A Normative Variational Autoencoder Scores Per-Cell Vulnerability in Parkinson's Disease

Guilherme Felype Herculano da Silva[1], **Karla Haryanna Santos Moura**[2], **Luana Dantas Pontes Espinola Casado**[3]

[1]Universidade Potiguar (UnP) – Natal, RN – Brazil

[2]Universidade Estadual do Ceará (UECE) – Fortaleza, CE – Brazil

[3]Universidade Federal Rural do Semi-Árido (UFERSA) – Mossoró, RN – Brazil

guilhermefelype9@gmail.com, karlaharyanna@gmail.com, luanadantasespinola@gmail.com

Abstract. *We recast per-cell vulnerability in Parkinson's disease (PD) as deviation from a healthy norm—a transcriptome's distance from healthy, not a prediction of which cells die. A variational autoencoder trained* only *on healthy midbrain nuclei gives each nucleus a deviation score, calibrated per cell type against a held-out healthy null. On the Smajić et al. (2022) cohort it separates PD from healthy nuclei (AUC* $= 0.623$*) with* no disease label*, concentrating in glia and vasculature; without marker genes, high-deviation glia carry the known reactive and disease-associated programs (*$\rho = 0.32, 0.30$*). The effect holds at donor level (exact permutation* $p = 0.002$*) and under an unseen-donor null, and reproduces in one external cohort but not a second.*

1. Introduction

Parkinson's disease (PD) is defined neuropathologically by the progressive loss of dopaminergic (DA) neurons of the substantia nigra pars compacta (SNpc). A central open question is *selective vulnerability*: why some neurons degenerate while neighbours in the same microenvironment persist [Surmeier et al. 2017, Poulin et al. 2020]. Single-nucleus RNA sequencing (snRNA-seq) lets us interrogate this at cellular resolution [Kamath et al. 2022, Smajić et al. 2022], yet the dominant paradigm remains the *case–control* contrast: cells are clustered, annotated, and the average expression of a predefined population compared between groups.

Two limitations of that paradigm are acute for neurodegeneration. It is *categorical*: a cell inherits a group-level effect from its cluster, whereas vulnerability is plausibly continuous and cell-intrinsic. And it is *label-dependent*: results hinge on clustering resolution and on the supply of disease cells—exactly the cells PD destroys. DA neurons are well below 1% of recovered nuclei in nigral snRNA-seq [Kamath et al. 2022, Smajić et al. 2022], so case–control power is scarcest for the population of greatest interest.

We borrow a different framing from clinical neuroscience: *normative modelling* [Marquand et al. 2016, Marquand et al. 2019]. Rather than contrasting two groups, a normative model learns the full distribution of a *healthy* reference population and expresses

each new observation as a deviation from that norm—like a paediatric growth chart. Deviations are continuous, defined per individual (here, per cell), and need *no* disease labels. We instantiate this with a deep generative model of the transcriptome.

Contributions. (i) We formulate per-cell vulnerability in PD as a normative-deviation problem, instantiated with a VAE trained *only* on healthy midbrain nuclei and scored by a k-NN and a density statistic, each calibrated per cell type against a held-out healthy null. (ii) The deviation concentrates, unsupervised, in the glial and vascular microenvironment, and the high-deviation glia carry known reactive-astrocyte and disease-associated-microglia programs down to individual genes—a label-free rediscovery of established disease biology. (iii) We benchmark against novelty detectors and a supervised reference, and separate the effect of covariate conditioning from that of the deviation statistic. (iv) We test whether the result is about donors rather than nuclei, using an exact donor-level permutation test and an unseen-donor null. (v) We test transfer to two independent cohorts, with and without batch-aware reference mapping: the *astrocytic* deviation reproduces in one [Kamath et al. 2022] but not the other [Martirosyan et al. 2024], which bounds label-free cross-cohort transfer.[1]

Structure. Section 2 reviews related work; Section 3 describes the cohorts, the normative model and the evaluation protocol; Section 4 reports the results, including donor-level inference and the robustness analyses; and Section 5 discusses what the deviation does and does not support.

2. Related Work

Single-cell studies of Parkinson's disease. snRNA-seq of the human midbrain has reshaped the study of selective vulnerability. The founding study of this cohort [Smajić et al. 2022] profiled the ventral midbrain of controls and idiopathic PD patients and reported, through supervised cluster-based analysis, a pronounced glial activation and a PD-specific neuronal state. A deep survey of nigral DA neurons [Kamath et al. 2022] identified a specific *SOX6*$^+$ *AGTR1*$^+$ subtype that is preferentially lost in PD—a result that required enriching for DA nuclei, precisely the effort a microenvironment-level readout avoids. A third study [Martirosyan et al. 2024] added an SNpc atlas of 15 PD and 14 control donors and resolved glial heterogeneity into subpopulations, including an astrocyte state carrying the reactive markers *C3* and *CD44* and a PD-enriched reactive state; its donor composition differs markedly from the other two, which becomes important in Section 4.6. Common to all three is a cluster-level case–control contrast: cells inherit a group effect, and no cell carries its own score. A complementary line of work casts the problem as *supervised* per-cell classification, reaching ≈ 0.92 mean balanced accuracy across four midbrain cohorts [Fiorini et al. 2025]—but such models need disease labels and inherit the case–control framing that motivates this work. We instead reuse these cohorts as a healthy reference (Smajić) and as independent validation (Kamath, Martirosyan) for a label-free, per-cell deviation.

