Learning Interpretable Disease-relevant Gene Programs for Personalized Disease Phenotype Prediction

Tala Yasenpoor[1,†], Marjan Hosseini[1], and Alyssa J Matz[2],
Beiyan Zhou[2,3], and Derek Aguiar[1,3,†]

[1]*School of Computing, University of Connecticut, Connecticut, USA*

[2]*Department of Immunology, School of Medicine, University of Connecticut, Connecticut, USA*

[3]*Institute for Systems Genomics, University of Connecticut, Connecticut, USA*

Prediction of personalized disease phenotypes can enable earlier and more accurate diagnosis, characterize disease severity, and inform treatment and disease management. Gene expression profiles capture disease-relevant molecular processes that link genetic and environmental effects to clinical disease status and severity. However, existing expression-based risk prediction models either do not explicitly model the coordinated gene activity underlying cellular processes or yield predictions that are difficult to interpret. We introduce a supervised Bayesian Poisson factorization model for jointly learning disease-relevant gene programs (DRGPs) and predicting personalized disease phenotypes from RNA-seq data. The DRGP model decomposes a gene expression matrix into latent factors representing gene programs that can be informed by prior pathway information, while an L1-penalized logistic regression encourages sparsity over DRGPs and links program activities and auxiliary covariates to disease labels. A variational Bayes algorithm is derived to enable efficient inference. We evaluate our method on simulated and experimental single-cell RNA-seq COVID-19 data and GTEx whole-blood bulk RNA-seq data, the latter including a genetic risk covariate. DRGP learns biologically coherent gene programs, with predictive performance closest to the strongest baseline on COVID-19 pseudo-bulk data and competitive performance on single-cell and GTEx data. The source code for the DRGP method, simulation scripts, and appendix are freely available at https://github.com/bayesomicslab/DRGP.

Keywords: Disease Phenotype Prediction; Bayesian Hierarchical Modelling; Disease-Relevant Gene Programs; Supervised Poisson Factorization

[†]To whom correspondence should be addressed: {tsyasenpoor, derek.aguiar}@uconn.edu

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

1. Introduction

Predicting disease is a central challenge in clinical decision-making, because it can enable earlier and more accurate diagnosis, characterize disease subtype or severity, and inform personalized treatment and disease management[1]. Disease prediction focuses on either estimating future disease risk or classifying an individual's current disease status. Traditional risk models, such as the Framingham risk score[2] and the pooled cohort equations[3], use clinical factors including age, sex, cholesterol, blood pressure, and comorbidities. Disease phenotype prediction (DPP) methods instead use patient-specific measurements to classify phenotypes currently exhibited by the patient. For example, critically ill COVID-19 patients can be partitioned into clinically distinct patient groups that exhibit differential responses to corticosteroid treatment[4], supporting precision medicine approaches[5]. Polygenic risk scores (PRS) summarize inherited susceptibility as weighted sums of disease-associated alleles and can inform both tasks[6,7]. While PRS have some clinical utility[8], the proportion of disease variance typically explained is substantially lower than estimated phenotypic variance from twin study estimates of heritability[9,10]. Further, although the genome is largely fixed, disease risk can change over time in response to sociobehavioral factors or gene-by-environment interactions[11,12].

Since genetic and environmental risk factors can be mediated through gene expression changes[13,14], expression profiles can complement clinical and genetic predictors by capturing disease-associated molecular processes[15–17]. For example, transcriptomic signatures from circulating monocytes have been shown to improve prediction of atherosclerotic cardiovascular disease risk[18]. Since expression measured in individuals with disease may reflect both contributing processes and responses to disease (unlike inherited genotype), we focus on DPP rather than predicting future disease onset. However, predicting phenotypes from expression poses several methodological and biological challenges. First, coordinated gene activity produces high-dimensional, correlated predictors, whereas traditional risk models use relatively few clinical covariates[2,3]. Second, model interpretability is important for understanding expression-based predictions, motivating inherently interpretable models[19]. Lastly, gene expression varies substantially due to factors not specific to the target phenotype, including cell-type[20,21], environmental exposures[22,23], and immune activation[24,25], so only a small subset of genes is expected to be disease-relevant.

To address these challenges, expression-based DPP typically pairs feature-selection methods with linear predictors, reducing overfitting while maintaining model interpretability. Generalized linear models are widely used in disease prediction[3,26,27]. Machine-learning models, such as boosting classifiers, have also been applied to disease prediction[28,29] and can capture more complex relationships among features[30], although often at the expense of interpretability (see Appendix §A for additional details). However, neither additive predictors[2,3] nor flexible interactions among individual genes[28,29] explicitly represent the coordinated gene activity underlying disease-relevant cellular processes.

Related but largely orthogonal unsupervised approaches aim to infer *gene programs*, sets of co-regulated genes representing cellular processes, from gene expression data. Relative to individual-gene predictors, program-level representations reduce dimensionality, mitigate multicollinearity, and provide direct biological interpretability. Matrix factorization methods

such as hierarchical Poisson factorization[31] model expression using latent factors, providing a structured representation of sparse count data that moves beyond the gene-independent assumptions common in expression-based DPP. Methods like f-scLVM (slalom) and PLIER use pathway annotations to identify interpretable factors[32,33]. Pathway-informed variational autoencoders, including expiMap, VEGA, and pmVAE, incorporate gene sets into representation learning[34–36], while Spectra uses gene sets and cell-type annotations to guide program discovery[37]. However, without phenotype supervision, factors are optimized to explain expression rather than distinguish disease phenotypes, motivating the integration of gene programs with disease supervision.

In this work, we develop a supervised Bayesian Poisson factorization model that combines the interpretable representations of gene programs with optimization for disease relevance (Fig. 1). We define disease-relevant gene programs (DRGPs), which are subsets of genes that collectively contribute to some biological process relevant to disease phenotype and are optionally informed by pathway information. We associate sample usage of DRGPs with disease using a regularized generalized linear model that jointly models DRGP activities, clinical covariates, and genetic risk scores. We derive a variational Bayes inference algorithm using Pólya-Gamma augmentation to restore conjugacy of the supervised term, to efficiently compute DRGPs. We benchmark DPP on simulated and experimental bulk and single-cell RNA sequencing (RNA-seq and scRNA-seq, respectively). Simulations combine a disease liability model with scDesign3[38] for single-cell and SPsimSeq[39] for bulk data. Experimental evaluations examine COVID-19 disease severity in scRNA-seq[40] and heart disease from GTEx whole blood data[21]. Simulations show that our DRGP method recovers disease-relevant programs more accurately than unsupervised factorization when informative pathway priors are available, and that the inclusion of multi-modal information (e.g., genetic covariates) improves expression-based DPP. On experimental data, DRGP achieves the highest mean single-cell AUC among factorization methods, while methods that do not model gene programs perform best overall on single-cell and bulk data.

