HIERARCHICAL DISEASE-STATE GENERATORS FOR NEURODEGENERATIVE GENOMICS: A BENCHMARK PROPOSAL FOR INTERVENTION-CONDITIONED MULTI-OMIC GENERATION

David Scott Lewis & Enrique Zueco
AIXC Research
reports@aiexecutiveconsulting.com

ABSTRACT

We present a *benchmark proposal and evaluation framework* for the disease-state generator task: intervention-conditioned generation of transcriptomic and epigenomic cell states, evaluated through mechanism-grounded acceptance criteria rather than generic sample-quality scores. Targeting neurodegeneration (AD/PD) as a biologically demanding test bed, we define (i) a formal task specification for conditional generation under drugs, CRISPR perturbations, and regulatory edits; (ii) an architecture blueprint—multimodal latent encoders coupled to conditional diffusion with hierarchical regulatory priors (enhancer→TF→gene); and (iii) a barrier-and-frontier evaluation suite testing hierarchy fidelity, perturbation prediction, cross-context generalization, and uncertainty-calibrated intervention ranking. The framework also serves as an *evaluation surface* for DNA foundation models, measuring whether sequence-derived priors improve intervention-conditioned generation. We report proof-of-concept experiments on the Norman 2019 CRISPRa dataset that validate the evaluation protocols, while identifying a key bottleneck—gene-regulatory-network sparsity—that must be resolved before hierarchy-fidelity testing is meaningful. This is a benchmark and evaluation contribution; the architecture is a proposed blueprint, not a fully validated system.

1 INTRODUCTION

Generative AI has transformed protein engineering, but the analogous goal in genomics—engineering cellular and tissue states—is still emerging. Neurodegenerative disorders such as Alzheimer's disease (AD) and Parkinson's disease (PD) are a natural proving ground because the objects of interest are *cell states* and *state transitions*: microglia activation trajectories, reactive astrocyte programs, and stress-linked neuronal subtypes. These programs are now observable at scale through single-nucleus RNA sequencing and multi-omic assays.

However, two gaps limit biological impact: (1) **Mechanism mismatch**: generative models can fit distributions but violate regulatory hierarchies that are core to genomics (e.g., enhancer→TF→target relationships). (2) **Evaluation mismatch**: generic sample-quality scores do not answer whether a model predicts *perturbation outcomes* or supports *intervention selection*.

This paper proposes a benchmark direction aligned with Gen2 workshop topics in single-cell omics (trajectory simulation; perturbation effect modeling; virtual cell models), regulatory genomics (sequence-to-function; regulatory element design), and evaluation (biologically grounded metrics). Our contributions are:

- **Task definition:** a disease-state generator formalism—a disease-specific *virtual cell model*—for intervention-conditioned generation of multi-omic cell states in AD/PD contexts.
- **Architecture blueprint:** conditional latent diffusion over multimodal latent states, constrained by hierarchical regulatory priors and paired with uncertainty estimation.

- **Barrier-and-frontier evaluation:** metrics that test regulatory hierarchy fidelity, perturbation accuracy, cross-context generalization, and uncertainty-calibrated intervention ranking.
- **Benchmark design:** concrete dataset ingredients and train/test splits emphasizing held-out interventions and contexts relevant to translational neurogenomics.

A key differentiator from prior perturbation-response models is that hierarchy fidelity and calibrated abstention are *first-class evaluation targets*, not post-hoc diagnostics. Where existing methods ask "how accurately can we predict?", our framework additionally asks "*when should we trust* a prediction?"—a question motivated by the finding that even linear baselines remain competitive on standard perturbation benchmarks (Ahlmann-Eltze et al., 2025), suggesting that raw accuracy alone is an insufficient evaluation axis.

Scope and maturity. This paper is a *benchmark proposal*, not a completed empirical study. The task formalization, evaluation suite, and benchmark design are fully specified; the architecture blueprint (Section 4) is a concrete but *proposed* design, not a trained system. Preliminary validation (Section 8) demonstrates that the evaluation protocols are implementable on real perturbation data, while surfacing a GRN-sparsity bottleneck that limits hierarchy-fidelity testing in its current form. We present this as a community resource for structuring future work on mechanism-aware generative genomics.

2 Background: why AD/PD is a hard but useful target

AD and PD involve coordinated programs across microglia, astrocytes, oligodendrocytes, vascular cells, and vulnerable neuronal populations. Single-cell studies have identified disease-associated microglia (DAM), neurotoxic reactive astrocytes, and PD-specific neuronal states with selective dopaminergic vulnerability. Large human atlases reveal substantial disease-stage and donor-dependent heterogeneity.