Generative models, reference mapping and anomaly detection. The variational autoencoder [Kingma and Welling 2014] underlies `scVI` [Lopez et al. 2018], which mod-

[1] Code and data to reproduce all results: `https://anonymous.4open.science/r/parkinson-normative-vae-FD45/`

els raw counts with a negative-binomial likelihood and has become standard for integration and denoising; such models are normally trained on *all* cells to obtain a shared embedding, whereas we train on healthy cells only and use the generative density as a normative reference. Transfer-learning methods such as `scArches` [Lotfollahi et al. 2022] project query cells onto a reference atlas, and recent work frames diseased cells as anomalies [Xu et al. 2024, Mallick et al. 2023]. Our goal differs: not labelling cells anomalous but assigning a *graded* deviation comparable across cell types and individuals, in the spirit of normative charts.

Normative modelling. Normative models are established in neuroimaging, where deviation maps from a healthy reference quantify individual-level pathology beyond case–control means [Marquand et al. 2016, Marquand et al. 2019]. To our knowledge their transfer to single-cell transcriptomics of neurodegeneration, with deviation defined at the resolution of individual nuclei, has not been explored.

3. Materials and Methods

3.1. Data

We use the human midbrain snRNA-seq atlas of the Smajić cohort [Smajić et al. 2022] (GEO `GSE157783`): 11 donors comprising 6 controls (22,433 nuclei) and 5 idiopathic PD patients (19,002 nuclei), annotated into 12 cell types spanning the glial microenvironment (oligodendrocytes, astrocytes, microglia, oligodendrocyte precursor cells, OPCs), vascular cells (endothelial cells, pericytes), neurons (excitatory, inhibitory, GABAergic, `CADPS2+`) and DA neurons, which the source annotation labels *DaNs*; we keep that label in the tables. Consistent with the difficulty of recovering nigral DA neurons, they are extremely rare here (74 nuclei), which motivates a method that does not depend on their abundance. Two further cohorts, used only as external validation, are summarised with the data divisions in Table 1.

Table 1. Cohorts and data divisions. Only healthy control nuclei of the discovery cohort train the model; its held-out control nuclei form the null. Every PD nucleus, and every nucleus of the two external cohorts, is test data scored with the frozen model. †Kamath donor counts are inferred from sample identifiers (38 control, 34 PD SNpc samples).

Cohort	GEO	Role	Donors	Nuclei	Train / null / test
Smajić	GSE157783	discovery	6 / 5	41,435	17,947 / 4,486 / 19,002
Kamath	GSE178265	ext. validation	9 / 6†	120,000	— / — / 99,147
Martirosyan	GSE243639	ext. validation	14 / 15	83,484	— / — / 60,007

3.2. Preprocessing

Raw unique-molecular-identifier (UMI) counts were retained in a dedicated layer. Genes detected in fewer than three nuclei and nuclei with fewer than 200 detected genes were removed. We selected the top 2,000 highly variable genes (HVGs) with the `seurat_v3` flavour computed on raw counts [Wolf et al. 2018]. This selection was run on the full object, PD nuclei included; because that lets test information reach the feature space, Section 4.7 repeats the entire analysis with HVGs recomputed on the training controls alone.

Control nuclei were partitioned at random *by nucleus* into a training set (17,947 nuclei, 80%) used to fit the model and a held-out set (4,486 nuclei, 20%) used *only* to calibrate the healthy null. All PD nuclei were held out for testing. Because this split is by nucleus, the healthy null and the training set share donors, whereas every PD nucleus comes from an unseen donor; Section 4.7 isolates that asymmetry with a leave-one-control-donor-out analysis.

3.3. Normative variational autoencoder

Figure 1 summarises the architecture and the scoring pipeline.

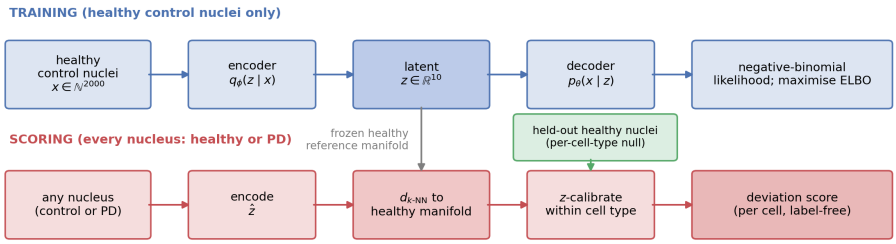


Figure 1. The normative model. Training (top) uses *only* healthy control nuclei. Scoring (bottom) encodes any nucleus, measures its k-NN distance to the frozen healthy manifold and standardises it within cell type against the held-out healthy null. No disease label is used at any stage.

