Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

scRED: Cell-Graph Reconstruction Error Prioritises Autism Risk Genes in Cortical Neurons

Norah Saeed Awn

noorasaeed.91@gmail.com

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Mengyuan Zhao

Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Mansoor S. M. Ba Mahel

College of Intelligence and Computing, Tianjin University

Abdulaziz S. Bamahel

Jiamusi University

Jijun Tang

University of Chinese Academy of Sciences

Research Article

Keywords: autism spectrum disorder, single-nucleus RNA-seq, graph neural network, anomaly detection, gene prioritisation, SFARI, negative control

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11194918/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Autism spectrum disorder (ASD) involves diffuse, cell-type-specific transcriptomic dysregulation in the cerebral cortex. Single-nucleus RNA sequencing (snRNA-seq) has begun to resolve it, but analysis is dominated by per-gene differential expression, which tests one gene at a time and ignores the manifold structure that separates patient from control cells. Graph neural networks offer an alternative, but most approaches place genes at graph nodes and score them by local co-expression rather than by deviation from healthy cell states.

Results

We present scRED, a graph-convolutional autoencoder built on a cell–cell similarity graph rather than a gene–gene graph. The model is trained on control cortical cells; the per-gene deviation in reconstruction error between patient and control cells provides the ranking statistic. Applied to the Velmeshev ASD cortex cohort across 17 cell types, scRED recovered SFARI-curated risk genes in neuronal populations at 1.98-fold enrichment against a matched, donor-disjoint control-versus-control negative control of 1.44-fold, exceeding that control in all 9 neuronal contexts (Wilcoxon signed-rank $p = 0.0039$); glial and endothelial populations showed no separation. Across all 17 contexts the disease contrast exceeded its negative control ($p = 0.0061$). The cell-graph and gene-graph formulations performed comparably (mean difference − 0.2%, $p = 0.82$); both exceeded a same-shape multilayer perceptron, though the primary-depth difference did not survive Holm correction. A curated gene-symbol filter applied before highly variable gene selection improved enrichment consistently (scRED vs no-filter + 1.9%, $p = 0.035$).

Conclusions

scRED extracts cell-type-conditional disease signal concentrated in cortical neurons and not detected in the examined glial populations. The negative control shows that roughly 45% of the raw enrichment is reproduced by the matched control — reflecting scoring asymmetry together with between-donor and technical variation — leaving a residual enrichment consistent with disease, a decomposition ordinary overlap statistics do not provide. Graph-based formulations outperformed a non-graph baseline, though not after Holm correction at the primary depth, while node choice did not matter and a simple Wilcoxon rank-sum test remained competitive; we report these results without minimisation.

Background

Autism spectrum disorder (ASD) is a neurodevelopmental condition affecting about 1 in 31 children [1], with a twin-study heritability exceeding 80% [2]. Despite strong genetic evidence, the molecular biology of ASD in the mature cortex remains diffuse: over 1,000 candidate genes have been identified through

whole-exome sequencing [3] and are curated in the SFARI Gene database [4, 5], and their cell-type-specific transcriptomic consequences are only beginning to be mapped.

Single-nucleus RNA sequencing (snRNA-seq) of post-mortem ASD cortex has been the enabling technology. Velmeshev et al. profiled 104,559 nuclei from the prefrontal and anterior cingulate cortex of 15 individuals with ASD and 16 controls and reported convergent dysregulation across cell types [6]. Downstream analysis of such cohorts, however, has been dominated by per-gene differential expression, whether through Scanpy's rank_genes_groups [7], Seurat's FindMarkers [8, 9], or the MAST hurdle model [10]. These approaches assess each gene in isolation and ignore both the modular structure of co-regulated transcripts and the manifold structure separating healthy from diseased cell states.

Graph convolutional networks (GCNs) [11, 12] are a natural alternative, since they learn representations that propagate across structurally related nodes. Several GCN applications to single-cell data place genes at nodes and gene–gene co-expression at edges [13, 14]. This formulation yields interpretable per-gene scores but has a structural limitation in the disease-versus-control setting: a co-expression graph built from healthy cells encodes only normal regulatory relationships and carries no patient-specific signal for message passing to amplify.

In this work we present scRED (single-cell Reconstruction-Error Deviation), a graph-convolutional autoencoder that operates on a cell–cell similarity graph rather than a gene–gene graph. The cell graph is constructed by k-nearest-neighbour search in the principal-component space of the control cells; patient cells are projected through the same control-fitted PCA, so their displacement from the healthy manifold becomes the disease signal. The autoencoder is trained to reconstruct per-cell expression vectors from messages aggregated over neighbours, and genes are ranked by the deviation in reconstruction error between patient and control cells. The pipeline is shown in Fig. 1.

We evaluate scRED against curated autism risk genes across 17 cortical cell types, using a per-context empirical null and a control-versus-control negative control to separate disease signal from the asymmetry inherent to the score. We find that (i) the method recovers SFARI-enriched genes specifically in neuronal populations and not in glia; (ii) the presence of a graph, rather than the choice of cells or genes at its nodes, is the component that improves over a non-graph baseline; and (iii) a Wilcoxon rank-sum test on the same gene universe remains competitive, which we report without minimisation. All code and per-cell-type ranked gene lists are provided.