From a modeling perspective, AD/PD introduces three "stressors" for GenAI-in-genomics:

1. **Context shifts:** donor, brain region, sex, ancestry, and disease stage produce distribution shifts that a generator should generalize across.
2. **Compositional interventions:** meaningful perturbations are multi-gene, pathway-level, and dose/time dependent.
3. **Hierarchical mechanisms:** regulatory elements shape TF activity which shapes gene programs and trajectories; models should respect this causal layering.

3 Problem statement: the disease-state generator task

We formalize *state engineering* as intervention-conditioned generation in single-cell genomics.

Observed state and context. Let $x \in \mathbb{R}^G$ be a gene-expression vector (counts or normalized), a optional additional modalities (e.g., ATAC peaks, protein counts), and c context (cell type, brain region, donor covariates, disease stage).

Interventions. Let u denote an intervention: a small molecule (identity, dose, time), a CRISPR perturbation (CRISPRi/a targets), or a regulatory edit (sequence/element design).

Learning objective. Given paired or partially paired data with pre/post measurements, the target is a conditional distribution

$$p(x', a' \mid x, a, c, u), \tag{1}$$

where (x', a') is the post-intervention state (counterfactual in the causal sense; see Section 4.3 for assumptions).

Design desiderata. A useful disease-state generator must support:

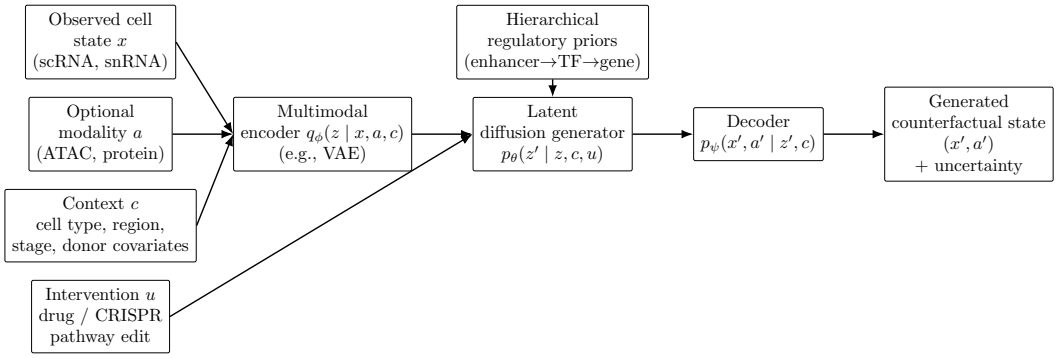

Figure 1: Disease-state generator blueprint. A multimodal encoder maps observed measurements to a latent state; a conditional generator produces counterfactual latent states under interventions; hierarchical regulatory priors guide generation toward mechanistically consistent trajectories.

- **Realism and identity preservation:** samples should resemble real post-intervention cells without collapsing cell identity.
- **Mechanistic consistency:** generated changes should be consistent with regulatory hierarchy constraints (Section 4.3).
- **Intervention generalization:** accurately predict held-out interventions and multi-perturbation compositions.
- **Cross-context generalization:** transfer across donors/regions/stages with calibrated uncertainty.

4 Model blueprint: conditional latent diffusion with hierarchical priors

4.1 Multimodal latent representation

We use a multimodal encoder $q_\phi(z \mid x, a, c)$ and decoder $p_\psi(x, a \mid z, c)$, building on established single-cell VAEs (Lopez et al., 2018; Gayoso et al., 2021). Latent representations reduce dimensionality, denoise sparse counts, and provide a continuous space for generative sampling. Disentangled latent factors—cell type, disease stage, intervention response—can be linked to known gene programs, enabling post-hoc interrogation of what the generator has learned.

4.2 Conditional diffusion in latent space

We generate intervention-conditioned latent states z' via a conditional diffusion model $p_\theta(z' \mid z, c, u)$ (Ho et al., 2020; Song et al., 2021). Latent diffusion avoids discrete count modeling, supports rich conditioning, and enables stochastic sampling for uncertainty-aware generation.

Conditioning and compositionality. The intervention space $\mathcal{U} = \mathcal{U}_{\text{gene}} \times \mathcal{U}_{\text{chem}} \times \mathcal{U}_{\text{dose}} \times \mathcal{U}_{\text{time}}$ reflects combinatorial biology. We encode interventions via an embedding $e(u)$ supporting combinations through set encoders or additive composition. The hierarchy prior (Section 4.3) constrains compositional generalization by requiring predicted shifts to respect pathway structure rather than unconstrained interpolation.