We model the healthy transcriptome with a VAE following `scVI` [Lopez et al. 2018]. For a nucleus with count vector $x \in \mathbb{N}^G$ over $G = 2{,}000$ genes, an encoder $q_\phi(z \mid x)$ maps x to a latent $z \in \mathbb{R}^{10}$ with a standard-normal prior, and a decoder parametrises a negative-binomial likelihood $p_\theta(x \mid z, \ell)$ with library size ℓ. Parameters are optimised by maximising the evidence lower bound (ELBO)

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x \mid z, \ell)\big] - \mathrm{KL}\big(q_\phi(z \mid x) \,\|\, p(z)\big). \tag{1}$$

where $\mathrm{KL}(\cdot \,\|\, \cdot)$ is the Kullback–Leibler divergence and $p(z) = \mathcal{N}(0, I)$ the standard-normal prior. Crucially, the expectation is taken *only* over healthy training nuclei, so p_θ becomes a generative density of *normality*. We used two hidden layers of 128 units, a 10-dimensional latent space, and trained for up to 200 epochs with early stopping (patience 20) on a CPU.

3.4. Covariate-conditional normative model

A normative model in the strict sense expresses the norm as a *function of covariates*, like an age-conditioned growth chart [Marquand et al. 2019]. That is not cosmetic here: control donors are on average ∼8 years older than PD donors (83 vs. 75), so an unconditioned deviation can absorb age. We therefore also train a *covariate-conditional* VAE whose encoder and decoder receive donor age and sex, so deviation is measured against the *matched* healthy norm.

3.5. Per-cell deviation score

A deviation measured against the *global* healthy mean would conflate cell type with disease, since rare types sit far from that mean whatever their condition. We therefore define

a *cell-type-aware* deviation. Nuclei are encoded to their latent mean $\hat{z} \in \mathbb{R}^{10}$, and each is given the average Euclidean distance to its $k = 30$ nearest neighbours *among the healthy training nuclei*,

$$d_{k\text{-NN}}(x) = \frac{1}{k} \sum_{j \in \mathcal{N}_k^{\text{healthy}}(\hat{z})} \|\hat{z} - \hat{z}_j\|_2, \tag{2}$$

where $\mathcal{N}_k^{\text{healthy}}(\hat{z})$ denotes the index set of the k healthy training nuclei nearest to $\hat{z}$. Being local, the score respects cell-type structure. It is then standardised *within each cell type* against the held-out healthy null: for a cell of type t, $\text{dev}(x) = (d_{k\text{-NN}}(x) - \mu_t)/\sigma_t$, with μ_t, σ_t estimated on the 4,486 validation-control nuclei of type t (a global null is used for types with fewer than ten). Each type's healthy null is thus centred at zero, so a positive deviation measures departure from the healthy manifold of the *same* cell type, in units of healthy variability.

Density-based deviation. As a probabilistic alternative to the non-parametric k-NN distance we fit a Gaussian-mixture density (15 components) on the control-train latent and score each cell by its negative log-density, which also yields a per-cell empirical p-value against the type-matched healthy null. We report both statistics.

3.6. Evaluation and baselines

Because training never sees disease labels, the separation of PD from held-out healthy nuclei is an unsupervised test of the hypothesis that PD cells depart from the healthy manifold. The held-out test set is class-imbalanced (all 19,002 PD nuclei vs. 4,486 held-out healthy nuclei, prevalence 0.81, a $4.2{:}1$ ratio). We deliberately do *not* oversample or synthesise minority-class cells (e.g. SMOTE): in a normative model the healthy training set *is* the reference manifold, and perturbing it would corrupt the very quantity being estimated. We instead report imbalance-robust metrics: the area under the ROC curve (AUC), which is prevalence-independent, with a $1000\times$ bootstrap 95% confidence interval; a *class-balanced* AUC that downsamples the majority class to confirm the score is not an imbalance artefact; the area under the precision–recall curve (PR-AUC) read against the prevalence baseline; balanced accuracy and Matthews correlation at the Youden-optimal threshold; and the sensitivity at 90% specificity, fixing the operating point on the healthy null (Table 3). These answer different questions: AUC and PR-AUC measure ranking quality; balanced accuracy and Matthews correlation measure decision quality at one threshold, and are near zero for a useless classifier; sensitivity at fixed specificity gives the practical yield once the false-alarm rate on healthy cells is pinned. Cliff's δ reports the size of a shift—the probability that a random PD cell is more deviant than a random healthy one, rescaled to $[-1, 1]$—which a p-value alone does not convey. We further report a one-sided Mann–Whitney U test and per-cell-type effect sizes (Cliff's δ) with Benjamini–Hochberg correction. To position the approach, we benchmark it against standard label-free novelty detectors fit on the same healthy nuclei—Isolation Forest, One-Class SVM, and PCA reconstruction error—and against a supervised logistic regression (which uses disease labels) as an upper-bound reference. Implementation used `scvi-tools`, `scanpy` [Wolf et al. 2018] and `scikit-learn`.