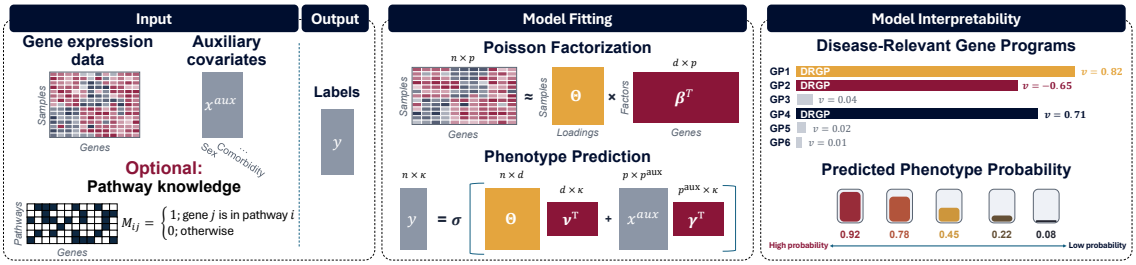


Fig. 1. **DRGP model overview.** From a gene expression count matrix $\boldsymbol{X}$, labels $\boldsymbol{y}$, covariates $\boldsymbol{x}^{\text{aux}}$, and an optional pathway mask $\boldsymbol{M}$ (left), the DRGP model simultaneously computes a Poisson factorization $\boldsymbol{X} \approx \boldsymbol{\Theta}\boldsymbol{\beta}^\top$ and predicted phenotype probabilities $\boldsymbol{y} = \sigma(\boldsymbol{\Theta}\boldsymbol{v}^\top + \boldsymbol{x}^{\text{aux}}\boldsymbol{\gamma}^\top)$ (center), yielding programs ranked by regression coefficient $|v|$ (large values indicate disease-relevance) and per-patient DPP scores (right).

2. Methods

Assume the data contain m individuals, where individual i contributes n_i expression samples (Fig. 1, Left). Function $\Pi : \{1,\ldots,j,\ldots,n\} \to \{1,\ldots,i,\ldots,m\}$ maps samples to individuals. In scRNA-seq, n_i represents the number of single-cell samples for individual i, whereas $n_i = 1$ for all i in bulk RNA-seq. Let $\boldsymbol{X}$ be an $n \times p$ gene expression matrix where $n = \sum_{i=1}^{m} n_i$ denotes the total number of samples, p is the number of genes, and row $\boldsymbol{x}_j$ and column $\boldsymbol{x}_{\cdot g}$ denote sample j and gene g, respectively. Our model assumes there exists a d-dimensional latent factor for each sample $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_j,\ldots,\boldsymbol{\theta}_n\}$ and each gene $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_g,\ldots,\boldsymbol{\beta}_p\}$. Sample- and gene-specific latent factors generate x_{jg}, *i.e.* the expression of gene g in sample j. Each individual i is associated with a κ-dimensional vector $\boldsymbol{y}_i \in \{0,1\}^{\kappa}$ denoting the phenotype labels and a vector of observed auxiliary covariates, $\boldsymbol{x}_i^{\text{aux}}$. We use inherited labels $y_{jk} = y_{\Pi(j)k}$ and covariates $x_j^{\text{aux}} = x_{\Pi(j)}^{\text{aux}}$.

2.1. *The DRGP Model*

The DRGP model generative process couples hierarchical Poisson factorization with logistic regression through the learned sample factor matrix (Fig. 1, Middle, Fig. 2, and Table 1).

For each sample j:

- Generate sample activity $\xi_j \sim \text{Gamma}(\alpha_\xi, \lambda_\xi)$.
- For each factor, generate sample factors $\theta_{j\ell} \sim \text{Gamma}(\alpha_\theta, \xi_j)$.

For each gene g:

- Generate gene inclusion rate $\pi_g \sim \text{Beta}(\alpha_\pi, \beta_\pi)$.
- Generate gene activity $\eta_g \sim \text{Gamma}(\alpha_\eta, \lambda_\eta)$.
- For each factor ℓ:
  - Generate gene–program membership $m_{g\ell} \sim \text{Bernoulli}(\pi_g)$.
  - Generate slab loading $\tilde{\beta}_{g\ell} \sim \text{Gamma}(\alpha_\beta, \eta_g)$.
  - Set spike-and-slab prior $\beta_{g\ell} = m_{g\ell}\,\tilde{\beta}_{g\ell}$.

For each sample j and gene g:

- Generate gene expression $x_{jg} \sim \text{Poisson}\left(\theta_j \beta_g^\top\right)$.

For each phenotype k and factor ℓ:

- Generate latent factor regression coefficients $\upsilon_{k\ell} \sim \text{Laplace}(\mu_\upsilon, b)$.

For each phenotype k:

- Generate auxiliary regression coefficients $\gamma_k \sim \text{Normal}(\mu_\gamma, I\sigma_\gamma^2)$.

For each sample j and phenotype k, generate $y_{jk} \sim \text{Bernoulli}\left(\sigma\left(\theta_j \upsilon_k^\top + x_j^{\text{aux}} \gamma_k^\top\right)\right)$, where σ is the logistic function.

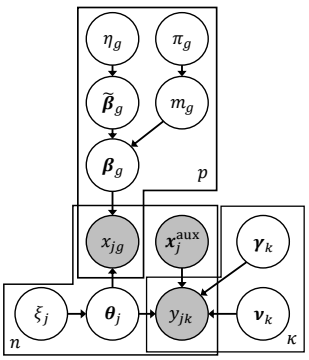


Fig. 2. **Supervised Poisson Factorization graphical model.** We omit the hyperparameters for brevity. We note that $\boldsymbol{X}^{\text{aux}}$ is an $n \times p^{\text{aux}}$ matrix of auxiliary covariates (e.g., age or sex).

2.2. *Variational Inference Algorithm*

Exact posterior inference is intractable due to the non-conjugacies introduced by the logistic supervision and sparsity inducing priors. We therefore derive a mean-field coordinate-

ascent variational inference algorithm[41], restoring conjugacy of the supervised term by Pólya-Gamma augmentation[42] (Algorithm 1). Training sequentially updates the gene program parameters, sample-specific program activities, and supervised regression parameters. Prior pathway knowledge enters through a binary mask $M \in \{0,1\}^{p\times d}$ with $M_{g\ell} = 1$ if gene g belongs to pathway ℓ; entries with $M_{g\ell} = 0$ are excluded from all updates, giving three modes: *de novo* (no mask), pathway-guided (all d columns masked), and combined (d_{path} masked columns alongside $d - d_{\text{path}}$ free columns). Pólya–Gamma augmentation preserves conjugacy of the regression coefficients, while the phenotype contributes to each sample activity through a supervised correction that couples gene program activity and disease status (Eq. B.10).

For a held-out sample, all global parameters are fixed and only its local program activities $(\boldsymbol{\theta}_{j^\star},\xi_{j^\star})$ are inferred from expression. Disease phenotype probabilities are then computed from the posterior mean and variance of the logistic linear predictor. Each training sweep is dominated by the sparse count-allocation update and scales as $\mathcal{O}(\text{nnz}(\boldsymbol{X})d)$ where $\text{nnz}(\boldsymbol{X}) = |\{(j,g) : x_{jg} \neq 0\}|$. Memory scales as $\mathcal{O}(\text{nnz}(\boldsymbol{X}) + (n + p)d)$. The variational parameter update derivations, detailed inference and prediction algorithms, and complexity analysis are given in Appendix §B (Algs.2,3).

Table 1. Summary of model variables.