Methods

Dataset and preprocessing

We applied scRED to the Velmeshev et al. snRNA-seq dataset of ASD cortex [6] (prefrontal and anterior cingulate cortex; 104,559 nuclei from 15 individuals with ASD and 16 controls). The original cell-type annotations were retained, giving 17 cell types: four excitatory layer types (L2/3, L4, L5/6, L5/6-CC), four inhibitory interneuron types (IN-PV, IN-SST, IN-VIP, IN-SV2C), two astrocyte types (AST-FB, AST-PP),

oligodendrocytes, oligodendrocyte precursor cells (OPC), microglia, endothelial cells, and three neurogenic/maturational types (Neu-mat, Neu-NRGN-I, Neu-NRGN-II). A pseudo-bulk context (ALL_CELLS) comprising all cells was also evaluated for the method comparison. UMI counts were normalised to counts per million and log1p-transformed using Scanpy [7]. Because the cohort is sex-imbalanced (10 male, 5 female among cases), XIST and Y-linked transcripts were removed.

Gene-symbol filter

Highly variable gene (HVG) selection on raw output is contaminated by transcript-annotation artifacts, including clone-library identifiers (RP-, CTD-, AC-, AL-), accession-only loci, mitochondrial transcripts, small structural and ribosomal RNAs, and pseudogenes, whose sparse, near-zero expression destabilises any deviation-based score. We implemented a curated symbol filter (gene_filter.py) that removes these classes before HVG selection, with a whitelist for canonical genes that incidentally match a pattern (e.g. TP53, FOXP1) and a per-run audit log. After filtering artifact content in the top-50 rankings fell to near-zero, with occasional clone-library identifiers (e.g. LL22NC03-*) remaining across all cell types.

Highly variable gene selection

For each context we selected 3,000 HVGs on the control cells only after symbol filtering, then applied that fixed gene universe to both control and patient cells so that no patient cell informs gene selection. Genes detected in fewer than 5% of control nuclei were removed before selection, and HVGs were ranked by mean-binned normalised dispersion (the Scanpy/Seurat procedure) rather than by raw variance-to-mean ratio. This step is consequential: a raw variance-to-mean statistic rises as the mean falls and therefore selects the sparsest genes available, producing a gene universe in which nearly every gene is detected in a small minority of nuclei and canonical cell-type markers are excluded. The detection floor and normalised dispersion together yield a universe of well-detected genes in which the markers of each population are retained (Section “Recovered genes”).

Cell-graph construction

A 50-component PCA was fitted on the control matrix, and both control and patient matrices were projected into that PC space. A symmetric k-nearest-neighbour graph (k = 15) was constructed in PC space, with weights $w_{ij} = \exp(-d_{ij} / \mathrm{median}(d))$, where d_{ij} is the Euclidean distance in PC space. Because patient cells are projected through the control-fitted PCA, they are embedded in the control-derived manifold and their displacement from it carries the disease signal. This is the methodological choice that distinguishes the cell-graph formulation from a gene-graph one: the control/patient distinction enters through the graph structure rather than through node identity.

Graph-convolutional autoencoder

The model is a two-layer graph convolutional autoencoder implemented in PyTorch Geometric [15]: $H = \mathrm{ReLU}(\mathrm{GCNConv}^{1}(X, A))$ and $\mathbb{X} = \mathrm{GCNConv}^{2}(\mathrm{Dropout}(H), A)$, where $X \in \mathbb{R}^{n \times g}$ is the cells-by-genes matrix (g = 3,000), A is the weighted symmetric kNN adjacency, and $\mathbb{X}$ is the reconstructed expression. The hidden dimension is 64 and dropout is 0. Training uses Adam [16] with learning rate 1×10^{-3} for 300 epochs on the control graph, with mean-squared-error reconstruction loss. The model is trained on control cells only.

Reconstruction-error deviation

The trained model is applied to both control and patient graphs. For each gene g we compute the mean squared per-cell reconstruction error in control and patient cells and define the reconstruction-error deviation $\mathrm{RED}_g = \boldsymbol{\varepsilon}_g(\mathrm{pat}) - \boldsymbol{\varepsilon}_g(\mathrm{ctrl})$. Genes are ranked by $|\mathrm{RED}^{g}|$. The sign is reported as a separate direction annotation (Up if mean patient expression exceeds mean control expression, Down otherwise), inferred from the log2 fold change on raw log1p(CPM) values to avoid signed reconstruction-error artifacts.

Training, validation and independent evaluation

The autoencoder is trained only on control-donor cells, using the control-fitted PCA basis and the control kNN graph. No patient cell and no SFARI label is visible during training or during construction of the PCA basis and graph. The reconstruction-error deviation is computed by applying the trained control-only model to patient cells, which never enter the training loss. The downstream evaluation metric — enrichment of the ranked list for the independent SFARI Gene reference [4, 5] — is external to the expression data and the training objective. Control and patient partitions are defined at the donor level, and cross-validation and the permutation null operate over donors rather than cells. Hyperparameters (latent dimension 64, k = 15, learning rate 1×10^{-3}, 300 epochs) were fixed a priori from Scanpy defaults and standard autoencoder settings and were not tuned on the evaluation metric. Per-context variability is obtained from five random seeds, and method comparisons use the paired Wilcoxon signed-rank test on per-cell-type seed-mean SFARI percentages. The family of hypotheses is the set of paired method comparisons at each ranking depth, to which Holm correction is applied within depth; the per-context enrichment and control-versus-control tests are pre-specified primary analyses and are reported with exact p-values.

