ are sensitive to weak confounding, whereas genes that remain significant until large κ exhibit greater robustness.

3. Experimental Setup

3.1. Analysis of the simulation dataset for method comparison

We evaluated sensGAN on simulated data generated from a model with a single unobserved binary confounder (n=100, p=100, d=4, k=1). Observed covariates X and the binary latent confounder Z jointly influenced a case–control variable D through a logistic regression model, while the outcome

matrix Y was generated from a NB generalized linear model depending on X, D, and Z. By varying the effects of D and Z on gene expression, we simulated three groups of genes: (i) genes truly associated with D but not Z, (ii) genes associated with Z but not D, and (iii) genes associated with neither. More details and parameter settings are provided in Appendix S8.

To evaluate different methods, we first compared the correlations between the learned confounder and the true confounder (Fig. 3). The most powerful confounder identified by sensGAN shows a correlation comparable to that of causarray (Du et al., 2025) and RUVr (Risso et al., 2014), two surrogate-variable methods that we benchmark against. As the predictive-gain knob, κ, is turned from 1 to 0, the correlation between the learned and true confounders decreases smoothly from approximately 0.75 to 0.4. 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 predictive gains.

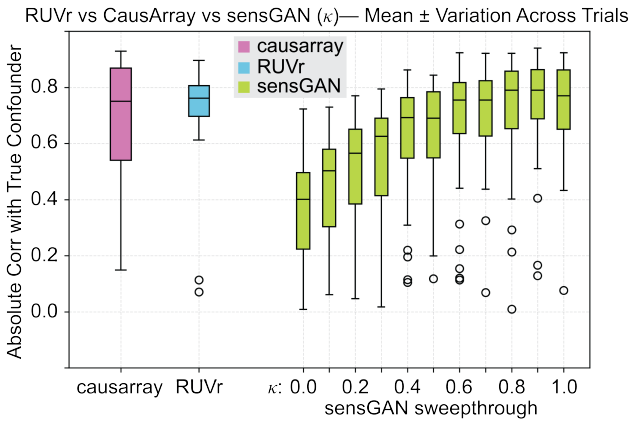

Figure 3. **Correlation comparisons among methods.** Correlations between the true confounder and the confounders estimated by causarray, RUVr, and sensGAN.

Next, we compare contingency tables of gene significance across methods (Table 1). All methods correctly identify always-significant genes, while causarray behaves conservatively for always-insignificant genes, likely due to its design for single-cell rather than donor-level pseudobulk data. Differences are most pronounced for nullified genes: the unadjusted GLM naively labels all such genes as significant (Fig. 3B), and causarray and RUVr split them between significant and insignificant categories. In contrast, sensGAN learns a family of plausible confounders with varying predictive gains and identifies a substantial fraction of nullified genes under worst-case confounding (8 out of 15; Table 1), while maintaining high accuracy for always-significant and always-insignificant genes (Fig. 4).

We further present the sensitivity diagnostic contour plot

Table 1. **Comparison of mean gene counts across simulated significance categories.** Rows: true gene categories, columns: inferred categories assigned by each method; values report mean counts averaged over 10 trials.

Method	Category	# Sig.	# Null.	# Insig.
Unadjusted	Significant	12.64	–	2.36
	Nullified	10.64	–	3.36
	Insignificant	4.06	–	64.94
causarray	Significant	13.18	–	1.82
	Nullified	6.88	–	7.12
	Insignificant	31.70	–	37.30
RUVr	Significant	11.04	–	3.96
	Nullified	6.92	–	7.08
	Insignificant	4.42	–	64.58
sensGAN	Significant	12.04	0.76	2.20
	Nullified	2.40	8.14	3.46
	Insignificant	2.20	3.90	62.90

based on sensGAN outputs (Fig. 3). We focus on nominal p-values rather than multiple-testing-adjusted p-values, as these allow us to better investigate how p-values change as the level of confounding increases. The y-axis is the logarithm of the worst-case p-values, the black dashed line is the significance threshold of 0.05, and the x-axis is the predictive knob (κ) for the treatment. Every gene has a contour curve that depicts its sensitivity to the estimated latent confounder, exemplified by the thin curves. The thick curves are the mean contour curves, and the bands are the 95% confidence intervals, colored by the simulated gene categories. If a gene crosses the significance threshold with a smaller κ, the gene is considered to be nullified by a weaker confounder. Compared to genes nullified by stronger confounders, this gene is less directly associated with the treatment. Therefore, the differential expression conclusion is less robust.