4. Results

4.1. Deviation separates PD from healthy nuclei without supervision

Fit on healthy nuclei alone, the latent space still organises the 12 midbrain cell types cleanly, and the deviation is low across the healthy-dominated core and elevated where PD nuclei concentrate. We first report the unconditional model with the k-NN deviation. The score is right-shifted for PD relative to held-out healthy nuclei (median 0.23 vs. -0.19). As a classifier of PD vs. held-out control nuclei it reaches AUC $= 0.623$ (95% CI $[0.614, 0.632]$, one-sided Mann–Whitney $p = 2.8 \times 10^{-146}$) with no disease label in training. This is not an imbalance artefact: downsampling the majority class to parity leaves the AUC unchanged, and the imbalance-robust panel (PR-AUC 0.87 against a 0.81 prevalence baseline, balanced accuracy 0.59; Table 3) confirms a modest but genuine per-cell signal. Sections 4.6 and 4.7 then put this number under pressure.

4.2. Deviation concentrates in the glial and vascular microenvironment

Stratifying by cell type (Table 2), the deviation is far from uniform. Six cell types are significantly more deviant in PD after Benjamini–Hochberg correction (FDR < 0.05), and they form a coherent group: oligodendrocytes (Cliff's $\delta = 0.42$), endothelial cells (0.38), astrocytes (0.24), microglia (0.23), ependymal cells (0.18) and pericytes (0.11). The glial and vascular compartment—precisely the populations whose activation defines this cohort [Smajić et al. 2022]—is thus recovered as the locus of PD deviation by a model that never saw a disease label. Excitatory and inhibitory neurons were not elevated, and DA neurons were too few (42 PD vs. 6 control) for a reliable test—the DA under-representation that limits nigral snRNA-seq, and that motivates a microenvironment-level rather than DA-centric readout.

Table 2. Per-cell-type deviation in PD vs. held-out healthy nuclei. Cliff's δ with BH-adjusted one-sided Mann–Whitney p, ordered by effect size.

Cell type	n_{PD}	n_{ctrl}	Cliff's δ	p_{adj}
Oligodendrocytes	8536	2618	0.42*	<.001
Endothelial cells	718	182	0.38*	<.001
Astrocytes	2565	410	0.24*	<.001
Microglia	2698	207	0.23*	<.001
Ependymal	325	45	0.18*	0.049
Pericytes	571	132	0.11*	0.049
GABA	249	59	0.10	0.174
OPCs	1216	310	0.02	0.439
Excitatory	1216	372	-0.30	1.000
DaNs	42	6	-0.33	1.000
Inhibitory	749	145	-0.39	1.000

4.3. High-deviation glia carry known disease-associated programs

We scored every nucleus for two literature-defined programs: the reactive / disease-associated astrocyte program [Liddelow et al. 2017, Habib et al. 2020] and the disease-associated microglia (DAM) program [Keren-Shaul et al. 2017]. The deviation tracks both without using these markers: within astrocytes it correlates with the reactive program (Spearman $\rho = 0.32$, $p < 10^{-100}$) and high-deviation astrocytes carry a far stronger

reactive signature than low-deviation ones (Cliff's $\delta = 0.43$); within microglia it tracks DAM ($\rho = 0.30$; high-vs-low $\delta = 0.38$; Figure 2). The unsupervised deviation thus rediscovers the *same* states supervised studies identify with curated markers—a validation independent of the AUC.

This holds gene by gene, so the score is interpretable rather than opaque. Ranking every gene by its correlation with the deviation *within* astrocytes puts four members of the reactive signature (*CP*, *VIM*, *CD44*, *GBP2*) in the top fifteen; the leading microglial gene is *CD163* ($\rho = 0.51$), a marker of activated phagocytic microglia, with the DAM gene *GPNMB* also in the top fifteen (Figure 2, right). Neither ranking uses a signature. Because library size can move both latent position and signature scores, we recomputed the correlations partialling out total counts and detected genes: the astrocyte association is essentially unchanged ($\rho = 0.32 \rightarrow 0.30$) and the microglial one weakens but persists ($0.30 \rightarrow 0.21$). Mitochondrial fraction cannot confound here: the matrix carries no mitochondrial genes.