Comparison methods

To isolate the contribution of each component, five alternatives were benchmarked on the identical upstream pipeline (same filtered gene universe, HVGs and cell partition): (i) a gene-node GCN, in which

nodes are genes, edges are gene–gene co-expression (Pearson |r| > 0.3 in controls) and node features are 64-quantile per-gene profiles; (ii) the same gene-node graph with features reduced to mean and standard deviation; (iii) a no-filter variant, as (i) but without the gene-symbol filter; (iv) a multilayer perceptron autoencoder with identical input/hidden/output dimensions (3,000–64–3,000) and no graph; and (v) a Wilcoxon rank-sum test (Scanpy rank_genes_groups, method='wilcoxon') on the same filtered universe, ranked by absolute score. All methods, including scRED, were run within a single pipeline so that every comparison shares one gene universe and one seed scheme. Five seeds were used per method per context.

Validation against curated autism risk genes

Reference set and confidence weighting

Ranked lists were evaluated against the SFARI Gene human module (1,277 curated symbols; 1,148 survive the symbol filter). Two statistics were computed at depths $k \in \{50, 100, 150, 200, 250\}$: the unweighted proportion of the top-k that are SFARI-curated, and a confidence-weighted proportion in which score-1 genes (n = 245) receive weight 1.00, score-2 (n = 710) weight 0.50, and score-3 or syndromic-only genes (n = 322) weight 0.25. The weighting was fixed before results were examined.

Empirical null distribution

Because HVG selection retains well-expressed, well-annotated genes, and curated risk genes are disproportionately of that kind, the SFARI content of the ranking universe substantially exceeds the genome-wide rate of 2.1%. A null was therefore measured per context: for a universe of A genes containing B curated genes, 2,000 random draws of k genes were taken without replacement, and enrichment is reported as the ratio of the observed value to the null mean with a two-sided empirical p-value. Measured baselines at k = 250 range from 8.1% to 15.6% across contexts (Table 1).

Control-versus-control negative control

The reconstruction-error deviation compares an error computed on control cells with an error computed on patient cells; because the model is fitted on control cells, these errors are asymmetric and their difference carries a per-gene generalisation term in addition to any disease effect. To quantify that term, the 16 control donors in each context were split into two disjoint halves of eight donors each; the model was fitted on the cells of the first donor group and the cells of the second group were scored exactly as the patient group would be. Because the two halves contain different individuals, this comparison removes disease and donor identity together, so enrichment above the null is attributable to the scoring asymmetry alone. The real contrast was scored against the same first-group model, and both were repeated over five random donor partitions per context (Microglia: 14 donors with sufficient cells, split 7 + 7).

Results

Enrichment relative to a measured null

The expected SFARI overlap under random selection is a property of the ranking universe, not of the genome, and varies almost twofold across cell types (Table 1). Reporting a single constant would misstate enrichment; the per-context null is used throughout.

Table 1. Measured SFARI baselines by cell-type context (k = 250, universe size 3,000 genes). Unweighted and confidence-weighted baselines are the mean overlap of 2,000 random draws.

Context	Curated	Unweighted %	Weighted %
ALL_CELLS	242	8.07	4.19
L5/6-CC	271	9.03	4.57
L5/6	277	9.23	4.84
IN-SV2C	278	9.27	4.88
L2/3	278	9.27	5.07
IN-PV	286	9.53	5.22
IN-VIP	288	9.60	4.99
L4	299	9.97	5.26
IN-SST	317	10.57	5.43
AST-PP	353	11.77	6.32
Endothelial	357	11.90	6.47
OPC	372	12.40	6.83
Neu-mat	392	13.07	7.24
Microglia	406	13.53	7.79
Neu-NRGN-I	426	14.20	8.00
Oligodendrocytes	447	14.90	8.28
AST-FB	454	15.13	8.53
Neu-NRGN-II	468	15.60	8.88

Disease-associated enrichment is separable from the scoring asymmetry and is concentrated in neurons

Table 2 reports the control-versus-control negative control alongside the real disease contrast, as fold enrichment over the per-context baseline, averaged over five random donor-disjoint partitions of the control cohort. Across all 17 cell types the disease contrast exceeded its matched negative control (Wilcoxon signed-rank, n = 17, p = 0.0061). The disease contrast and matched negative control across cell types are shown in Fig. 2. The summary by cell-class group is shown in Table 3. The effect was concentrated in neuronal populations, where the real contrast reached a mean of 1.98-fold against a negative-control floor of 1.44-fold and exceeded the control in all 9 contexts (p = 0.0039). Glial and endothelial populations showed no separation as a group (1.16-fold versus 1.18-fold, p = 1.0), and the two NRGN neuronal subtypes behaved like the glial group; with only two subtypes, the NRGN class is reported descriptively and no group-level significance test is computed for it.

Table 2. Control-versus-control negative control and real disease contrast (k = 250, five random donor-disjoint partitions per context), as fold enrichment over the per-context baseline. Values are mean ± SD across the five partitions.