Param.	Dim.	Description
$\boldsymbol{\theta}$	$n \times d$	factor matrix for $\boldsymbol{X}$ and design matrix for $\boldsymbol{Y}$
$\boldsymbol{\beta}$	$p \times d$	loading matrix for $\boldsymbol{X}$
$\boldsymbol{\upsilon}$	$\kappa \times d$	regression coefficients for $\boldsymbol{\theta}$
$\boldsymbol{\gamma}$	$\kappa \times p^{\text{aux}}$	regression coefficients for x_{aux}
$\boldsymbol{y}$	$n \times \kappa$	individual phenotype
$\boldsymbol{x}$	$n \times p$	gene expression matrix
$\boldsymbol{X}^{\text{aux}}$	$n \times p^{\text{aux}}$	auxiliary factors (e.g., sex)

Algorithm 1 DRGP model training and test set prediction

Input: Expression $\boldsymbol{X}$, labels $\boldsymbol{Y}$, covariates $\boldsymbol{X}^{\text{aux}}$, number of programs d, and pathway mask $\boldsymbol{M}$
Output: Fitted variational parameters and test set disease phenotype probabilities

Initialize variational parameters for $(\tilde{\boldsymbol{\beta}},\boldsymbol{m}_g,\boldsymbol{\pi},\boldsymbol{\eta},\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\upsilon},\boldsymbol{\gamma})$ and enforce the pathway mask $\boldsymbol{M}$
while the ELBO has not converged **do** ▷ Training loop
    Update variational parameters for $(\tilde{\boldsymbol{\beta}},\boldsymbol{m}_g,\boldsymbol{\pi},\boldsymbol{\eta},\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\upsilon},\boldsymbol{\gamma})$ using equations B.3 to B.14
end while
for each held-out sample $j^\star$ **do** ▷ Test set prediction
    while the local Poisson objective has not converged **do**
        Update $(\boldsymbol{\theta}_{j^\star},\xi_{j^\star})$ from test set expression data $\boldsymbol{x}_{j^\star}$ using Equations B.8 and B.16
    end while
    $\hat{y}_{j^\star k} \leftarrow \sigma\left(\dfrac{\mathbb{E}_q[A_{j^\star k}]}{\sqrt{1 + \frac{\pi}{8}\,\text{Var}_q(A_{j^\star k})}}\right)$ ▷ Equation B.17
end for

2.3. *Simulation Algorithm*

We assess gene program recovery, phenotype prediction, and auxiliary-effect attribution using synthetic benchmarks built from experimental data. The simulation generates background expression using scDesign3[38] on single-cell and SPsimSeq[39] on bulk data, assigns nuisance and DRGPs, perturbs sample gene expression, and generates phenotype labels from program activity and auxiliary covariates. We defer a full description of the modality-specific instantiations to the results (Sec. 3).

Simulated Gene Programs. Given a simulated expression matrix X^0, we simulate d^* gene programs $\{S_\ell\}_{\ell=1}^{d^*}$. Each program ℓ is defined by a support set of genes S_ℓ and a parameter $\phi_{g\ell} \in \mathbb{R}$ that determines the magnitude of expression changes induced when the program is active. Programs are partitioned into disease-relevant ($\ell \in \mathcal{D}$) and nuisance ($\ell \in \mathcal{N}$) programs. Both program types perturb gene expression through the same mechanism, but only DRGPs contribute to phenotype generation. This separation enables evaluation of whether methods can distinguish transcriptional variation associated with disease from unrelated biological variation. *De novo* program genes are sampled uniformly without replacement or they can be defined by biological pathways. The gene support for pathway-guided programs, S_ℓ, is determined by a binary mask M defined by a known pathway (§2.2); *de novo* programs remain hidden from all methods and must be discovered directly from the simulated expression data.

Assigning Programs to Samples. Disease-associated transcriptional programs ($S_\ell : \ell \in \mathcal{D}$) are assigned to individuals rather than to cells. We draw each individual's latent disease propensity as $p_i \sim \text{Beta}(\alpha_p,\beta_p)$ where larger values increase the prevalence of disease-relevant programs across that individual's cells. To allow for program-specific variability across programs, we define $\tilde{p}_{i\ell} = \sigma\big(\sigma^{-1}(p_i) + u_{i\ell}\big)$ where $u_{i\ell} \sim \mathcal{N}(0,\sigma_u^2)$. Disease-relevant programs are then assigned to individuals according to $c_{i\ell} \sim \text{Bernoulli}\left(q_0 + (q_1 - q_0)\lambda_\ell \tilde{p}_{i\ell}\right), \ell \in \mathcal{D}$, where $(0 \le q_0 \le q_1 \le 1)$ are the baseline and maximum carrier probabilities, respectively, and $\lambda_\ell \in [0,1]$ modifies program prevalence. As a result, individuals with larger disease propensities have a higher probability of carrying DRGPs. In contrast, nuisance programs are assigned independently of disease propensity, ensuring that nuisance programs affect gene expression but not disease status: $c_{i\ell} \sim \text{Bernoulli}(\pi_\ell)$ where $\ell \in \mathcal{N}$.

Gene Expression Perturbation. After program assignment, we perturb expression based on program activation and gene-specific effects. Let μ_{jg}^0 denote the background expression mean for sample j and gene g. Active programs perturb the background expression on the log-mean scale according to

$$\log \mathbb{E}[x_{jg} \mid X^0] = \log \mu_{jg}^0 + \sum_\ell a_{j\ell}\phi_{g\ell}, \tag{1}$$

where $a_{j\ell}$ denotes the activation of program ℓ in sample j and $\phi_{g\ell}$ perturbs the expression of gene g in program S_ℓ additively; $\phi_{g\ell} = 0$ for genes outside the program support. We then resample counts from the perturbed means using the modality-specific procedures in §3.2.

Mediation and Label Generation. After expression perturbations, phenotype labels are generated from DRGP activity and auxiliary covariates using a liability-threshold model. To

allow expression-mediated effects, DRGP activations are $a_{j\ell} = a_{j\ell}^{(0)} + \kappa_{\text{med}} \operatorname{sign}(\omega_\ell)\tilde{s}_{\Pi(j)}$ for $\ell \in \mathcal{D}$, while nuisance activations remain $a_{j\ell} = a_{j\ell}^{(0)}$. Here $a_{j\ell}^{(0)}$ is the baseline activation, $\tilde{s}_i$ is a standardized auxiliary covariate, and ω_ℓ denotes the program's liability weight. The program burden is $Z_i = \sum_{\ell \in \mathcal{D}} \omega_\ell A_{i\ell}$, where $A_{i\ell}$ is the individual-level program activity constructed as described in §3.2. The resulting burden is standardized as $\tilde{Z}_i = (Z_i - \overline{Z})/\operatorname{sd}(Z)$. Disease liability is then defined as

$$l_i = \sqrt{w_Z}\tilde{Z}_i + \sqrt{(1-\tilde{m})w_s}\tilde{s}_i + \sqrt{1 - w_Z - (1-\tilde{m})w_s}\,\epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,1), \tag{2}$$

where $\tilde{m} \in [0,1]$ controls mediation, w_Z weights the program burden, and w_s is the nominal auxiliary weight. The coupling scale κ_{med} is calibrated as described in Appendix §E.2. At $\tilde{m} = 0$, $\kappa_{\text{med}} = 0$ and the independent program and auxiliary contributions have liability variance shares w_Z and w_s. Increasing $\tilde{m}$ introduces expression-mediated effects and reduces the direct auxiliary contribution, which vanishes at $\tilde{m} = 1$. Because both contributions share $\tilde{s}_i$, their combined variance is not constant across mediation settings. Disease labels are generated by thresholding the liability, $D_i = \mathbb{1}[l_i > \tau]$, where τ sets disease prevalence. Because experimental data lack cell-level labels, each cell inherits its individual's label.