4.3 Hierarchical regulatory priors

We regularize generation using a directed, signed regulatory graph $\mathcal{G}$ with nodes for enhancers, TFs, and genes, and edges for activation/repression.

Where does $\mathcal{G}$ come from? $\mathcal{G}$ can be inferred from multi-omics (enhancer–gene links, motif evidence, TF activity) or from DNA foundation models—Enformer (Avsec et al., 2021), Nucleotide Transformer (Dalla-Torre et al., 2025), DNABERT (Ji et al., 2021)—which provide sequence-derived

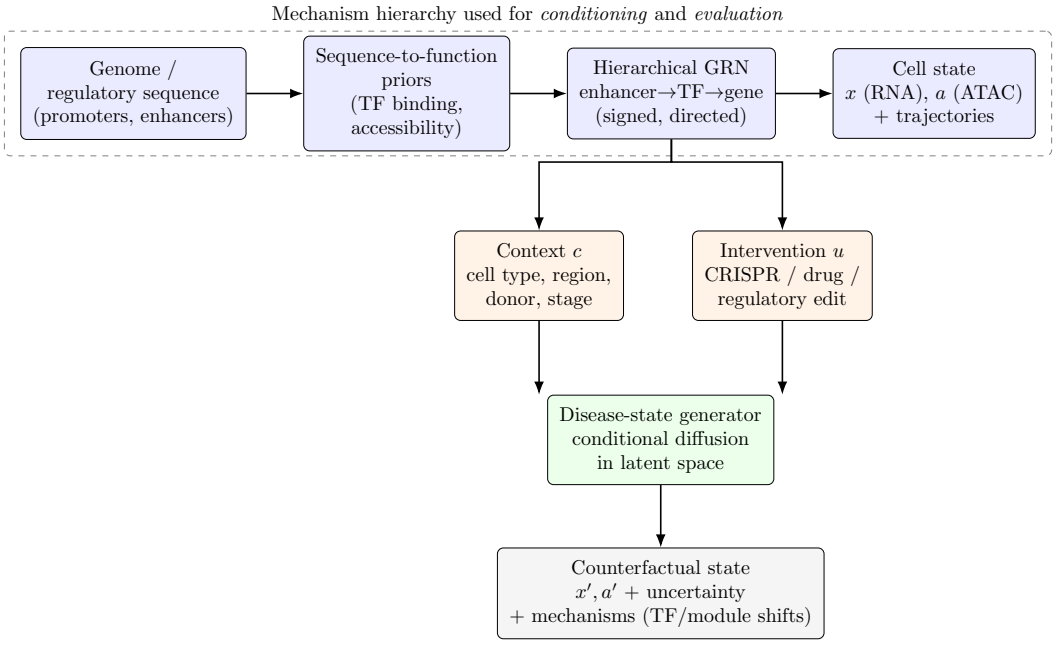

Figure 2: Multi-scale hierarchy used for *conditioning* and *evaluation*. Sequence-to-function priors inform enhancer/TF activity; a signed GRN provides mechanistic constraints; the generator produces counterfactual states under intervention and context.

edge weights and enhancer-activity predictions. The disease-state generator thus serves as an *evaluation surface* for these FMs: hierarchy fidelity and perturbation accuracy (Section 5) measure whether FM-derived priors improve generation over motif-only baselines.

A simple hierarchy-consistency loss. Let $\Delta\hat{t}_i$ be a predicted TF-activity shift (from expression and/or chromatin), and $\Delta\hat{x}_j$ the gene-expression shift for target gene j. For each signed edge $(i \to j)$ with sign $s_{ij} \in \{\pm 1\}$ and weight w_{ij}:

$$\mathcal{L}_{\text{hier}} = \sum_{(i\to j)\in\mathcal{G}} w_{ij}\ \max\big(0, -s_{ij}\,\Delta\hat{t}_i\,\Delta\hat{x}_j\big). \tag{2}$$

This penalizes "wrong-direction" target responses given TF shifts, turning mechanistic faithfulness into a measurable *barrier*.

Causal framing and assumptions. We distinguish *intervention-conditioned prediction* from *causal identification*. Under potential outcomes (Rubin, 1974), valid counterfactual inference requires consistency (SUTVA), conditional exchangeability, and positivity; under structural causal models (Pearl, 2009), a correct DAG. These assumptions are *not fully testable* in observational atlases where batch effects, latent confounders, and selection bias are pervasive. The hierarchy prior partially addresses identifiability by restricting predictions to trajectories consistent with $\mathcal{G}$, but this is a soft constraint, not a causal guarantee. Outputs should be interpreted as *mechanism-informed conditional predictions*, and uncertainty estimates should reflect this epistemic limitation.