Context	Negative control (mean ± SD)	Real contrast (mean ± SD)	Difference	Group
L5/6-CC	1.27 ± 0.15	2.41 ± 0.18	+1.15	Neuronal
IN-SV2C	1.55 ± 0.38	2.38 ± 0.17	+0.83	Neuronal
IN-PV	1.15 ± 0.26	1.82 ± 0.22	+0.67	Neuronal
L5/6	1.49 ± 0.56	2.11 ± 0.22	+0.62	Neuronal
L4	1.73 ± 0.73	2.22 ± 0.26	+0.49	Neuronal
IN-SST	1.57 ± 0.37	1.92 ± 0.09	+0.35	Neuronal
IN-VIP	1.33 ± 0.56	1.65 ± 0.12	+0.32	Neuronal
L2/3	1.14 ± 0.22	1.41 ± 0.26	+0.27	Neuronal
OPC	1.04 ± 0.38	1.30 ± 0.28	+0.26	Glial
Neu-mat	1.70 ± 0.43	1.90 ± 0.28	+0.20	Neuronal
AST-FB	1.05 ± 0.12	1.10 ± 0.19	+0.05	Glial
Neu-NRGN-I	1.22 ± 0.28	1.25 ± 0.27	+0.03	NRGN
AST-PP	1.22 ± 0.36	1.23 ± 0.10	+0.01	Glial
Microglia	1.19 ± 0.25	1.19 ± 0.08	+0.00	Glial
Neu-NRGN-II	0.97 ± 0.33	0.95 ± 0.17	−0.02	NRGN
Oligodendrocytes	1.21 ± 0.18	1.02 ± 0.18	−0.19	Glial
Endothelial	1.37 ± 0.10	1.11 ± 0.23	−0.26	Glial

Table 3. Summary of Table 2 by cell-class group (Wilcoxon signed-rank, real contrast versus negative control).

Group	n	Neg. control	Real	Favouring real	p
Neuronal	9	1.44	1.98	9 of 9	0.0039
Glial and endothelial	6	1.18	1.16	3 of 6	1.0
NRGN neurons	2	1.09	1.10	1 of 2	—
All contexts	17	1.31	1.59	13 of 17	0.0061

The negative control was non-trivial in every neuronal context (1.14- to 1.73-fold). Approximately 45% of the raw enrichment observed in neuronal populations is therefore reproduced by the matched negative control — reflecting the scoring asymmetry between in-sample control error and out-of-sample patient error together with between-donor and technical variation — leaving a residual enrichment consistent with a disease-associated component; reporting raw enrichment alone would overstate that disease-associated component by the reproduced margin. Critically, the disease-specific neuronal gap was 0.54-fold and was essentially unchanged (0.54–0.55) across three variants — splitting control cells, splitting control donors, and additionally restricting gene selection to control cells only — so neither donor and batch differences between individuals nor the choice of gene-selection universe inflates the disease signal. Every neuronal context, including the most abundant (L2/3), exceeded its negative control, and per-partition standard deviations are reported in Table 2.

The graph contributes at shallower depths, but node choice does not

Paired comparisons at the primary depth (k = 250, n = 17 cell types, donor-level analysis) are summarised in Table 4. The cell-graph and gene-graph formulations were statistically indistinguishable (mean difference −0.2%, p = 0.82), so node identity is not the operative variable. Both graph formulations exceeded the non-graph multilayer perceptron (cell-graph versus MLP +3.1%, p = 0.031), and at shallower depths this advantage was larger and survived Holm correction (for example +6.9%, p_holm = 0.038 at k = 100). The gene-symbol filter improved enrichment consistently (cell-graph versus no-filter +1.9%, p = 0.035; the gene-graph showed the same effect at p_holm = 0.008). A Wilcoxon rank-sum test on the same universe was competitive and marginally ahead at the primary depth (−2.2%, p = 0.05). The method comparisons across ranking depths and cell types are shown in Fig. 3.

Table 4. Paired comparisons of scRED against each alternative at k = 250 (Wilcoxon signed-rank, n = 17 cell types; meanD = mean difference in SFARI overlap, positive favouring scRED; r = rank-biserial effect size).

Comparison	meanD	r	p	Holm p
scRED vs gene-graph	−0.2%	−0.07	0.82	1.00
scRED vs mean/std pool	+0.9%	+0.10	0.75	1.00
scRED vs no-filter	+1.9%	+0.58	0.035	0.32
scRED vs MLP (no graph)	+3.1%	+0.59	0.031	0.31
scRED vs Wilcoxon	−2.2%	−0.55	0.05	0.37

Recovered genes

The top-ranked genes across four representative neuronal contexts are shown in Fig. 4. Interpretation of individual genes is drawn from eight of the nine neuronal contexts (all but L2/3, which under the donor-disjoint control now also exceeds its negative control, +0.27). Across these eight, 25 of 120 top-15 positions are SFARI-curated, against a mean baseline of 10.0%, and the recovered genes are dominated by synaptic adhesion molecules, axon guidance receptors and ion channels: GRID2, CSMD1, PCDH9, KIRREL3 and RBFOX1 in IN-VIP; PTPRD, PLCB1, CSMD1, SYT1 and KCND2 in Neu-mat; ROBO2, ZMYM3, KIRREL3 and RORA in IN-PV; PTPRT, ROBO2 and ZNF385B in L5/6. KIRREL3 recurs in IN-PV and IN-VIP, ROBO2 in IN-PV and L5/6, and CSMD1 in IN-VIP and Neu-mat. An immediate-early transcriptional program is prominent in excitatory populations, with NPAS4 the top-ranked gene in L5/6 and L5/6-CC and FOS, FOSB and GADD45G within the top 15. Cell-type-defining genes are recovered in their own populations — SST ranks first in IN-SST and SV2C fifth in Neu-mat — confirming that the corrected HVG selection retains and prioritises markers rather than discarding them for sparsely detected transcripts.

Three features of the rankings are unlikely to reflect disease biology and are reported rather than omitted. MALAT1, a nuclear-retained lncRNA whose apparent abundance is sensitive to nuclear isolation and sequencing depth, ranks at or near the top of IN-SV2C, IN-VIP and Neu-mat. A cluster of heat-shock and chaperone transcripts (HSPA1A, HSPB1, DNAJB1) leads IN-VIP, which is also among the neuronal contexts with the smallest margin over its negative control. Ribosomal and housekeeping transcripts appear across several contexts, consistent with the residual expression-magnitude sensitivity quantified by the negative control. Individual top-ranked genes should therefore not be treated as candidate autism genes without independent support.