3. Results

We evaluated our DRGP method on both single-cell and bulk RNA-seq cohorts, each used for both synthetic and experimental evaluation. Synthetic benchmarks enable evaluation of program recovery and auxiliary-effect attribution because the true DRGPs are known. The experimental cohorts assess prediction of observed disease phenotypes and interpretation of the associated expression programs. We compared three DRGP configurations (unmasked, masked, and combined) with two gene program methods, scHPF[31] and Spectra[37], and two general matrix-factorization methods, NMF and PCA. The unsupervised baselines were followed by L_1-regularized logistic regression. Where applicable, we also included gene-level L_1-regularized logistic regression; model-selection details are provided in Appendix §D.

3.1. *Datasets and Preprocessing*

COVID-19 PBMC scRNA-seq. Single-cell experiments used the COVID-19 peripheral blood mononuclear cell (PBMC) atlas[40], comprising 846,847 cells from 152 biological samples collected from 131 COVID-19 patients spanning "mild or moderate" (61 patients) and "severe or critical" (70 patients) disease labels. All data splits were grouped by patient, ensuring that all samples and cells from the same patient remained in a single partition within each split. Unless otherwise noted, analyses used unnormalized UMI counts. Gene symbols were mapped to Ensembl (v. 116)[43]. We removed duplicate genes, genes not mapped to Ensembl, non-protein-coding genes, and genes with 0 expression across all samples. For single-cell analyses, we additionally required detection in at least 0.1% of cells, retaining 13,791 genes. No additional detection-frequency restriction was imposed on pseudo-bulk profiles, which retained 15,754 genes. Severity labels were defined at the patient level, with sex and three comorbidity indicators (asthma or chronic obstructive pulmonary disease, cardiovascular disease, and diabetes) included as auxiliary covariates.

GTEx Whole Blood. Bulk experiments used GTEx v8 whole-blood RNA-seq data[21]. Because GTEx lacks a single cardiovascular phenotype, we defined a binary label equal to 1 for any history of prior myocardial infarction or acute coronary syndrome (`MHHRTATT`), ischemic heart disease or coronary artery disease (`MHHRTDIS`), or secondary heart-disease history indicator (`MHHRTDISB`). After removing individuals with missing transcriptomic and phenotype information, 708 individuals remained. Restricting to those with whole-genome sequencing required for the polygenic risk-score covariate, 631 individuals remained in the cohort, with 148 cases and 483 controls. Predictive analyses used raw read counts from protein-coding genes detected in at least 0.1% of samples, resulting in 16,246 genes. Separate GTEx program-interpretation fits used the full input set of 27,693 genes without protein-coding restriction (Table E1). Auxiliary covariates included sex, four race indicators, standardized age and BMI, smoking status, hypertension, type 2 diabetes, and the coronary artery disease PRS described in §3.2.

Pathway masks. Predictive analyses used human Reactome release 96 for GTEx (depth-2 pathways with at least 15 genes) and a 500-pathway selection from human MSigDB v2024.1, collection C2:CP[44], for COVID-19. After gene mapping and filtering, the masks retained 309, 351, and 346 pathways for GTEx, COVID-19 pseudo-bulk, and single-cell analyses, respectively. The exact GMT files and construction details are provided with the code (Appendix §F).

3.2. *Simulation Instantiation*

Here, we describe modality-specific instantiations of the general simulation framework (§2.3).

COVID-19 Simulation. We generated realistic background expression by fitting scDesign3 to healthy-control PBMCs[38]. Each simulated dataset consists of 80 individuals, 100 cells per individual across six cell types, and 10,000 genes per cell. Simulations include two annotated and two *de novo* DRGPs, plus one *de novo* nuisance program; each *de novo* program contains 50–100 genes. Individuals are assigned DRGPs, which are activated only in predefined responder cell types for a random fraction eligible cells (default $\rho = 0.3$). Program perturbations are added to the background expression (Eq. 1). For label generation, activations are standardized across cells and averaged within individual by $A_{i\ell} = |C_i|^{-1} \sum_{j \in C_i} \tilde{a}_{j\ell}$, where C_i is individual i's set of cells. We use $\tilde{m} = 0$ and target prevalence 0.5; each cell inherits its individual's label. We vary the perturbation strength relative to cell-type variation $r = \delta^2 \frac{\mathrm{SS}_{\mathrm{pert}}}{\mathrm{SS}_{\mathrm{type}}} \in \{0.05, 0.15, 0.30\}$, where $\mathrm{SS}_{\mathrm{pert}}$ is the variation induced by the simulated program, $\mathrm{SS}_{\mathrm{type}}$ is the between cell-type variation, and δ is calibrated to achieve the target value of r. We parametrize the single-cell arm by r rather than a raw effect size because cell-type identity dominates single-cell expression variance, so a program's detectability depends on its magnitude relative to that structure. Finally, we vary the transcriptomic contribution to disease liability $w_Z \in \{0.1, 0.3, 0.5, 0.7\}$, generating five independent datasets for each combination of w_Z and r.

GTEx Simulation. Background expression is generated using SPsimSeq fit to GTEx whole blood samples[39], yielding 2,000 samples each with 5,000 genes. Each simulation uses six DRGPs, two annotated and four *de novo*, together with two nuisance programs, each supported on 50–150 genes. Programs are assigned to samples and directly perturb bulk ex-

pression. Samples include genetic covariates based on a coronary artery disease PRS (ID: PGS000018, metaGRS_CAD)[45] computed from GTEx donor genotypes (aligned to GRCh38 using `plink2`[46]). For each GTEx donor with genetic data, we computed $\mathrm{PRS}_i = \sum_v w_v\, g_{iv}$, where w_v is the PRS effect weight of variant v and genotype $g_{iv} \in \{0, 1, 2\}$ denotes the effect-allele dosage of individual i at variant v. The 866 donor PRS values were z-scored and upsampled to 2,000 synthetic expression samples by independently sampling with replacement from the PRS pool. Each individual contributes one expression sample, with baseline activations $a_{i\ell}^{(0)} = c_{i\ell}$ and $A_{i\ell} = a_{i\ell}$; only the summed program burden is standardized. Labels follow the liability model in § 2.3 at target prevalence 0.26. For program recovery and prediction, we set $\tilde{m} = 0$ and generate five independent replicates at $w_Z = 0.3$, $w_s = 0.1$, and each effect size $\phi \in \{\log_2 3{,}3.0{,}4.5\}$. Parameter settings for these benchmarks are provided in Table D1; the attribution experiment varies mediation as described in § 3.3.

3.3. *Simulation Results*

Recovery of Simulated Gene Programs. We evaluated DRGP recovery at $\tilde{m} = 0$ by matching inferred programs to ground truth programs using the Hungarian algorithm with cosine similarity. In our model, gene program score vectors are defined by $\hat{s}_{g\ell} = \mathbb{E}_q[\beta_{g\ell}]\mathbb{E}_q[\eta_g]$. The area under the precision-recall curve (AUPRC) is then computed for the inferred programs matched to the true DRGPs (see Appendix §C for more details).

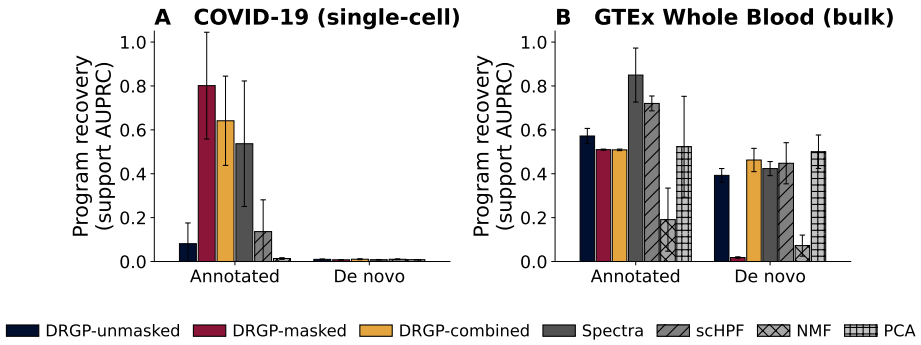


Fig. 3. **Recovery of simulated DRGPs across data modalities and prior-knowledge settings.** Support AUPRC is compared for **(A)** COVID-19 single-cell and **(B)** GTEx whole-blood simulations with $\tilde{m} = 0$. Bars are means over five simulated draws and error bars show ±1 SD.