4.4 Uncertainty and specification-gated generation

Wrong confident predictions are worse than abstention. We pair stochastic generation with uncertainty estimation (ensembles; dropout; diffusion variance) and *calibrate* predictive intervals via conformal prediction (Vovk et al., 2005). A *specification-gated* (spec-gated) function $S(x', \mathcal{G}, u) \in \{0, 1\}$ returns 1 iff the prediction x' satisfies: (a) hierarchy-consistency above τ_h, (b) interval width below τ_w, and (c) conformal coverage $\geq 1-\alpha$. Only predictions with $S = 1$ are released; otherwise the system abstains. This coverage–accuracy tradeoff is absent from GEARS (Roohani et al., 2024) and

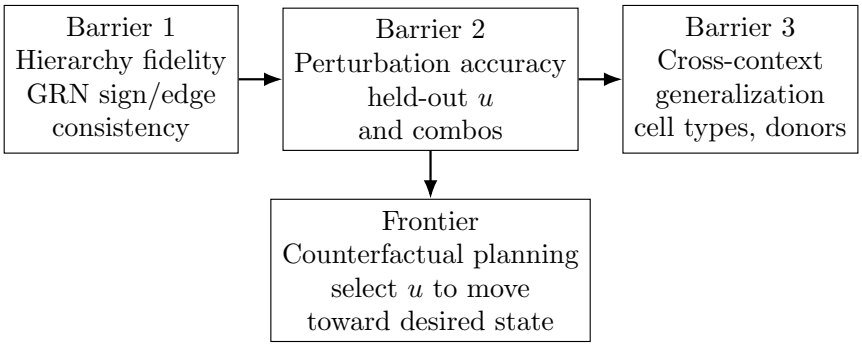

Figure 3: Barrier-and-frontier evaluation for disease-state generators. Barriers test biological consistency and generalization. The frontier asks whether models support intervention planning under uncertainty.

CPA (Lotfollahi et al., 2023), which lack formal acceptance criteria. We recommend reporting the full coverage–abstention curve rather than a single operating point, tuning τ_h, τ_w, α on a held-out calibration set.

5 Mechanism-grounded evaluation: barriers and frontiers

A core workshop theme is evaluation aligned with biology rather than proxy scores. We propose three barrier tests and one frontier test. Full experimental protocols are in the Appendix.

5.1 Barrier 1: regulatory-hierarchy fidelity

Given $\mathcal{G}$, evaluate whether generated shifts respect edge directions and signs. Metrics: (i) sign-consistency rate across TF→target edges; (ii) rank correlation between predicted and observed TF-activity changes; (iii) enhancer–gene coupling preservation in multimodal outputs.

Anti-circularity requirement: when $\mathcal{G}$ also informs training via $\mathcal{L}_{\text{hier}}$, evaluation must use held-out edges not seen during training, or an independently derived GRN (e.g., motif-only evaluation against a multi-omic–trained model). Without this separation, high sign-consistency may reflect prior memorization rather than genuine mechanistic faithfulness.

5.2 Barrier 2: perturbation prediction (held-out u)

Hold out interventions and evaluate both distributional fidelity and effect-size accuracy on differentially expressed genes. Recommended metrics: (i) mean absolute error on log-fold changes for top-K DE genes ($K \in \{20, 50, 100\}$); (ii) module-level agreement (pathway enrichment overlap; regulon activation correlation); and (iii) calibration (empirical coverage of predicted intervals vs. nominal level).

Generic distributional metrics. For sanity checks, two-sample distances between generated and real post-intervention cells (e.g., MMD (Gretton et al., 2012)) and sample-quality proxies (e.g., FID-style embedding scores (Heusel et al., 2017)) should be computed *in learned biological embedding spaces* rather than raw feature space.

5.3 Barrier 3: cross-context generalization (held-out c)

Neurodegeneration requires transfer across donors, regions, stages, ancestry groups, and biological sex. Risk alleles (APOE, TREM2) show population-level frequency variation; a generator trained on one ancestry may miscalibrate for others. Evaluate on held-out contexts from atlas-scale datasets. A practical metric is whether predicted intervention effects preserve ordering along disease trajectories (pseudotime or severity gradients) without identity collapse.