Astrocytes, oligodendrocytes, microglia, endothelial cells and the NRGN subtypes showed no separation from the negative control as a group. OPC was the one non-neuronal population with a positive individual margin (1.30 versus 1.04), but the glial–endothelial group did not separate overall (its mean sat at or below the negative control, p = 1.0) and OPC is not interpreted further. Ranked lists for these populations are provided for completeness but no biological claim is made from them.

Discussion

This study yields three findings for snRNA-seq analysis in complex brain disorders, together with a methodological caution.

First, the disease signal recovered by a control-trained reconstruction model is cell-type-conditional and concentrated in neurons. Against a matched control-versus-control negative control — a comparison in which no disease exists — the method enriched for curated risk genes in all 9 neuronal populations and in no glial population as a group (3 of 6 positive individually, but not significant, $p = 1.0$). Because SFARI is dominated by synaptic and transcriptional-regulation genes expressed in neurons, this localisation is the pattern a method sensitive to genuine neuronal dysregulation should produce. The negative control also shows that a substantial fraction of the raw enrichment — roughly 45% in neuronal contexts — is reproduced by the matched negative control, reflecting the train/test asymmetry of the score together with between-donor and technical variation rather than disease; to our knowledge this decomposition is not routinely reported in disease-gene prioritisation from snRNA-seq, and we recommend it.

Second, the presence of a graph, not the choice of entity at its nodes, accounts for the higher enrichment over a non-graph baseline at shallower ranking depths. Both cell-graph and gene-graph formulations showed higher enrichment than a same-shape multilayer perceptron — significantly so at shallower depths, though not after Holm correction at the primary depth — but they did not differ from each other. This corrects a natural expectation — that placing cells rather than genes at the nodes would be decisive — and indicates that the benefit of the graph in this setting is more robust to node choice than anticipated.

Third, a Wilcoxon rank-sum test on the same filtered gene universe is competitive with, and at the primary depth marginally ahead of, the graph methods. This is consistent with single-cell benchmark studies in which simple baselines match deep methods on annotation, clustering and trajectory inference [17–19]. We report it without minimisation: the practical value of the graph model is confined to the neuronal contexts where the cell-similarity manifold is structured, and elsewhere a standard differential-expression test is an adequate and cheaper choice.

The methodological caution concerns preprocessing. Highly variable gene selection by a raw variance-to-mean statistic selects the sparsest genes in the matrix; in per-cell-type universes this excludes canonical markers and fills the ranking universe with genes detected in a small minority of nuclei, on which a deviation-based score is dominated by sampling noise. Substituting a detection floor and mean-binned normalised dispersion removes this failure mode. This choice is upstream of, and independent from, the scoring method, and we note it because it materially affects any deviation-based ranking.

Several limitations apply. The benchmark uses a single cohort, and the relative ordering of methods may differ on other neuropsychiatric snRNA-seq datasets; validation against expression-derived signatures from PsychENCODE [20] or the original cohort differential-expression lists would broaden the assessment. SFARI overlap, derived from rare-variant exome studies, is skewed toward developmental loss-of-function genes and is enriched in the gene universe itself, which is why enrichment is reported against a per-context null. Hyperparameters were fixed a priori and not tuned. Because scRED fits its

parameters on control cells only, the trained model is reusable on new patient sets without refitting, a property a differential test does not share and which we did not evaluate across cohorts here.

Conclusions

scRED is a control-trained, graph-aware alternative to per-gene differential expression for single-nucleus transcriptomics of complex brain disorders. By placing cells at the graph nodes and training on healthy cells only, it recovers SFARI-enriched genes specifically in cortical neurons, at an enrichment that exceeds a matched negative control in all 9 neuronal contexts. Graph-based formulations show higher enrichment than a non-graph baseline at shallower ranking depths, though not after Holm correction at the primary depth, while node choice does not matter, and a Wilcoxon rank-sum test remains competitive. We provide the code, the gene-symbol filter and the per-cell-type ranked gene lists for the community.

Declarations

Ethics approval and consent to participate

Not applicable. This study used publicly available, de-identified single-nucleus RNA sequencing data from a previously published cohort.

Consent for publication

Not applicable.

Availability of data and materials

The snRNA-seq data analysed in this study were generated by Velmeshev et al. [6] and are available through the UCSC Cell Browser. The SFARI Gene reference is available at https://gene.sfari.org. Source code, the gene-symbol filter, the analysis scripts and the per-cell-type ranked gene lists are freely available at https://github.com/NorahAwn/scRED and archived at Zenodo (DOI **10.5281/zenodo.23010003**) under the MIT license.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Shenzhen Science and Technology Program [grant numbers JCYJ20241202130212016 and KQTD20200820113106007 to J.T., and JCYJ20230807140709020 to M.Z.].

Authors' contributions

NSA conceived the study, implemented the method and analyses, and drafted the manuscript. MZ contributed to the analysis and interpretation. MSMB contributed to validation and to review and editing of the manuscript. ASB contributed to the analysis. JT supervised the work and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank the Velmeshev et al. group for making their snRNA-seq cortex dataset publicly available, and the SFARI Gene curators for maintaining the ASD-risk gene reference.