In the scRNA-seq simulations, methods that leverage pathway annotations (DRGP-masked, DRGP-combined, and Spectra) recovered ground-truth gene programs more accurately than methods without them (DRGP-unmasked, scHPF, NMF, and PCA; Fig. 3A). Program recovery AUPRC was insensitive to varying w_Z and r except for a slight upward trend for Spectra (Fig. E1); methods performed similarly across varying responder fractions ρ (Fig. E2). All methods recovered *de novo* programs at near-chance levels, likely because their signals were obscured by pathway and cell-type variation inherited from the experimental scRNA-seq background. Recovery was generally higher in bulk data, where programs perturb expression directly at the subject level (Fig. 3B). Spectra and scHPF performed best for annotated programs, while PCA, DRGP-combined, scHPF, Spectra, and DRGP-unmasked were

broadly comparable for *de novo* recovery.

Disease Phenotype Prediction. In the scRNA-seq simulations, patient-level AUC increased with ρ, r, and w_Z (Figs. E3, E4), but cell-level performance remained near chance across methods, likely because cells inherited noisy patient-level phenotype labels (Fig. E11). After aggregating cell-level predictions by individual, DRGP-combined and Spectra achieved the highest AUC (Fig. 4A). In the bulk simulations, our method remained competitive with the strongest factorization baselines (Fig. 4B).

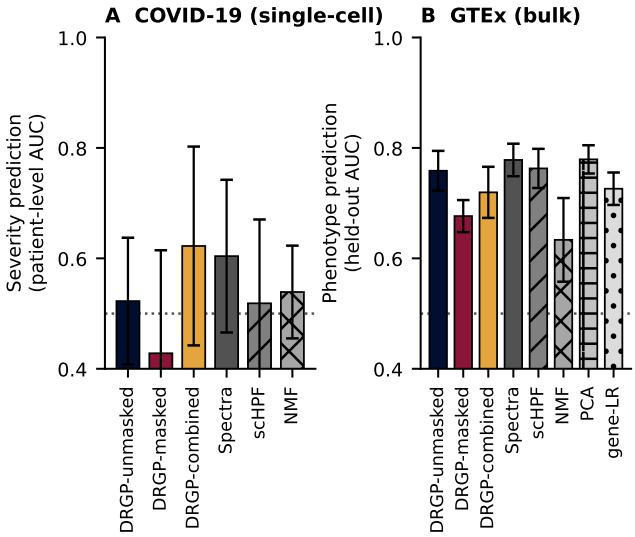


Fig. 4. **The DRGP method remains competitive with baselines on simulated RNA-seq phenotype prediction.** Both simulation arms use $\tilde{m} = 0$. Left: COVID-19 scRNA-seq simulations evaluated on cells from held-out patients using inherited labels, aggregated to form patient-level prediction. Right: GTEx whole-blood simulations evaluated on held-out subjects. Bars show means with error bars denoting ±1 SD. Dotted lines mark chance performance (AUC = 0.5); the displayed y-axes span 0.4–1.0.

Attribution of Multi-modal Auxiliary Effects in GTEx simulations. We next varied the mediation setting in the GTEx simulation, using $\tilde{m} \in \{0,0.5,1\}$ and $w_s \in \{0.05,0.1,0.2\}$ at $w_Z = 0.3$, with five independent replicates per combination. These settings represent direct-only, partially mediated, and fully expression-mediated PRS contributions, respectively. Coupling calibration and evaluation procedures are provided in Appendix §E.2.

Program–PRS correlations were close to the permutation reference without mediation and increased as mediation strengthened (Fig. 5A). The auxiliary PRS coefficient decreased but did not consistently vanish when the direct contribution was removed (Fig. 5B). Held-out AUC also declined with increasing mediation (Fig. 5C). Thus, inferred programs captured PRS-dependent expression, but direct and mediated effects were not fully separated.

3.4. *Experimental Results*

Experimental prediction used five repeats of stratified, individual-grouped five-fold cross-validation with identical stored partitions across methods; 15% of each outer training cohort was reserved for validation (Appendix §F). We averaged test-fold ROC-AUC and AUPRC across 25 evaluations, scoring cells for single-cell data and individuals for bulk data (Table 2). All methods were refitted per fold for bulk analyses. For single-cell analyses, scHPF and Spectra representations were learned once from the full expression matrix with fold-specific logistic classifiers due to runtime constraints (Fig. F1); other methods were refitted per fold. Hyperparameters were selected before evaluation, without outer-fold retuning (Appendix §E.3). Factor counts and prior hyperparameters were selected by Optuna[47] using grouped validation

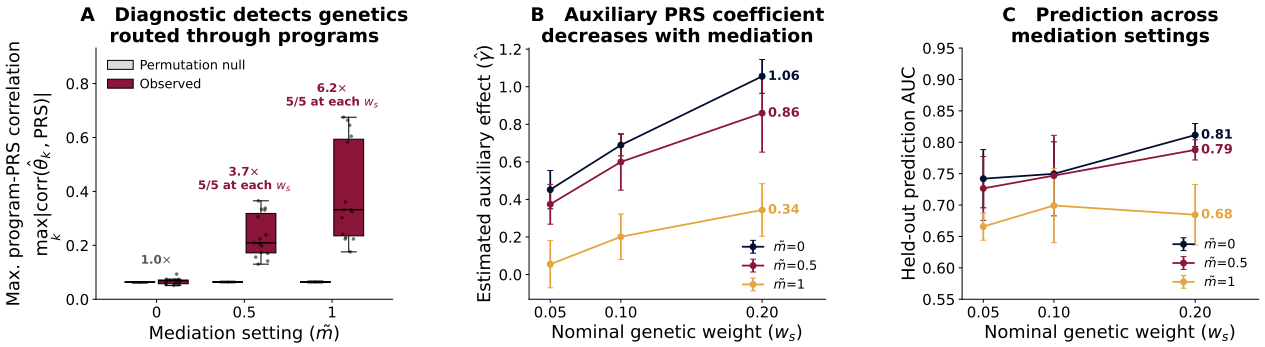


Fig. 5. **DRGP behavior under simulated PRS–program coupling.** GTEx-based simulations with $w_Z = 0.3$, $w_s \in \{0.05,0.1,0.2\}$, and mediation settings $\tilde{m} \in \{0,0.5,1\}$. **(A)** Maximum absolute correlation between the PRS and any fitted program, compared with a reference distribution from 50 PRS permutations. Boxes summarize 15 simulation outcomes per $\tilde{m}$: five independent truth replicates evaluated at three w_s settings. Outcomes sharing a truth replicate are dependent. Count annotations give the number of replicates with observed correlation exceeding their permutation reference at each w_s. **(B)** Posterior mean auxiliary PRS coefficient. **(C)** Held-out phenotype-prediction AUC. In (B) and (C), points and error bars show means and ±1 SD across five simulation replicates. The simulation model is described in §2.3; coupling calibration and evaluation procedures appear in Appendix §E.2.