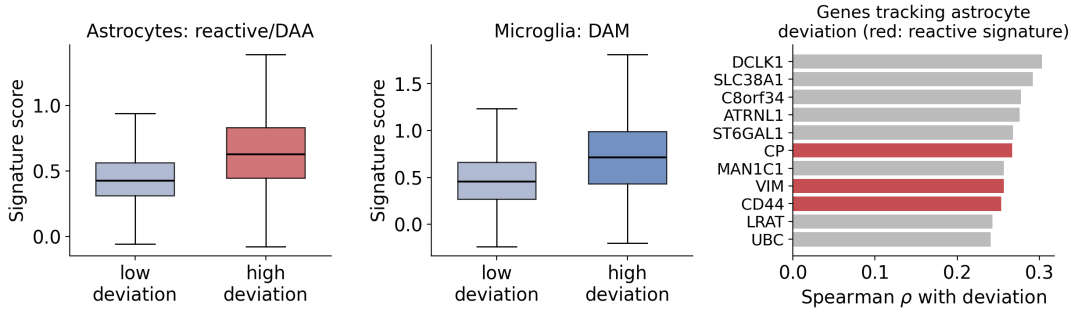


Figure 2. High-deviation glia express known disease-associated programs, down to individual genes. Reactive/DAA signature in astrocytes (left) and DAM signature in microglia (centre), in low- vs. high-deviation tertiles; genes most correlated with the deviation within astrocytes (right), with members of the reactive signature in red. Nothing here uses the marker lists.

4.4. Disentangling covariate conditioning from the deviation statistic

Because the controls are markedly older than the PD donors, part of the unconditional deviation could reflect age rather than disease. Conditioning the normative model on age and sex tests this—but conditioning and the choice of deviation statistic must be varied separately, since the conditional model was originally paired with the density statistic. Crossing both factors gives AUC 0.623 (unconditional, k-NN), 0.561 (unconditional, GMM), 0.555 (conditional, k-NN) and 0.539 (conditional, GMM). The two main effects are comparable: conditioning costs 0.045 AUC and switching from k-NN to the mixture density costs 0.039. The drop from 0.62 to 0.54 reported for the conditional density model is therefore *not* attributable to age alone, as we previously implied; roughly half of it is the weaker statistic. What survives is the qualitative claim: conditioning on age does remove part of the naive signal, so an unadjusted normative deviation in this cohort is partly an age contrast. The astrocytic effect is positive in three of the four cells ($\delta = 0.24$, 0.16, 0.12) and vanishes only when the mixture density is paired with the unconditional latent (-0.01), where that statistic is weakest.

4.5. Comparison with baseline novelty detectors

Table 3 places the approach among standard label-free detectors, and the picture is twofold. *No* unsupervised method is a strong global classifier: all lie between AUC 0.51 and 0.62, far below the supervised bound (0.82). But the *astrocytic* deviation is positive across every method (largest for a simple PCA reconstruction error, $\delta = 0.29$), so the signal does not depend on the particular model. The normative framing adds not a higher AUC but a calibrated, per-cell, covariate-adjusted and interpretable readout.

Table 3. Evaluation. (a) Label-free deviation methods, in-cohort: all are fit on healthy nuclei with the same per-cell-type calibration; the supervised row is an upper-bound reference. (b) Imbalance-robust metrics for the two normative models (prevalence 0.81, a 4.2:1 ratio); the class-balanced AUC matches the raw AUC to three decimals, so the score is not an imbalance artefact. No oversampling is used: it would corrupt the reference manifold.

(a)

Method (label-free unless noted)	PD-vs-ctrl AUC	Astrocyte δ
scVI latent + k-NN (unconditional)	0.623	$+0.24$
PCA reconstruction error	0.569	$+0.29$
Isolation Forest	0.544	$+0.09$
Covariate-conditional VAE + GMM density	0.539	$+0.12$
One-Class SVM	0.507	-0.06
Supervised logistic regression †	0.817	—

† uses disease labels; reference upper bound, not a normative method.

(b)

Model	AUC [95% CI]	PR-AUC	Bal. acc.	MCC	Se@90Sp
Unconditional VAE + k-NN	0.623 [0.614, 0.632]	0.866	0.591	0.143	0.206
Covariate-conditional VAE + GMM	0.539 [0.530, 0.549]	0.832	0.531	0.054	0.143

Positioning against the literature. The 0.623 should be read against *label-free* methods. Supervised per-cell predictors trained *with* labels on these cohorts reach ≈ 0.92 balanced accuracy [Fiorini et al. 2025] (our supervised reference recovers 0.82), whereas unsupervised *normative* models sit far lower—deviation-based autoencoders in neuroimaging report AUROC 0.62–0.69 [Marquand et al. 2019]. Our 0.623 sits squarely there: the contribution is not a stronger classifier but the transfer of a normative, label-free, per-cell readout to single-nucleus transcriptomics.