References

1. Shaw KA, Williams S, Patrick ME. Prevalence and early identification of autism spectrum disorder among children aged 4 and 8 years — Autism and Developmental Disabilities Monitoring Network, 16 sites, United States, 2022. MMWR Surveillance Summaries. 2025;74(2):1–22.
2. Bai D, Yip BHK, Windham GC. Association of genetic and environmental factors with autism in a 5-country cohort. JAMA Psychiatry. 2019;76(10):1035–43.
3. Satterstrom FK, Kosmicki JA, Wang J. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. Cell. 2020;180(3):568–84.
4. Abrahams BS, Arking DE, Campbell DB. *SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders.* Molecular Autism, 2013. 4(1): p. 36.
5. Banerjee-Basu S, Packer A. SFARI Gene: an evolving database for the autism research community. Dis Models Mech. 2010;3(3–4):133–5.
6. Velmeshev D, Schirmer L, Jung D. Single-cell genomics identifies cell type-specific molecular changes in autism. Science. 2019;364(6441):685–9.
7. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15.
8. Hao Y, Hao S, Andersen-Nissen E. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573–87.
9. Stuart T, Butler A, Hoffman P. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–902.
10. Finak G, McDavid A, Yajima M. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16:278.
11. Kipf TN, Welling M. *Semi-supervised classification with graph convolutional networks*, in *International Conference on Learning Representations (ICLR)*. 2017.
12. Kipf TN, Welling M. Variational graph auto-encoders. in NeurIPS Bayesian Deep Learn Workshop. 2016.

13. Ciortan M, Defrance M. GNN-based embedding for clustering scRNA-seq data. Bioinformatics. 2022;38(4):1037–44.
14. Wang J, Ma A, Chang Y. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. Nat Commun. 2021;12:1882.
15. Fey M, Lenssen JE. *Fast graph representation learning with PyTorch Geometric*, in *ICLR Workshop on Representation Learning on Graphs and Manifolds*. 2019.
16. Kingma DP, Ba J. *Adam: a method for stochastic optimization*, in *International Conference on Learning Representations (ICLR)*. 2015.
17. Abdelaal T, Michielsen L, Cats D. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol. 2019;20:194.
18. Duò A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Research, 2018. 7: p. 1141.
19. Saelens W, et al. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019;37(5):547–54.
20. Gandal MJ, Zhang P, Hadjimichael E. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science. 2018;362(6420):eaat8127.

Figures

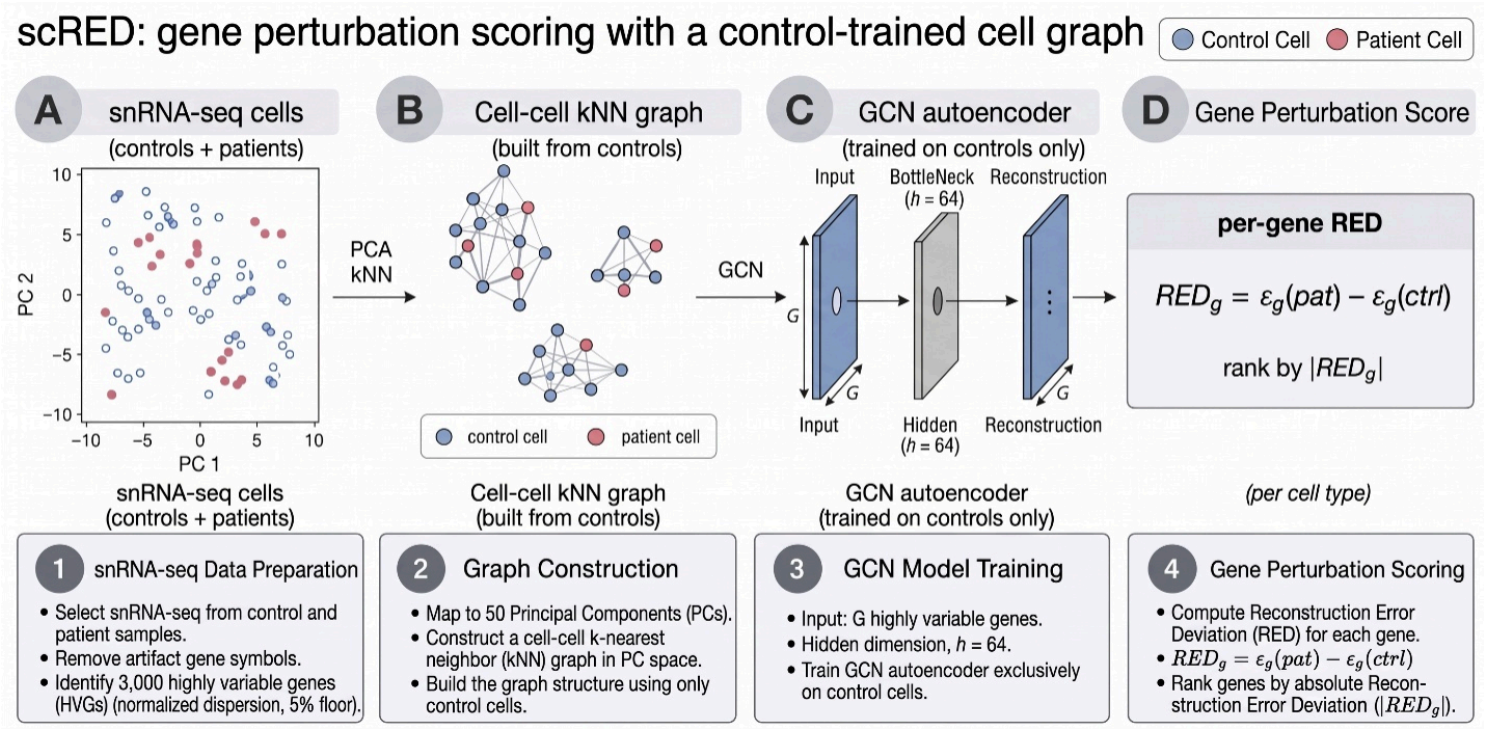


Figure 1

The scRED pipeline. (A) snRNA-seq cells are symbol-filtered and HVG-selected with a 5% detection floor and normalised dispersion. (B) A k-nearest-neighbour graph is built in PCA space, fitted on control cells;