3.2. Validation of contour curve accuracy

Having shown that sensGAN recovers latent confounders correlated with those identified by existing methods, we next evaluate its ability to accurately characterize sensitivity contour curves, which prior approaches do not provide. Our primary focus is to investigate whether sensGAN can faithfully track how gene-level p-values evolve with increasing confounding strength. To this end, we simulate data as before ($n = 100$, $p = 104$, $d = 4$, $k = 1$), but consider a continuous confounder Z to enable finer control over predictive gains, and restrict attention to genes that are associated with Z but not with D (Category (ii)). We first construct reference sensitivity contour curves using a crawler-based search that manually explores the predictive-gain space to identify the "true" worst-case confounder for a particular setting of (κ, η) (see Appendix S8). We then

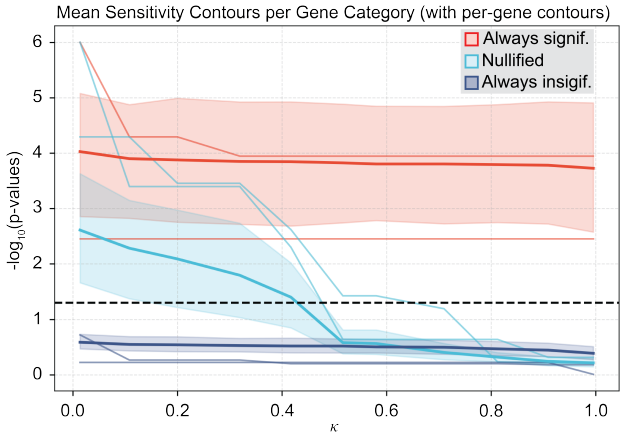

Figure 4. **Simulation demonstrates sensGAN's accuracy. Sensitivity diagnostic contour plot** With fixed $\eta = 0.5$, the thick contour curve represents the mean across genes, the thin curve shows an individual gene for illustration, and the shaded band denotes the 95% confidence interval.

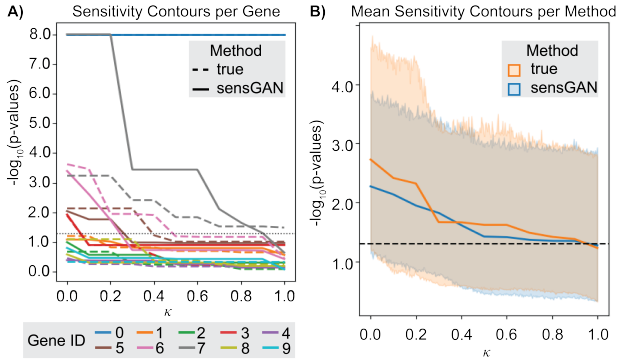

Figure 5. **Sensitivity contour plot comparison between the reference and the sensGAN output. A)** Per-gene sensitivity contour plots showing $-\log_{10}(p)$ as a function of the predictive-gain parameter κ for individual genes. Dashed curves denote reference contours obtained via the crawler, while solid curves denote contours recovered by sensGAN. **B)** Mean sensitivity contour curves across genes, with shaded regions indicating 95% confidence intervals.

apply sensGAN to the same simulated data and compare the resulting sensitivity contours, assessing whether sensGAN accurately reproduces the true nullification behavior across predictive-gain levels.

We first evaluate sensGAN at the gene level by comparing per-gene sensitivity contour curves against the reference contours (Fig. 4), with a primary focus on the difference in κ^* at the nullification point. Quantitatively, the difference is modest, with a mean difference of 0.05 and a 95% confidence interval of $(0.00, 0.22)$, indicating close alignment in the predictive-gain level at which genes lose significance.

We next compare the mean sensitivity contour curves across genes (Fig. 5B). Overall, the mean curves produced by sensGAN closely track the reference contours, with similar shapes and comparable crossing points near $\kappa \approx 0.95$. The 95% confidence intervals largely overlap across the predictive-gain range, suggesting that sensGAN captures the dominant trend in how significance decays with increasing confounding strength. See Appendix S9 for more results.