4.6. Cross-cohort transfer: one replication, one failure, and a batch-aware test

We applied the *frozen* discovery model to two independent atlases (Table 1): query nuclei embedded with the discovery encoder, cell types transferred by nearest neighbours, deviation calibrated within each predicted type against *that cohort's own* controls so a constant platform offset cancels. Global discrimination degraded in both (AUC ≈ 0.52 and 0.44), as expected without batch correction, so the informative quantity is the per-cell-type effect (Figure 3).

On Kamath the astrocytic deviation reproduced the discovery effect almost exactly (Cliff's $\delta = 0.24$ in both; 0.27 with the conditional density model), while the other glial

and vascular effects largely vanished ($\delta \leq 0.06$). On Martirosyan nothing replicated: the astrocyte effect reversed ($\delta = -0.21$, n.s.). We traced this rather than asserting a cause: it is not an artefact of label transfer—the authors' *own* annotations give the same result—nor of age alone, since conditioning only shrinks δ to -0.13.

Because our own explanation was batch, we tested it rather than assuming it, building a batch-aware reference with scArches-style architectural surgery [Lotfollahi et al. 2022] and mapping each query cohort onto it before scoring. Batch correction helps where signal exists and cannot create it where it does not: on Kamath the astrocytic effect survives ($\delta = 0.24 \rightarrow 0.21$, $p_{\text{adj}} = 6 \times 10^{-68}$) and global discrimination improves ($0.52 \rightarrow 0.57$), whereas on Martirosyan every cell type stays negative ($\delta = -0.21 \rightarrow -0.19$, AUC 0.42). The non-replication therefore survives all three principled explanations—label transfer, covariate structure and batch—and is best read as a genuine between-cohort difference.

Crucially, the discrepancy is statistical rather than biological: that cohort's own authors report reactive astrocyte states in PD SNpc—an *Astrocytes1* subpopulation carrying the canonical reactive markers *C3* and *CD44*, and a PD-enriched *Astrocytes3* reactive state [Martirosyan et al. 2024]. The reactivity our score targets (*CD44* also drives our grounding signature) is thus corroborated there by supervised subtyping; what fails to transfer is the *label-free normative mapping*, not the disease biology.

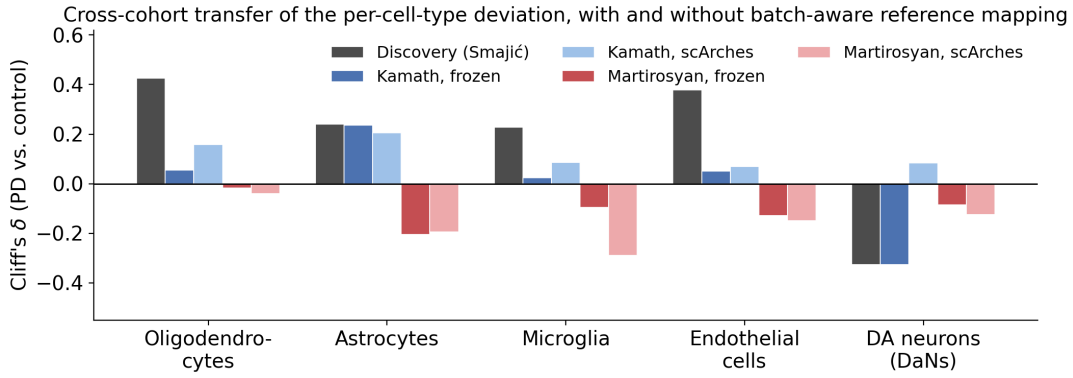


Figure 3. Cross-cohort transfer. Only the astrocytic effect reproduces on Kamath, and it survives batch-aware mapping (scArches); on Martirosyan every effect is negative with and without batch correction, so batch does not explain the non-replication.

4.7. Is this a statement about nuclei, or about donors?

Every test so far is computed over tens of thousands of nuclei from only 11 donors, so a p-value of 2.8×10^{-146} describes nuclei, not people: one atypical donor could in principle produce it. Averaging the deviation within each donor and enumerating all 462 ways of assigning five PD labels among 11 donors gives an *exact* permutation test, and the observed separation is the most extreme attainable ($p = 0.002$): every PD donor's mean exceeds every control donor's (Figure 4). A mixed model over nuclei with donor as a random intercept agrees ($\beta = 0.44$, $p = 4.6 \times 10^{-4}$). Per cell type the evidence thins, as it should with $n = 11$: oligodendrocytes remain significant under both tests, astrocytes under the mixed model ($p = 0.009$) and marginally under the permutation test

($p_{\mathrm{adj}} = 0.052$), whereas microglia do not ($p = 0.51$). The nucleus-level statistics had overstated the microglial case, and we report the cell-type claims at donor level from here on.