patient cells are projected through the control-fitted PCA. (C) A two-layer graph convolutional autoencoder is trained to reconstruct expression from control cells. (D) Per-gene reconstruction-error deviation, RED = patient error − control error, ranks genes by |RED|.

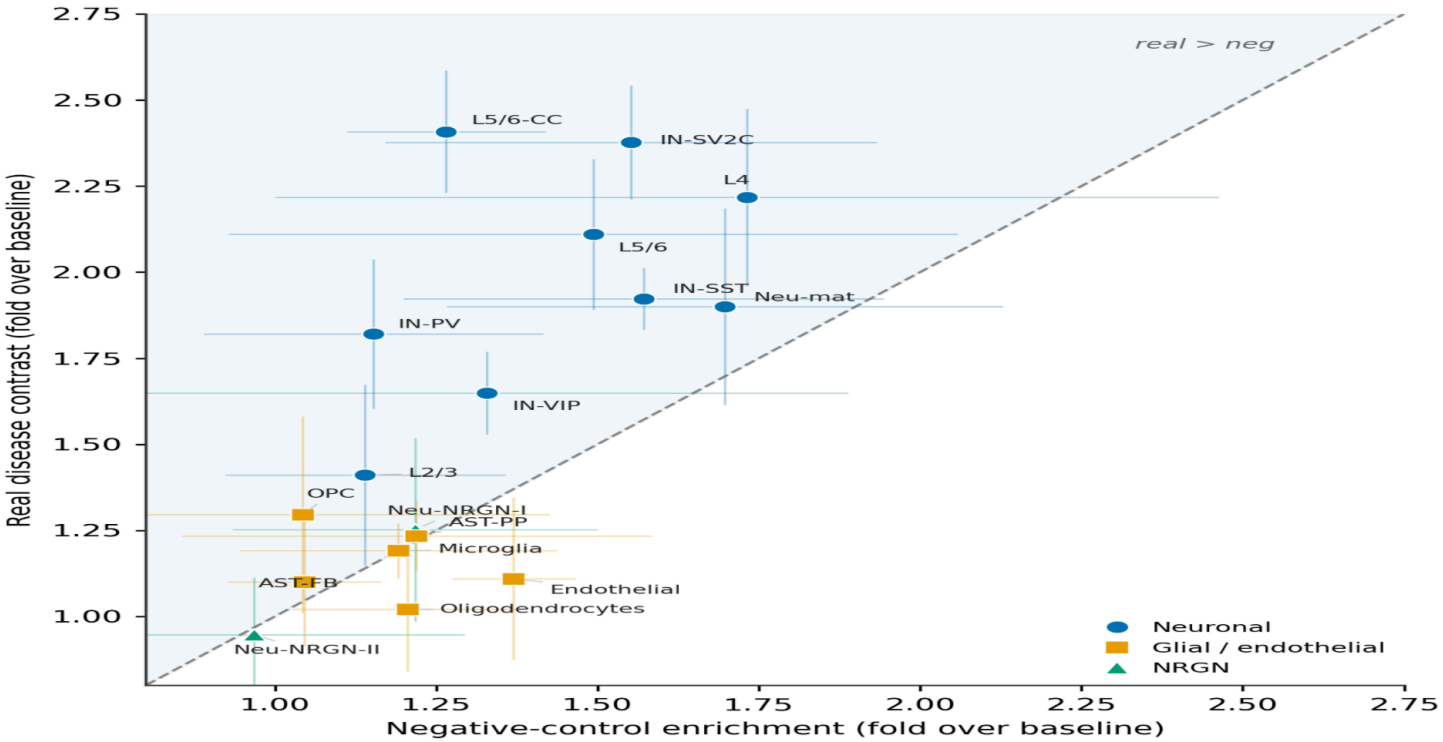


Figure 2

Enrichment against the negative control. For each cell type, real disease contrast (fold over baseline) is plotted against the matched control-versus-control negative control; neuronal populations lie above the identity line, and glial and NRGN populations on or below it, with OPC the single non-neuronal population above the line.