AUC; masked factor counts were determined by retained pathways, while COVID-19 combined runs used 50 *de novo* factors (Appendix §D; Table D2). Calibration analyses appear in Appendix §E.4.

COVID-19 severity prediction. We evaluated disease severity prediction using individual scRNA-seq profiles, with cells inheriting their patient's label, and pseudo-bulk profiles obtained by summing raw UMI counts across each patient's cells (Table 2; Figs E5 and E6). On single-cell data, gene-level logistic regression achieved the highest mean predictive performance, while DRGP-unmasked had the highest mean AUC and AUPRC among factorization methods, closely followed by PCA. On pseudo-bulk data, PCA achieved the highest mean scores, followed by DRGP-unmasked and gene-LR. Across both representations, pathway-constrained DRGP configurations had lower mean AUC than DRGP-unmasked. This may reflect restrictions on the expression patterns they can represent: masked fits exclude genes outside the retained pathways, while combined fits retain all genes but constrain most programs to predefined pathways. Platt scaling improved calibration for Spectra and scHPF, while effects on DRGP varied by configuration and metric (Figs. E8 and E9).

GTEx heart disease prediction. We evaluated heart disease status prediction from GTEx whole-blood expression with clinical covariates and a coronary artery disease PRS (Table 2; Fig E7). PCA achieved the highest mean AUC and AUPRC. All methods performed worse on GTEx than on COVID-19, with PCA remaining the top performer. Platt scaling brought DRGP reliability curves closer to the calibration diagonal, although deviations remained (Fig. E10).

Table 2. **Experimental phenotype prediction.** Mean ± standard deviation of ROC-AUC and AUPRC across the 25 test folds of 5 × 5 individual-grouped CV. Single-cell metrics score cells; bulk metrics score individuals. Bold and italics indicate the highest and second-highest values in each column, respectively. † In the COVID-19 single-cell column only, scHPF and Spectra representations were learned once from the full expression matrix with fold-specific logistic heads; both bulk columns use per-fold inductive fits in which the factorization is estimated on training samples, the gene programs are frozen, and held-out samples are scored by projection. (Appendix §E.3).

	COVID single-cell		COVID pseudo-bulk		GTEx	
Method	AUC	AUPRC	AUC	AUPRC	AUC	AUPRC
DRGP-unmasked	*0.720 ± 0.120*	*0.737 ± 0.109*	*0.822 ± 0.074*	*0.864 ± 0.064*	0.662 ± 0.065	0.366 ± 0.066
DRGP-masked	0.688 ± 0.111	0.695 ± 0.103	0.816 ± 0.065	0.858 ± 0.054	0.636 ± 0.043	0.355 ± 0.050
DRGP-combined	0.699 ± 0.115	0.708 ± 0.112	0.789 ± 0.105	0.824 ± 0.087	0.621 ± 0.066	0.365 ± 0.071
NMF	0.710 ± 0.079	0.718 ± 0.094	0.743 ± 0.127	0.809 ± 0.106	0.723 ± 0.051	*0.436 ± 0.072*
PCA	0.716 ± 0.077	0.730 ± 0.089	**0.840** ± 0.059	**0.885** ± 0.045	**0.742** ± 0.056	**0.444** ± 0.079
gene-LR	**0.786** ± 0.060	**0.799** ± 0.062	0.821 ± 0.075	0.862 ± 0.061	0.647 ± 0.053	0.328 ± 0.053
Spectra†	0.704 ± 0.060	0.728 ± 0.062	0.746 ± 0.103	0.803 ± 0.099	0.718 ± 0.053	0.401 ± 0.067
scHPF†	0.689 ± 0.061	0.697 ± 0.071	0.792 ± 0.096	0.824 ± 0.090	*0.727 ± 0.054*	0.430 ± 0.074

Interpretation of Learned Gene Programs. For each cohort, we fit DRGP-combined under five initialization seeds (Table E1) and interpreted programs present in at least three runs with mean pairwise top-50 gene-set Jaccard similarity ≥ 0.30 (Appendix §E.6). GTEx interpretation used separate fits on the full gene set, including noncoding genes. Within programs, genes are ranked by $\hat{s}_{g\ell} = \mathbb{E}_q[\beta_{g\ell}]\mathbb{E}_q[\eta_g]$ (Eq. C.1), and programs are ordered by an effect-scaled importance score $I_\ell = |\bar{v}_{k\ell}|\mathrm{sd}(\bar{\boldsymbol{\theta}}_{\cdot\ell})$ averaged over all seeds. Masked programs are more stable due to their pathway priors, but only 16/300 GTEx and 7/50 COVID-19 *de novo* programs were stable (Fig. E12). Appendix Table E2 summarizes supporting gene-function associations.

In the COVID-19 experiments, DRGP-combined identified gene programs representing critical cellular processes, including chemotaxis, protein translation, DNA replication, and microautophagy (Fig. E13). The program most strongly associated with severe disease (GP7) was composed of immune-related genes, including those involved in cell proliferation (11/50 genes), monocyte and macrophage markers (8/50 genes), and important immune-regulatory metallothionein proteins (7/50). Interestingly, the two top programs (GP7 and GP3) suggest metal ion homeostasis markers are important for predicting disease state, with seven metallothionein proteins as well as proteins that require binding to zinc or other metals to function (15/100). This is particularly relevant given the important role of metal ion homeostasis in infection susceptibility via immune cell function[48,49]. Lastly, both GP7 and GP3 contain cytoskeletal genes, which may overlap with the top pathway, "Mechanical Load Activates Signaling By PIEZO1 and Integrins in Osteocytes", which may suggest a role for cell recruitment and chemotaxis in prediction, despite opposite associations.

In the GTEx results, the pathways with the largest effect-scaled importance are critical cell processes including DNA repair, mRNA transcription, and mRNA decay (Fig. E14). The leading programs include several novel transcripts and pseudogenes, which may be related to differences in the central cell process pathways that are top predictors. In addition, GP13 has 24/50 membrane receptors, including 5 involved in cell adhesion, which has a known role in

atherogenesis and plaque progression[50]. GP3 and GP5 are enriched in markers for megakaryocytes/platelets and B cells, respectively. Platelets are crucial to vascular inflammation and vessel integrity[51,52] while B cell antibodies can be either progressive or protective in atherogenesis depending on the epitopes targeted[53].

Computational resources. Experiments used a heterogeneous SLURM cluster. DRGP and single-cell Spectra used GPUs; other jobs used CPUs. Median DRGP job times were 5–8 min for GTEx, 2–3 min for COVID-19 pseudo-bulk, and 251–568 min for single-cell data. Full-cohort single-cell Spectra and scHPF representation fits took 649 and 2,006 min, respectively. Hardware, software, memory usage, and timing protocols are detailed in Appendix F (Fig. F1; Table F1).

4. Discussion

Several limitations and directions for future work warrant consideration. Learned programs represent disease-associated expression, not causal mechanisms; pathway interpretation depends on annotation quality and relevance. In scRNA-seq, cells inherit patient labels despite potentially cell-type-specific disease activity; patient-grouped splits do not explicitly model this hierarchy. Future work should develop probabilistic pathway priors and hierarchical models separating patient-, cell-type-, and cell-level variation. Although DRGP extends to non-binary outcomes, our evaluation covers only binary outcomes in two transcriptomic cohorts.