3.3. Analysis of microglia studying Alzheimer's disease, adjusting for unmeasured co-occurring neurodegenerative disease

To demonstrate how sensGAN can clarify the biological signals, we showcase our analysis of systemic lupus erythematosus (SLE) and Alzheimer's disease (AD). We discuss the AD case-study here in detail, while deferring the SLE case-study to Appendix S10. For our AD case-study, we leverage the Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) consortium (Gabitto et al., 2024) and specifically focus on the microglia in the prefrontal cortex (PFC) due to its critical role in clearing AD pathology and its implication from established GWAS studies (Karch & Goate, 2015; Scheltens et al., 2016). We construct a pseudobulk gene expression matrix per donor (i.e., Y) and perform our sensGAN analysis to determine which genes are impacted based on whether the donor has AD pathology after post-mortem dissection of their brain tissue (i.e., D), adjusting for sex, age, APOE4 status, ethnicity, and post-mortem interval (PMI) (i.e., X). Since we are only interested in whether or not significant genes remain significant during our sensitivity analysis, we focus on a key set of genes that had a p-value < 0.2 when doing a differential expression test via DESeq2 (Love et al., 2014), resulting in $p = 448$ genes in our analysis. Our goal is to determine which nominally significant genes are nullified by the latent confounders identified by sensGAN (Goal #1) and what these latent confounders represent biologically (Goal #2).

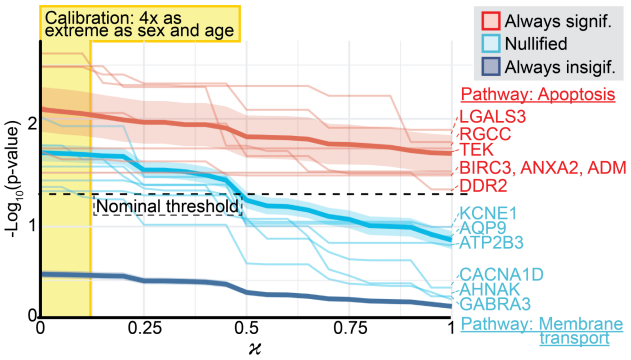

Figure 6. **sensGAN prioritizes DEGs specific to microglia in AD.** Diagnostic contour, showing genes that are always significant (red), nullified (light blue), and always insignificant (dark blue) across different values of κ. The calibration with respect to age and sex is shown (yellow).

Our sensGAN analysis, sweeping across κ from 0 to 1, disentangles potential microglial pathways that may be confounded by co-pathologies from those that are specific to AD. Our diagnostic contour plots reveal that genes that are always nominally significant are enriched for an apoptosis pathway (red; Fig. 5) (Dou et al., 2024). For instance, the *LGALS3* gene is strongly upregulated in plaque-associated microglia in AD models and human tissue (Tan et al., 2021), and has been implicated to be AD-specific through knockouts in AD mouse models, which demonstrated reduced plaque burden and improved cognitive behaviors (Siew et al., 2024). In stark contrast, genes that are eventually nullified by sensGAN's learned latent confounders Z are enriched for transmembrane transport (light blue; Figure 5). For instance, AHNAK regulates voltage-gated calcium channels (Matza et al., 2008). Although it is nominally significant without confounder adjustment, this gene is quickly nullified by sensGAN. This is plausible, since its association with AD is not solely through microglia but also through its interaction with neurons (Wang et al., 2025). Furthermore, it has also been suggested to be involved in Lewy bodies (Santpere et al., 2018) and Frontal Temporal Dementia (Lorenzini et al., 2023), two other co-occurring pathologies typically associated with AD. Additionally, AQP9, a gene that regulates membrane channels to conduct water, is also nullified, which is canonically associated with an AD-relevant gene for astrocytes, not microglia (Liu et al., 2018). Together, these results demonstrate that sensGAN can meaningfully "purify" DEG results, retaining only microglia-specific DEGs robust to non-AD latent confounders. See Appendix S11 for further investigations of these genes that remain significant after adjusting for latent confounders.