A second asymmetry cuts deeper. Because the control split is by nucleus, held-out healthy nuclei share donors with the training set while every PD nucleus comes from an unseen donor—out of distribution in two ways at once. We separated the two by retraining the model six times, each time excluding one control donor entirely and using it as the healthy null. Discrimination falls from 0.623 to 0.577 ± 0.029 (Figure 4): part of the original figure was donor novelty rather than disease, though the separation does not disappear. Effect sizes shrink with it and the ranking thins: oligodendrocytes remain the strongest population ($\delta = 0.30 \pm 0.13$ across folds) and astrocytes stay positive though markedly more variable (0.13 ± 0.13), whereas the endothelial effect—strong in discovery—does not survive an unseen-donor null at all (0.00). This also explains the modest cross-cohort AUCs of the previous section: an unseen cohort is an unseen donor, amplified.

Table 4 collects the remaining checks; two deserve emphasis. Feature selection had been run on the full object, PD nuclei included; recomputing the highly variable genes on the training controls alone changes 11% of the gene set and leaves discrimination at 0.602. And the score does *not* predict which populations are lost: mean per-type deviation is uncorrelated with the PD-to-control abundance ratio ($\rho = 0.33$, $p = 0.30$)—the honest limit of the word *vulnerability*, and the reason we narrowed the paper's claims.

Table 4. Robustness and sensitivity analyses. Each row states what the analysis puts at risk and what came out; the first three address whether the effect is a statement about donors rather than about nuclei.

Analysis	Question	Outcome
Donor-level exact permutation	all 462 label assignments over 11 donors	$p = 0.002$; donor means fully separated
Mixed model, donor random intercept	nucleus-level effect with donor clustering	$\beta = 0.44$, $p = 4.6 \times 10^{-4}$
Leave-one-control-donor-out	healthy null from an *unseen* donor	AUC $= 0.577 \pm 0.029$ vs. 0.623
HVG recomputed on training controls	feature-selection leakage	88.8% gene overlap; AUC $= 0.602$
Retraining under new seeds	does re-initialisation change the ordering?	rank $\rho \geq 0.84$; AUC 0.623–0.623
Neighbourhood size $k \in [10, 100]$	stability of the cell-type ordering	rank $\rho \geq 0.94$; AUC 0.596–0.641
2×2: latent $\times$ statistic	attribution of the conditional-model drop	conditioning $+0.045$, statistic $+0.039$ AUC
Partial correlation on library size	technical confounding of the grounding	astrocytes $\rho\, 0.32 \rightarrow 0.30$
Deviation vs. abundance ratio	does deviation predict cell loss?	$\rho = 0.33$, $p = 0.30$ (n.s.)

5. Discussion

Our results support the central claim that vulnerability in PD can be expressed as a *continuous, per-cell deviation* from a healthy norm, learned without disease labels—while delimiting what such a deviation delivers. On the positive side it is *spatially structured*

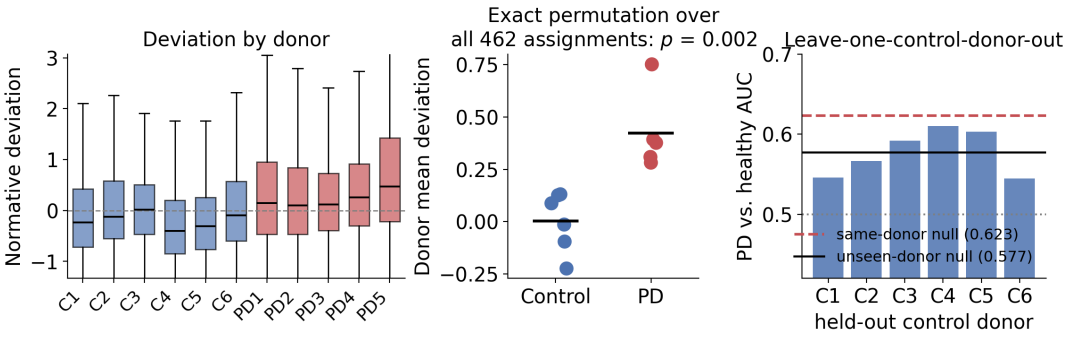


Figure 4. Donor-level inference. Left: per-donor deviation (blue control, red PD). Centre: donor means with the exact permutation test over all 462 assignments. Right: leave-one-control-donor-out; the dashed line is the published same-donor result.

across cell types, concentrating unsupervised in glia and the vasculature and so recovering the glial activation that the original analysis of this cohort [Smajić et al. 2022] obtained by supervised clustering. Referencing each cell to a healthy null of its own type also sidesteps dependence on rare DA-neuron recovery. Most importantly, the high-deviation glia are not a statistical artefact: they carry the reactive astrocyte and DAM programs that supervised studies define with curated markers [Liddelow et al. 2017, Habib et al. 2020, Keren-Shaul et al. 2017], gene by gene and after adjusting for library size—and the effect is a statement about donors, not merely nuclei, surviving both an exact donor-level permutation test and an unseen-donor null.