Ingredient	Candidate sources	What it enables
AD/PD state space	SEA-AD (Gabitto et al., 2024); human snRNA-seq atlases (Mathys et al., 2019; Smajić et al., 2022)	Contextual trajectories; conditioning/evaluation targets
Perturbations u	Norman 2019 (Norman et al., 2019); Replogle 2022 (Replogle et al., 2022); sci-Plex (Srivatsan et al., 2020)	Supervision for shifts; held-out u generalization
Regulatory priors	SCENIC+ (Bravo González-Blas et al., 2023); CellOracle (Kamimoto et al., 2023); Enformer (Avsec et al., 2021)	Hierarchy constraints; interpretability; edits

Table 1: Minimal ingredients for a disease-state generator benchmark.

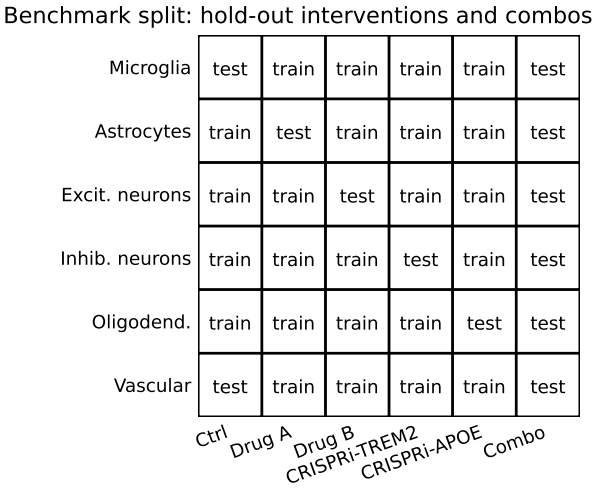

Figure 4: Schematic benchmark split (illustrative). The primary axis is held-out interventions (including combinations) evaluated across multiple cellular contexts.

5.4 Frontier: intervention ranking under uncertainty

Given a desired target state (e.g., shift microglia from DAM-like to homeostatic programs), the frontier question is: *Can a model rank interventions u by expected improvement, with calibrated uncertainty, and with mechanistically interpretable rationales?* This aligns evaluation with actionable biology: selecting perturbations for follow-up rather than merely generating plausible cells. A key diagnostic for this frontier is *ranking stability*: whether intervention orderings change substantially under (a) bootstrap resampling, (b) perturbation of GRN edge weights, or (c) removal of individual hierarchy priors—identifying which ranking decisions are robust and which depend on fragile assumptions.

6 Benchmark design: data ingredients and splits

We propose a benchmark that can be instantiated with any trio: (1) an AD/PD atlas (state space), (2) perturbation screens with single-cell readouts (supervision for u), and (3) regulatory priors (mechanistic constraints).

6.1 Splits: hold out interventions and contexts

Figure 4 illustrates a split strategy designed to stress both compositional intervention modeling and cross-context generalization.

Recommended split families:

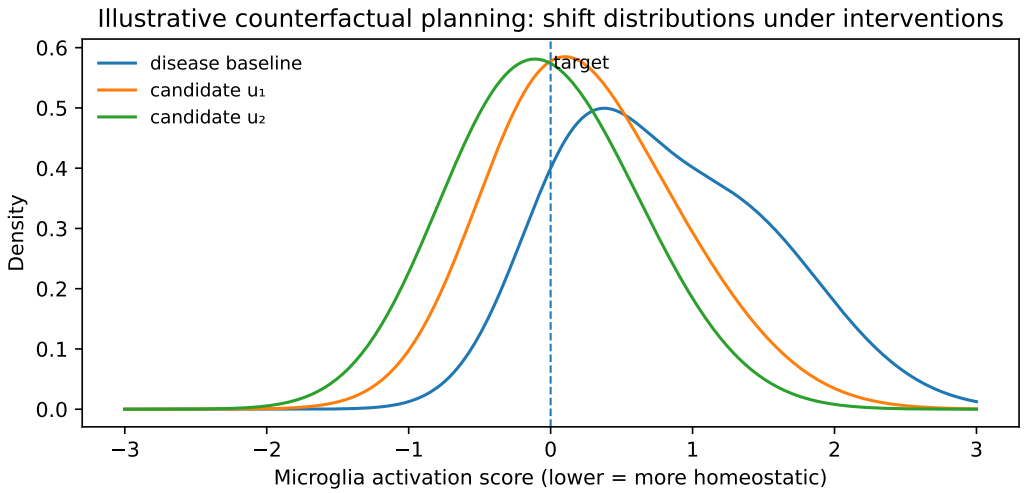

Figure 5: Illustrative counterfactual planning output (toy schematic). A generator ranks interventions by shifting the distribution of a microglia activation score toward a target, while tracking uncertainty and mechanistic constraints.