Strict pathway constraints may limit prediction when annotations are incomplete or poorly matched to the disease or cellular context. Longitudinal, prospective, and external-cohort studies could assess generalization and whether programs precede disease progression; our observational data cannot establish causal direction. Additionally, fairness and robustness under data distribution shift should be evaluated before clinical use. Because transcriptomic and auxiliary data may contain identifying information, secure data governance and access controls remain necessary to mitigate re-identification and unauthorized use risks[54,55].

5. Conclusions

DRGP combines sparse supervised Bayesian Poisson factorization, pathway information, and auxiliary covariates for interpretable DPP. Experimental results show a trade-off between program interpretation and prediction, with DRGP-unmasked most competitive on COVID-19 pseudo-bulk data. Simulations show that inferred programs capture PRS-dependent expression and that auxiliary coefficients decline as the explicit PRS contribution decreases, without establishing exact separation of direct and mediated effects.

6. Acknowledgments

This work was supported in part by NIH NHLBI (1R01HL172239, awarded to Dr. Beiyan Zhou), whose funding is gratefully acknowledged.

References

1. A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia and A. Y. Ng, Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network, *Nature medicine* **25**, 65 (2019).
2. P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz and W. B. Kannel, Prediction of coronary heart disease using risk factor categories, *Circulation* **97**, 1837 (1998).
3. D. C. Goff, D. M. Lloyd-Jones, G. Bennett, S. Coady, R. B. D'agostino, R. Gibbons, P. Greenland, D. T. Lackland, D. Levy, C. J. O'donnell *et al.*, 2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines, *Journal of the American College of Cardiology* **63**, 2935 (2014).
4. C. López-Martínez, P. Martín-Vicente, J. Gómez de Oña, I. López-Alonso, H. Gil-Peña, E. Cuesta-Llavona, M. Fernández-Rodríguez, I. Crespo, E. Salgado del Riego, R. Rodríguez-García *et al.*, Transcriptomic clustering of critically ill covid-19 patients, *European Respiratory Journal* **61**, p. 2200592 (2023).
5. R. Knoll, E. T. Helbig, K. Dahm, O. Bolaji, F. Hamm, O. Dietrich, M. van Uelft, S. Müller, L. Bonaguro, J. Schulte-Schrepping *et al.*, The life-saving benefit of dexamethasone in severe covid-19 is linked to a reversal of monocyte dysregulation, *Cell* **187**, 4318 (2024).
6. C. Iyegbe *et al.*, The emerging molecular architecture of schizophrenia, polygenic risk scores and the clinical implications for GxE research, *Soc Psychiatry Psychiatr Epidemiol* **49**, 169 (2014).
7. S. W. Choi, T. S.-H. Mak and P. F. O'Reilly, Tutorial: a guide to performing polygenic risk score analyses, *Nature Protocols* **15**, 2759 (September 2020).
8. N. Mavaddat *et al.*, Polygenic risk scores for prediction of breast cancer and breast cancer subtypes, *The American Journal of Human Genetics* **104**, 21 (2019).
9. T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti *et al.*, Finding the missing heritability of complex diseases, *Nature* **461**, 747 (2009).
10. R. Plomin and S. Von Stumm, Polygenic scores: prediction versus explanation, *Molecular psychiatry* **27**, 49 (2022).
11. M. Whitesell, A. Bachand, J. Peel and M. Brown, Familial, social, and individual factors contributing to risk for adolescent substance use, *Journal of addiction* **2013** (2013).
12. A. Giannoula, A. Gutierrez-Sacristán, Á. Bravo, F. Sanz and L. I. Furlong, Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study, *Scientific reports* **8**, 1 (2018).
13. C. M. Karch, A. T. Jeng, P. Nowotny, J. Cady, C. Cruchaga and A. M. Goate, Expression of novel alzheimer's disease risk genes in control and alzheimer's disease brains, *PloS one* **7**, p. e50976 (2012).
14. E. T. Dermitzakis, From gene expression to disease risk, *Nature genetics* **40**, 492 (2008).
15. A. M. Westerlund, J. S. Hawe, M. Heinig and H. Schunkert, Risk prediction of cardiovascular events by exploration of molecular data with explainable artificial intelligence, *International Journal of Molecular Sciences* **22**, p. 10291 (2021).
16. R. Perchard, T. Garner, P. G. Murray, A. Roslan, L. E. Higgins, E. D. Johnstone, A. Stevens and P. E. Clayton, A transcriptomic signature that predicts prehypertension in adolescence and higher systolic blood pressure in childhood, *JCI insight* **10** (2025).
17. Y. Zhao, M. G. Savelieff, X. Li, K. Guo, K. Wang, M. Li, B. Li, G. Iyer, S. A. Sakowski, L. Zhao *et al.*, Gene expression signatures from whole blood predict amyotrophic lateral sclerosis case status and survival, *Nature Communications* **16**, p. 9631 (2025).
18. C. Li, L. Qu, A. J. Matz, P. A. Murphy, Y. Liu, A. W. Manichaikul, D. Aguiar, S. S. Rich, D. M. Herrington, D. Vu *et al.*, Atherospectrum reveals novel macrophage foam cell gene signatures

associated with atherosclerotic cardiovascular disease risk, *Circulation* **145**, 206 (2022).
19. C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature machine intelligence* **1**, 206 (2019).
20. H. M. Natri, C. B. Del Azodi, L. Peter, C. J. Taylor, S. Chugh, R. Kendle, M.-i. Chung, D. K. Flaherty, B. K. Matlock, C. L. Calvi *et al.*, Cell-type-specific and disease-associated expression quantitative trait loci in the human lung, *Nature Genetics* **56**, 595 (2024).
21. G. Consortium, The GTEx consortium atlas of genetic regulatory effects across human tissues, *Science* **369**, 1318 (2020).
22. A. Baccarelli, D. C. Dolinoy and C. L. Walker, A precision environmental health approach to prevention of human disease, *Nature communications* **14**, p. 2449 (2023).
23. L. Maitre, M. Bustamante, C. Hernández-Ferrer, D. Thiel, C.-H. E. Lau, A. P. Siskos, M. Vives-Usano, C. Ruiz-Arenas, D. Pelegrí-Sisó, O. Robinson *et al.*, Multi-omics signatures of the human early life exposome, *Nature communications* **13**, p. 7024 (2022).
24. T. S. Sumida, S. Dulberg, J. C. Schupp, M. R. Lincoln, H. A. Stillwell, P.-P. Axisa, M. Comi, A. Unterman, N. Kaminski, A. Madi *et al.*, Type i interferon transcriptional network regulates expression of coinhibitory receptors in human t cells, *Nature Immunology* **23**, 632 (2022).
25. C. Li, A. Menoret, C. Farragher, Z. Ouyang, C. Bonin, P. Holvoet, A. T. Vella and B. Zhou, Single-cell transcriptomics–based macspectrum reveals macrophage activation signatures in diseases, *JCI insight* **4**, p. e126453 (2019).
26. J. Lindstrom and J. Tuomilehto, The diabetes risk score: a practical tool to predict type 2 diabetes risk, *Diabetes care* **26**, 725 (2003).
27. Y. Edlitz and E. Segal, Prediction of type 2 diabetes mellitus onset using logistic regression-based scorecards, *Elife* **11**, p. e71862 (2022).
28. M. R. Atreya, S. Banerjee, A. J. Lautz, M. N. Alder, B. M. Varisco, H. R. Wong, J. A. Muszynski, M. W. Hall, L. N. Sanchez-Pinto, R. Kamaleswaran *et al.*, Machine learning-driven identification of the gene-expression signature associated with a persistent multiple organ dysfunction trajectory in critical illness, *EBioMedicine* **99** (2024).
29. S. Shi, X. Pan, L. Zhang, X. Wang, Y. Zhuang, X. Lin, S. Shi, J. Zheng and W. Lin, An application based on bioinformatics and machine learning for risk prediction of sepsis at first clinical presentation using transcriptomic data, *Frontiers in Genetics* **13**, p. 979529 (2022).
30. N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr and J. M. O'Sullivan, A review of feature selection methods for machine learning-based disease risk prediction, *Frontiers in Bioinformatics* **2**, p. 927312 (2022).
31. H. M. Levitin, J. Yuan, Y. L. Cheng, F. J. Ruiz, E. C. Bush, J. N. Bruce, P. Canoll, A. Iavarone, A. Lasorella, D. M. Blei *et al.*, De novo gene signature identification from single-cell RNA-seq with hierarchical poisson factorization, *Molecular systems biology* **15**, p. MSB188557 (2019).
32. F. Buettner, N. Pratanwanich, D. J. McCarthy, J. C. Marioni and O. Stegle, f-scLVM: scalable and versatile factor analysis for single-cell rna-seq, *Genome biology* **18**, p. 212 (2017).
33. W. Mao, E. Zaslavsky, N. Bhardwaj, D. K. Bhatt, J. S. Tsang, M. Pellegrini, H. Hibshoosh, D. K. Bhatt, K. Hao, D. K. Bhatt *et al.*, Pathway-level information extractor (PLIER) for gene expression data, *Nature Methods* **16**, 607 (2019).
34. M. Lotfollahi, S. Rybakov, K. Hrovatin, S. Hediyeh-Zadeh, C. Talavera-López, A. V. Misharin and F. J. Theis, Biologically informed deep learning to query gene programs in single-cell atlases, *Nature Cell Biology* **25**, 337 (2023).
35. L. Seninge, I. Anastopoulos, H. Ding and J. Stuart, Vega is an interpretable generative model for inferring biological network activity in single-cell transcriptomics, *Nature communications* **12**, p. 5684 (2021).
36. G. Gut, S. G. Stark, G. Rätsch and N. R. Davidson, pmVAE: Learning interpretable single-cell representations with pathway modules, *BioRxiv* **01**, 2021 (2021).