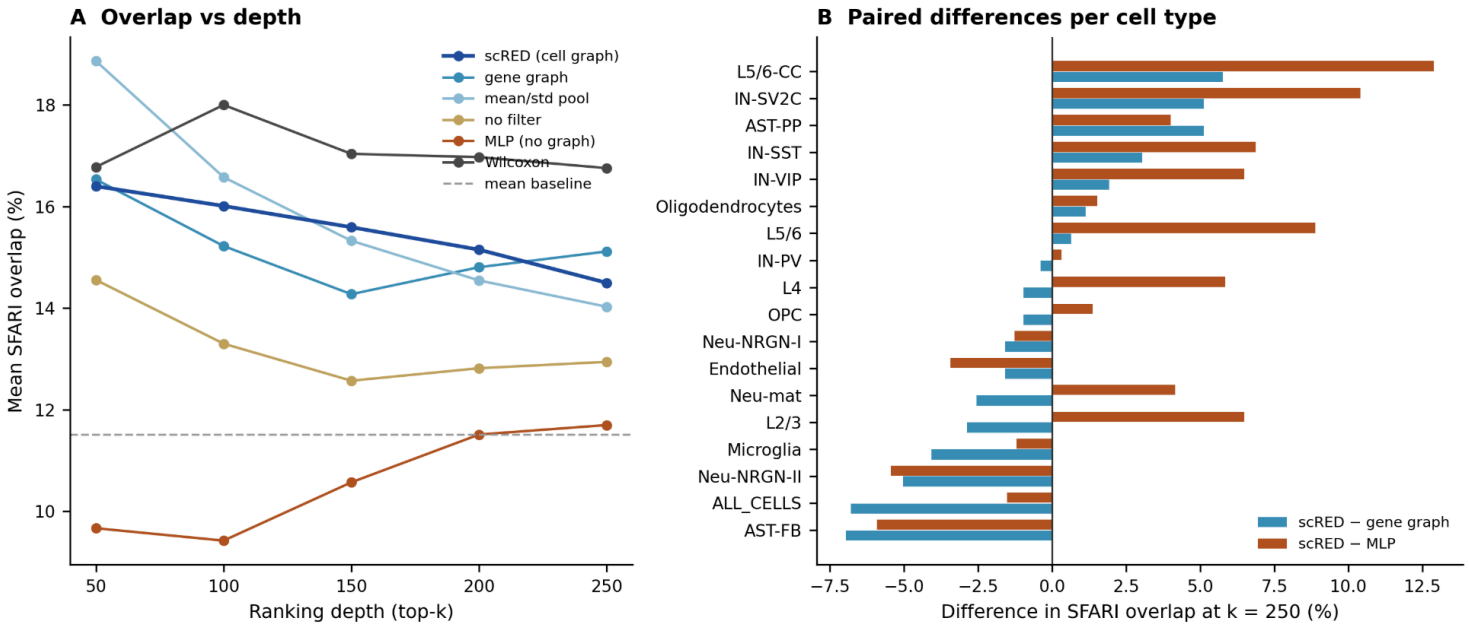


Figure 3

Method comparison. (A) Mean SFARI overlap versus ranking depth per method, with the mean measured baseline. (B) Per-cell-type paired difference at k = 250 between scRED and the gene-graph variant and between scRED and the MLP.

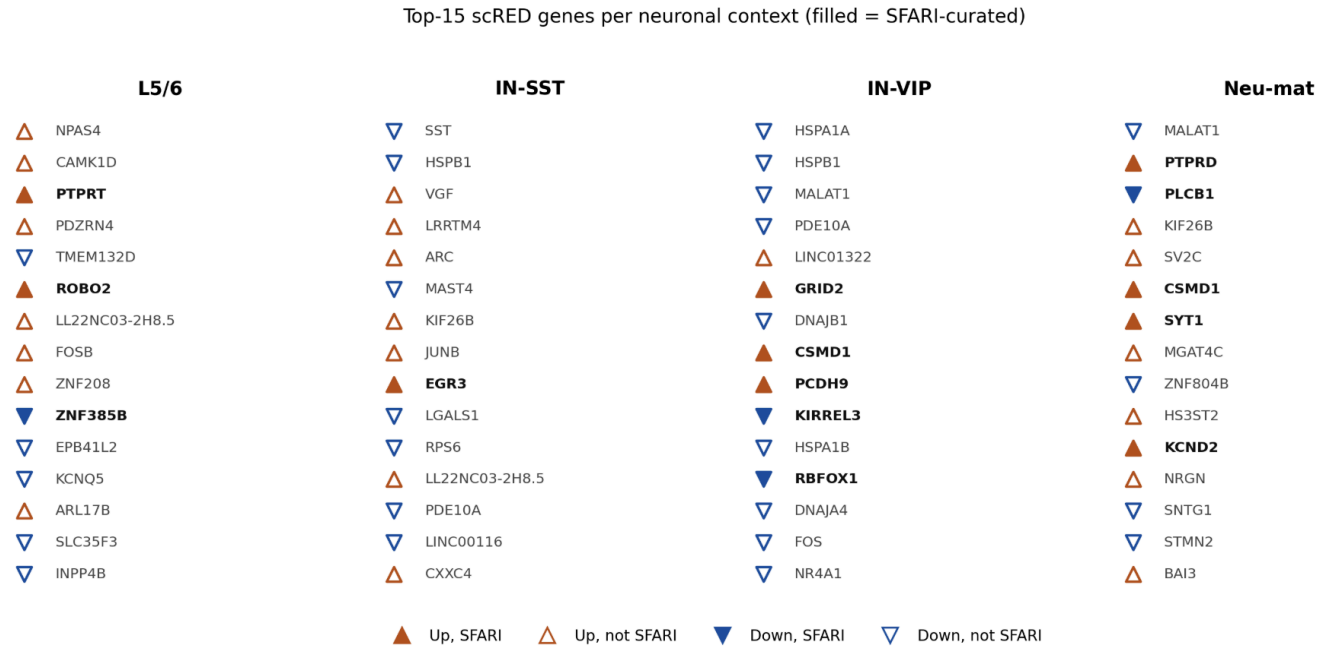


Figure 4

Recovered genes in four neuronal contexts. The top-15 scRED genes per context, with direction of change (up/down) and SFARI status (filled = curated).