- **Intervention hold-out:** train on a subset of interventions; test on unseen drugs/targets and unseen combinations.
- **Context hold-out:** train on donors/regions/stages; test on held-out contexts.
- **Joint hold-out:** hold out both u and c to probe transportability and uncertainty calibration.

6.2 Connecting sequence and state engineering

Topic (1) in the prompt emphasizes bridging regulatory sequence/perturbations to cellular states. In the benchmark, this connection appears in two ways: (i) *regulatory editing* as an intervention u (e.g., enhancer design), and (ii) sequence-derived priors that help constrain and interpret state changes. This integration of regulatory genomics with single-cell state modeling makes the disease-state generator task naturally suited to the Gen2 workshop scope.

Spatial context. Spatial transcriptomics adds a critical dimension: microglial activation and amyloid neighborhoods exhibit spatial organization that dissociated data cannot capture. Cell2location (Kleshchevnikov et al., 2022) and Tangram (Biancalani et al., 2021) enable mapping single-cell states onto tissue coordinates, providing spatially resolved contexts for future benchmark extensions.

7 Illustrative case study: microglia-state modulation planning

As a concrete AD/PD translation target, consider planning perturbations that reduce microglial activation while maintaining microglial identity. A disease-state generator can be used as follows:

1. Fit the generator on atlas + perturbation data; infer baseline microglia activation trajectories.
2. Propose candidate interventions u (single or combinatorial).
3. Sample counterfactuals $p(x', a' \mid x, a, c, u)$; compute activation scores and mechanistic summaries (TF/regulon shifts).
4. Rank interventions by expected movement toward a target state, penalizing uncertainty and hierarchy violations.

Why this is useful even without perfect ground truth. Real biological interventions are expensive. A benchmark that measures (i) held-out perturbation accuracy, (ii) hierarchy fidelity, and (iii) uncer-

tainty calibration provides actionable triage: it identifies *where* a generator is trustworthy enough to propose experiments, and *where* it should abstain.

8 PRELIMINARY VALIDATION

To test whether the evaluation framework is implementable on real data, we designed proof-of-concept experiments on the Norman et al. 2019 CRISPRa dataset (Norman et al., 2019) (111,255 cells, 19,018 genes; 105 single and 2-gene CRISPRa perturbations in K562 cells) paired with a K562 gene regulatory network derived from TRRUST v2. While K562 is not an AD/PD system, it provides a well-characterized testbed with real perturbation readouts, combinatorial interventions, and known regulatory relationships—exactly the properties needed to stress-test Barriers 1–3 and the Frontier. We designed four experiments aligned with Barriers 1–3 and the Frontier (E1: hierarchy fidelity, E2: perturbation prediction, E3: cross-context, E4: intervention ranking); implementation status is summarized in Table 2 (Appendix). The key finding is that the TRRUST-derived GRN yielded only 10 TF→target edges after filtering—too sparse for powered hierarchy-fidelity analysis (E1 blocked; E2–E4 protocols ready). This reveals a central lesson: *curated but generic GRN databases are insufficient for hierarchy evaluation in specific cellular contexts*; context-specific inference via SCENIC+ (Bravo González-Blas et al., 2023) or CellOracle (Kamimoto et al., 2023) is the critical next step. Full protocols and per-experiment status are in Appendix A.

9 LIMITATIONS AND OPEN PROBLEMS

As a benchmark proposal, this work identifies several open problems that must be addressed as the framework is instantiated:

- **Counterfactual identifiability:** separating causal effects from confounding and batch effects in observational atlases. Benchmark metrics should be accompanied by sensitivity analyses quantifying how metric values change under simulated confounding or batch-effect perturbations.
- **Mechanism uncertainty:** GRNs are incomplete and context-dependent; hierarchy constraints must be soft and uncertainty-aware.
- **Evaluation alignment:** designing scores that predict downstream experimental utility rather than in-silico similarity alone.
- **Evaluation circularity:** when the same $\mathcal{G}$ informs both training and evaluation, high scores may reflect prior memorization. We recommend three mitigations: (a) edge hold-out splits analogous to intervention hold-outs; (b) evaluation with independently derived GRNs (e.g., motif-only vs. multi-omic–derived); and (c) ablation studies removing $\mathcal{L}_{\text{hier}}$ to measure the marginal value of hierarchy constraints.