37. R. Z. Kunes, T. Walle, M. Land, T. Nawy and D. Pe'er, Supervised discovery of interpretable gene programs from single-cell data, *Nature biotechnology* **42**, 1084 (2024).
38. D. Song, Q. Wang, G. Yan, T. Liu, T. Sun and J. J. Li, scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics, *Nature Biotechnology* **42**, 247 (2024).
39. A. T. Assefa, J. Vandesompele and O. Thas, Spsimseq: semi-parametric simulation of bulk and single-cell RNA-sequencing data, *Bioinformatics* **36**, 3276 (2020).
40. X. Ren, W. Wen, X. Fan, W. Hou, B. Su, P. Cai, J. Li, Y. Liu, F. Tang, F. Zhang *et al.*, COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas, *Cell* **184**, 1895 (2021).
41. D. M. Blei, A. Kucukelbir and J. D. McAuliffe, Variational inference: A review for statisticians, *Journal of the American statistical Association* **112**, 859 (2017).
42. N. G. Polson, J. G. Scott and J. Windle, Bayesian inference for logistic models using pólya–gamma latent variables, *Journal of the American statistical Association* **108**, 1339 (2013).
43. S. C. Dyer, O. Austine-Orimoloye, A. G. Azov, M. Barba, I. Barnes, V. P. Barrera-Enriquez, A. Becker, R. Bennett, M. Beracochea, A. Berry, J. Bhai, S. K. Bhurji, S. Boddu, P. R. Branco Lins, L. Brooks, S. B. Ramaraju, L. I. Campbell, M. C. Martinez, M. Charkhchi, L. A. Cortes, C. Davidson, S. Denni, K. Dodiya, S. Donaldson, B. El Houdaigui, T. El Naboulsi, O. Falola, R. Fatima, T. Genez, J. G. Martinez, T. Gurbich, M. Hardy, Z. Hollis, T. Hunt, M. Kay, V. Kaykala, D. Lemos, D. Lodha, N. Mathlouthi, G. A. Merino, R. Merritt, L. P. Mirabueno, A. Mushtaq, S. N. Hossain, J. G. Pérez-Silva, M. Perry, I. Piližota, D. Poppleton, I. Prosovetskaia, S. Raj, A. I. Salam, S. Saraf, N. Saraiva-Agostinho, S. Sinha, B. Sipos, V. Sitnik, E. Steed, M.-M. Suner, L. Surapaneni, K. Sutinen, F. F. Tricomi, I. Tsang, D. Urbina-Gómez, A. Veidenberg, T. A. Walsh, N. L. Willhoft, J. Allen, J. Alvarez-Jarreta, M. Chakiachvili, J. Cheema, J. B. daRocha, N. H. DeSilva, S. Giorgetti, L. Haggerty, G. R. Ilsley, J. Keatley, J. E. Loveland, B. Moore, J. M. Mudge, G. Naamati, J. Tate, S. J. Trevanion, A. Winterbottom, B. Flint, A. Frankish, S. E. Hunt, R. D. Finn, M. A. Freeberg, P. W. Harrison, F. J. Martin and A. D. Yates, Ensembl 2025, *Nucleic Acids Research* **53**, D948 (01 2025).
44. A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo and J. P. Mesirov, Molecular signatures database (msigdb) 3.0, *Bioinformatics* **27**, 1739 (2011).
45. M. Inouye, G. Abraham, C. P. Nelson, A. M. Wood, M. J. Sweeting, F. Dudbridge, F. Y. Lai, S. Kaptoge, M. Brozynska, T. Wang *et al.*, Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention, *Journal of the American College of Cardiology* **72**, 1883 (2018).
46. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell and J. J. Lee, Second-generation plink: rising to the challenge of larger and richer datasets, *Gigascience* **4**, s13742 (2015).
47. T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, (ACM, 2019).
48. A. J. Monteith and E. P. Skaar, The impact of metal availability on immune function during infection, *Trends in Endocrinology & Metabolism* **32**, 916 (2021).
49. C. C. Murdoch and E. P. Skaar, Nutritional immunity: the battle for nutrient metals at the host–pathogen interface, *Nature Reviews Microbiology* **20**, 657 (2022).
50. E. Galkina and K. Ley, Vascular adhesion molecules in atherosclerosis, *Arteriosclerosis, thrombosis, and vascular biology* **27**, 2292 (2007).
51. H. M. Nording, P. Seizer and H. F. Langer, Platelets in inflammation and atherogenesis, *Frontiers in immunology* **6**, p. 98 (2015).
52. B. HO-TIN-NOÉ, M. Demers and D. D. Wagner, How platelets safeguard vascular integrity, *Journal of Thrombosis and Haemostasis* **9**, 56 (2011).

53. A. P. Sage, D. Tsiantoulas, C. J. Binder and Z. Mallat, The role of b cells in atherosclerosis, *Nature Reviews Cardiology* **16**, 180 (2019).
54. B. Berger and H. Cho, Emerging technologies towards enhancing privacy in genomic data sharing, *Genome biology* **20**, 1 (2019).
55. M. Naveed, E. Ayday, E. W. Clayton, J. Fellay, C. A. Gunter, J.-P. Hubaux, B. A. Malin and X. Wang, Privacy in the genomic era, *ACM Computing Surveys (CSUR)* **48**, 1 (2015).