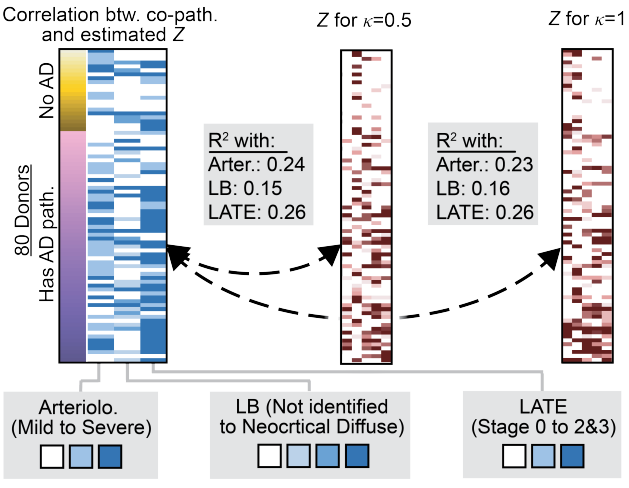

Figure 7. **sensGAN prioritizes DEGs specific to microglia in AD.** Donors (rows) and their measured co-pathology, as well as the estimated latent confounders for $\kappa \in \{0.5, 1\}$.

Next, we illustrate that sensGAN's latent confounders are suggestive of currently measured co-occurring neurodegenerative disease. While we did not use any co-pathologies in our sensGAN analysis, the SEA-AD consortium also staged all the post-mortem tissues for arteriolosclerosis (i.e., a vascular disease), Lewy bodies, and Limbic-predominant age-related TDP-43 encephalopathy (LATE). We compute how predictive sensGAN's learned confounders predict these co-pathologies (Fig. 5C). Surprisingly, for $\kappa \in [0.5, 1]$, the predictive power measured by R^2 between the predicted (linear regression) and observed co-pathology stages is significantly non-zero and not changing. This suggests that sensGAN latent confounders are capturing co-pathology impacts from snRNA-seq data, even with a modest value of $\kappa = 0.5$. We hypothesize that the latent confounders estimated by sensGAN additionally capture effects beyond co-pathology, for instance, the effects of spatial microenvironments, genetics, or environmental factors (e.g., diet) on the donor that affect AD. Further investigations are shown in Appendix S11.

3.4. Discussion

We presented sensGAN, an adversarial framework that shifts the focus of high-dimensional genomic analysis from binary significance to quantitative robustness. Unlike traditional surrogate-variable methods that provide single-point estimates of latent factors, sensGAN explicitly explores the confounding spectrum by learning "worst-case" latent variables under controlled predictive-gain constraints. Our simulations demonstrate that sensGAN maintains high fidelity in separating robust biological signals from those easily nullified by hidden confounders, effectively capturing a continuum of plausible confounding scenarios that point-estimation methods often miss. By treating confounding as a formal machine learning optimization problem, sensGAN provides a principled metric for determining how strong a latent process must be to explain away a discovery.

Applied to complex diseases, sensGAN successfully "purified" differential expression results, isolating pathways specifically associated with AD from signals likely driven by co-occurring neurodegenerative disease. For instance, our framework prioritized apoptosis-related processes as robust AD-specific drivers, while identifying transmembrane transport genes as highly sensitive to unmeasured co-occurring neurodegenerative disease, such as Lewy body pathology. While currently limited by its reliance on pseudobulk aggregation and calibrated (rather than absolute) effect sizes, sensGAN establishes a unified, future-oriented approach to interpreting single-cell genomics in the presence of pervasive, evolving confounding.

Impact statement

This work introduces an adversarial framework to quantify the robustness of high-dimensional genomic findings against latent confounding. By transitioning from binary significance to a quantitative exploration of the confounding spectrum, this method facilitates more reliable interpretation of biological mechanisms and the prioritization of robust therapeutic targets. This advancement enhances the reproducibility and trustworthiness of computational biology, with potential long-term societal benefits for precision medicine across diverse complex diseases.

References

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):39–67, 2020.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22 (1):173–203, 1959.

Dou, R.-x., Zhang, Y.-m., Hu, X.-j., Gao, F.-L., Zhang, L.-L., Liang, Y.-h., Zhang, Y.-y., Yao, Y.-p., Yin, L., Zhang, Y., et al. Aβ1-42 promotes microglial activation and apoptosis in the progression of AD by binding to TLR4. *Redox Biology*, 78:103428, 2024.