Four findings keep the claims honest. First, *no* label-free method—ours or standard novelty detectors—is a strong global classifier (AUC ≤ 0.62 vs. 0.82 supervised), so the per-cell-type effect sizes, not the pooled AUC, are the substantive result. Second, conditioning on age *reduces* global separation, exposing that part of the naive signal tracks the age gap between groups. Third, transfer is uneven, and the honest reading splits by population: oligodendrocytes carry the largest and most robust *within-cohort* effect—they alone survive both donor-level tests and the unseen-donor null—whereas astrocytes, weaker in discovery, are the only population reproducing in an external cohort, and not in the second one, whose own subtyping nonetheless reports reactive astrocytosis (Section 4.6). Neither is yet a cohort-invariant biomarker, and a batch-naive frozen model should not be expected to give one. Fourth, *vulnerability* must be read in the restricted sense we test: mean per-type deviation does *not* predict which populations are depleted in PD ($\rho = 0.33$, $p = 0.30$). The score measures how far a transcriptome sits from the healthy norm, which is not the same as measuring which cells die; these data cannot bridge that gap, and we make no claim that they do.

6. Conclusion

We introduced a normative variational autoencoder of the healthy human midbrain that gives each nucleus a calibrated, continuous deviation score without disease supervision. The deviation concentrates in the glial and vascular microenvironment, carries known reactive-astrocyte and DAM programs down to individual genes, and holds at donor level and under an unseen-donor null. It reproduces in one independent cohort but not a second, and does not predict which populations are lost—limits we state rather than smooth over.

Deviation-from- normality yields a label-free, per-cell instrument whose generalisation is an explicit, measurable property; making it platform-robust is the next step.

References

Fiorini, G., Li, Y., Fon, E. A., Farhan, S. M. K., Thomas, R. A., et al. (2025). Neural networks reveal novel gene signatures in parkinson disease from single-nuclei transcriptomes. *npj Parkinson's Disease*, 11.

Habib, N., McCabe, C., Medina, S., Varshavsky, M., Kitsberg, D., et al. (2020). Disease-associated astrocytes in alzheimer's disease and aging. *Nature Neuroscience*, 23(6):701–706.

Kamath, T., Abdulraouf, A., Burris, S. J., Langlieb, J., Gazestani, V., Nadaf, N. M., Balderrama, K., Vanderburg, C., and Macosko, E. Z. (2022). Single-cell genomic profiling of human dopamine neurons identifies a population that selectively degenerates in parkinson's disease. *Nature Neuroscience*, 25(5):588–595.

Keren-Shaul, H., Spinrad, A., Weiner, A., Matcovitch-Natan, O., Dvir-Szternfeld, R., et al. (2017). A unique microglia type associated with restricting development of alzheimer's disease. *Cell*, 169(7):1276–1290.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*.

Liddelow, S. A., Guttenplan, K. A., Clarke, L. E., Bennett, F. C., Bohlen, C. J., et al. (2017). Neurotoxic reactive astrocytes are induced by activated microglia. *Nature*, 541(7638):481–487.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058.

Lotfollahi, M., Naghipourfar, M., Luecken, M. D., et al. (2022). Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40:121–130.

Mallick, K., Chakraborty, S., Mallik, S., and Bandyopadhyay, S. (2023). A scalable unsupervised learning of scrnaseq data detects rare cells through integration of structure-preserving embedding, clustering and outlier detection. *Briefings in Bioinformatics*, 24(3):bbad125.

Marquand, A. F., Kia, S. M., Zabihi, M., Wolfers, T., Buitelaar, J. K., and Beckmann, C. F. (2019). Conceptualizing mental disorders as deviations from normative functioning. *Molecular Psychiatry*, 24(10):1415–1424.

Marquand, A. F., Rezek, I., Buitelaar, J., and Beckmann, C. F. (2016). Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. *Biological Psychiatry*, 80(7):552–561.

Martirosyan, A., Ansari, R., Pestana, F., Hebestreit, K., Gasparyan, H., et al. (2024). Unravelling cell type-specific responses to parkinson's disease at single cell resolution. *Molecular Neurodegeneration*, 19(1):7.

Poulin, J.-F., Gaertner, Z., Moreno-Ramos, O. A., and Awatramani, R. (2020). Classification of midbrain dopamine neurons using single-cell gene expression profiling approaches. *Trends in Neurosciences*, 43(3):155–169.

Smajić, S., Prada-Medina, C. A., Landoulsi, Z., Ghelfi, J., Memet, S., et al. (2022). Single-cell sequencing of human midbrain reveals glial activation and a parkinson-specific neuronal state. *Brain*, 145(3):964–978.

Surmeier, D. J., Obeso, J. A., and Halliday, G. M. (2017). Selective neuronal vulnerability in parkinson disease. *Nature Reviews Neuroscience*, 18(2):101–113.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15.

Xu, K., Ding, Y., Hou, S., Zhan, W., Chen, N., Wang, J., and Sun, X. (2024). Domain adaptive and fine-grained anomaly detection for single-cell sequencing data and beyond. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*.