10 CONCLUSION

We have presented a benchmark proposal for disease-state generators—a concrete instantiation of "genomics as state engineering" for neurodegeneration. The central idea is to evaluate intervention-conditioned multi-omic generation by mechanism- and task-aligned barriers: regulatory hierarchy fidelity, perturbation prediction, cross-context generalization, and uncertainty-calibrated intervention ranking. Preliminary experiments on the Norman 2019 CRISPRa dataset demonstrate that these evaluation protocols are implementable, while surfacing GRN sparsity as a critical bottleneck requiring context-specific regulatory inference. We expect this framing to clarify research priorities for GenAI in genomics and to provide a practical bridge from generative modeling to actionable AD/PD experimentation.

Software and data. Evaluation protocols, data-loading scripts, and baseline implementations for all four experiments are available in the supplementary repository. The benchmark is designed to be instantiated using open-source single-cell toolkits (`scanpy`, `pertpy`, `scvi-tools`) and publicly available perturbation and atlas datasets; all data sources used in the proof-of-concept are freely accessible (see Appendix A).

REFERENCES

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8):1657–1661, 2025.

Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Hana Imrichova, Gert Hulselmans, Florian Rambow, et al. Scenic: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11): 1083–1086, 2017.

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna- and rna-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, 2015.

Ricard Argelaguet, Anna S E Cuomo, Oliver Stegle, and John C Marioni. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(111), 2020.

Tal Ashuach, Mariano I Gabitto, Michael I Jordan, and Nir Yosef. Multivi: deep generative model for the integration of multimodal data. *Nature Methods*, 20(8):1222–1231, 2023.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, 2021.

Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Rahul Avasthi, Ziqing Lu, Anna Sanger, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nature Methods*, 18 (11):1352–1362, 2021.

Carmen Bravo González-Blas, Seppe De Winter, Gert Hulselmans, et al. Scenic+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature Methods*, 20:1355–1367, 2023.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.

Paul Datlinger, André F Rendeiro, Christian Schmidl, et al. Pooled crispr screening with single-cell transcriptome readout. *Nature Methods*, 14(3):297–301, 2017.

Bart De Strooper and Eric Karran. The cellular phase of alzheimer's disease. *Cell*, 164(4):603–615, 2016.

Aleksandra Deczkowska, Hadas Keren-Shaul, Assaf Weiner, Marco Colonna, Michal Schwartz, and Ido Amit. Disease-associated microglia: A universal immune sensor of neurodegeneration. *Cell*, 173(5):1073–1081, 2018.

Atray Dixit, Oren Parnas, Biyu Li, et al. Perturb-seq: Dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016.

Mariano I Gabitto, Kyle J Travaglini, Victoria M Rachleff, et al. Integrated multimodal cell atlas of alzheimer's disease. *Nature Neuroscience*, 27(12):2366–2383, 2024.

Adam Gayoso, Zoë Steier, Romain Lopez, et al. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature Methods*, 18(3):272–282, 2021.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander J Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Alexandra Grubman et al. A single-cell atlas of entorhinal cortex from individuals with alzheimer's disease reveals cell-type-specific gene expression regulation. *Nature Neuroscience*, 22(12):2087–2097, 2019.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, 2017.

Miguel A Hernán and James M Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

Tushar Kamath, Abdulraouf Abdulraouf, S J Burris, et al. Single-cell genomic profiling of human dopamine neurons identifies a population that selectively degenerates in parkinson's disease. *Nature Neuroscience*, 25 (5):588–597, 2022.

Kenji Kamimoto, Blerta Stringa, Christy M Hoffmann, et al. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614:742–751, 2023.

David R Kelley. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, 2018.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, 2017.

Hadas Keren-Shaul et al. A unique microglia type associated with restricting development of alzheimer's disease. *Cell*, 169(7):1276–1290.e17, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Vitalii Kleshchevnikov, Andrii Shmatko, Emma Dann, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, 40(5):661–671, 2022.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Shane A Liddelow et al. Neurotoxic reactive astrocytes are induced by activated microglia. *Nature*, 541(7638): 481–487, 2017.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, 2019.

Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D Luecken, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40(1):121–130, 2022.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023.

Hansruedi Mathys et al. Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, 570(7761):332–337, 2019.

Thomas M Norman, Max A Horlbeck, Joseph M Replogle, et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2016.

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

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Sanjay R Srivatsan, Jose L McFaline-Figueroa, Vineet Ramani, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, 2015.

A Proof-of-concept experiment details

A.1 Data and preprocessing

We used the Norman et al. 2019 CRISPRa combinatorial perturbation dataset (Norman et al., 2019), obtained via the `pertpy` package. The dataset contains 111,255 cells with 19,018 genes, measuring responses to single and combination CRISPRa perturbations in K562 leukemia cells. Preprocessing: (i) gene filtering (retain genes expressed in ≥ 10 cells), (ii) cell filtering (retain cells with ≥ 200 detected genes), (iii) library size normalization to 10^4 counts per cell, (iv) $\log(1 + x)$ transformation.