Du, J.-H., Shen, M., Mathys, H., and Roeder, K. Causal differential expression analysis under unmeasured confounders with causarray. *bioRxiv*, pp. 2025–01, 2025.

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):2366–2383, 2024.

Karch, C. M. and Goate, A. M. Alzheimer's disease risk genes and mechanisms of disease pathogenesis. *Biological Psychiatry*, 77(1):43–51, 2015.

Leek, J. T. and Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

Leek, J. T. and Storey, J. D. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008.

Liu, J.-Y., Chen, X.-X., Chen, H.-Y., Shi, J., Leung, G. P.-H., Tang, S. C.-W., Lao, L.-X., Yip, H. K.-F., Lee, K.-F., Sze, S. C.-W., et al. Downregulation of aquaporin 9 exacerbates beta-amyloid-induced neurotoxicity in Alzheimer's disease models in vitro and in vivo. *Neuroscience*, 394: 72–82, 2018.

Lorenzini, I., Alsop, E., Levy, J., Gittings, L. M., Lall, D., Rabichow, B. E., Moore, S., Pevey, R., Bustos, L. M., Burciu, C., et al. Moderate intrinsic phenotypic alterations in C9orf72 ALS/FTD iPSC-microglia despite the presence of C9orf72 pathological features. *Frontiers in Cellular Neuroscience*, 17:1179796, 2023.

Love, M. I., Huber, W., and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

Matza, D., Badou, A., Kobayashi, K. S., Goldsmith-Pestana, K., Masuda, Y., Komuro, A., McMahon-Pratt, D., Marchesi, V. T., and Flavell, R. A. A scaffold protein, AHNAK1, is required for calcium signaling during T cell activation. *Immunity*, 28(1):64–74, 2008.

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):479–479, 1991.

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):1–11, 2012.

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):1503–1527, 2019.

Park, Y. P. and Kellis, M. Cocoa-diff: Counterfactual inference for single-cell gene expression analysis. *Genome Biology*, 22(1):228, 2021.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9):896–902, 2014.

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):2557–2569, 2023.

Santpere, G., Garcia-Esparcia, P., Andres-Benito, P., Lorente-Galdos, B., Navarro, A., and Ferrer, I. Transcriptional network analysis in frontal cortex in Lewy body diseases with focus on dementia with Lewy bodies. *Brain Pathology*, 28(3):315–333, 2018.

Sarkar, A. and Stephens, M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nature Genetics*, 53(6):770–777, 2021.

Scheltens, P., Blennow, K., Breteler, M. M., De Strooper, B., Frisoni, G. B., Salloway, S., and Van der Flier, W. M. Alzheimer's disease. *The Lancet*, 388(10043):505–517, 2016.

Siew, J. J., Chen, H.-M., Chiu, F.-L., Lee, C.-W., Chang, Y.-M., Chen, H.-L., Nguyen, T. N. A., Liao, H.-T., Liu, M., Hagar, H.-T., et al. Galectin-3 aggravates microglial activation and tau transmission in tauopathy. *The Journal of Clinical Investigation*, 134(2), 2024.

Spina, S., La Joie, R., Petersen, C., Nolan, A. L., Cuevas, D., Cosme, C., Hepker, M., Hwang, J.-H., Miller, Z. A., Huang, E. J., et al. Comorbid neuropathological diagnoses in early versus late-onset Alzheimer's disease. *Brain*, 144(7):2186–2198, 2021.

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J., Barraud, Q., et al. Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12(1):5692, 2021.

Tan, Y., Zheng, Y., Xu, D., Sun, Z., Yang, H., and Yin, Q. Galectin-3: A key player in microglia-mediated neuroinflammation and Alzheimer's disease. *Cell & Bioscience*, 11(1):78, 2021.

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*, 33:10999–11009, 2020.

Wang, E., Yu, K., Cao, J., Wang, M., Katsel, P., Song, W.-m., Wang, Z., Li, Y., Wang, X., Wang, Q., et al. Multiscale proteomic modeling reveals protein networks driving Alzheimer's disease pathogenesis. *Cell*, 188(22): 6186–6204, 2025.