Gene regulatory network edges were obtained from the TRRUST v2 database (Han et al., 2018), filtered to TFs present in the perturbation data. After filtering, only 10 TF→target edges (from 4 source TFs: GATA1, SPI1, CEBPB, NFKB1, STAT5A) remained—substantially fewer than anticipated.

This sparsity is the primary bottleneck for the hierarchy fidelity experiment (E1) and motivates the use of richer GRN sources in future work.

A.2 E1: HIERARCHY FIDELITY PROTOCOL

For each perturbation u with ≥ 20 cells, we compute:

- TF-activity shift: $\Delta t_i = \overline{x}_{\text{targets}(i)|u} - \overline{x}_{\text{targets}(i)|\text{ctrl}}$
- Gene-expression shift: $\Delta x_j = \log_2(\overline{x}_{j|u} + 1) - \log_2(\overline{x}_{j|\text{ctrl}} + 1)$

Sign-consistency for edge $(i \to j)$ with sign s_{ij}: $\mathbb{I}[\text{sign}(\Delta t_i \cdot \Delta x_j) = s_{ij}]$. Real GRN consistency is compared to 100 shuffled GRNs (randomized target assignments) via one-sample t-test.

Execution result: With only 10 GRN edges, no perturbations passed the minimum-cell filter after matching TF names to perturbation labels (0 perturbations evaluated; all metrics NaN). A GRN with $\geq$100 edges from SCENIC+ (Bravo González-Blas et al., 2023) or CellOracle (Kamimoto et al., 2023) is required.

A.3 E2: PERTURBATION PREDICTION PROTOCOL

Perturbations split 80/20 (stratified by single vs. combination). Baselines:

- **Mean shift:** $\widehat{\Delta x} = \frac{1}{|U_{\text{train}}|} \sum_u \Delta x_u$
- **Linear:** Ridge regression ($\alpha = 1$) per gene from one-hot perturbation encoding
- **scGen:** VAE-based latent arithmetic (Lotfollahi et al., 2019)

Metrics: Pearson R on top-K DE genes ($K \in \{20, 50, 100\}$), MSE on all genes.

A.4 E3: CROSS-CONTEXT PROTOCOL

Leiden clustering (resolution 0.5, 30 PCs) defines transcriptomic contexts. Train on Context A, evaluate on both Context A (in-context) and Context B (cross-context). Paired t-test assesses degradation.

A.5 E4: UNCERTAINTY-AWARE PLANNING PROTOCOL

For each candidate intervention, bootstrap 50 iterations ($n = 50$ cells with replacement), compute cosine similarity to a target state, and derive 90% confidence intervals. Calibration assessed by empirical coverage on held-out perturbations.

Barrier / Frontier	Evaluation approach	Status & next step
B1: Hierarchy fidelity	Sign-consistency of $\mathcal{L}_{\text{hier}}$ on real vs. shuffled GRN	GRN too sparse (10 edges); scale via SCENIC+
B2: Perturbation prediction	Held-out u; mean-shift, linear, scGen baselines	Protocol ready; awaiting GRN-scaled run
B3: Cross-context	Leiden-split train/test across cellular contexts	Protocol ready; requires sufficient per-context cells
Frontier: Planning	Bootstrap-based uncertainty ranking	Protocol ready; calibration analysis pending

Table 2: Proof-of-concept experiments on Norman 2019 K562 data: evaluation approach and implementation status. Results are preliminary; see text for GRN sparsity limitations.

A.6 CHALLENGES AND LESSONS LEARNED

1. **GRN sparsity:** TRRUST provides curated but incomplete interactions. After filtering to K562-expressed genes, coverage drops dramatically. Context-specific GRN inference (SCENIC+, CellOracle) is essential for hierarchy evaluation.

2. **Gene identifier harmonization:** Perturbation labels in the Norman dataset use gene symbols that do not always match TRRUST TF identifiers, requiring careful mapping.
3. **Code bug:** An initial implementation error (variable name mismatch in E1) was identified and corrected during development.

A.7 REPRODUCIBILITY

- **Data:** Norman 2019 via `pertpy` (`pip install pertpy`); TRRUST v2 from `https://www.grnpedia.org/trrust/`
- **Software:** Python 3.10+, `scanpy` 1.9+, `numpy`, `scipy`, `scikit-learn`
- **Compute:** All experiments designed to run on a single CPU node ($<$1 hour)
- **Code:** Available in the supplementary repository