MULTIMODAL LATENT CAUSAL VAE FOR JOINT INFERENCE OF GENE REGULATORY AND PROTEIN INTERACTION NETWORKS

Shervin Goudarzi*
Department of Electrical Engineering and Computer Science
University of California, Berkeley

Nuo Liu
Goodarzi Lab, Arc Institute
Palo Alto, CA

Hani Goodarzi
Department of Biochemistry and Biophysics
University of California, San Francisco
Arc Institute

ABSTRACT

Learning directed causal relationships between genes and proteins from observational single-cell data remains a fundamental challenge, as correlation-based methods cannot distinguish cause from effect. We introduce **CrossModalCausalVAE**, a variational autoencoder that embeds a structural causal model (SCM) in the latent space, learning directed acyclic graphs (DAGs) within modalities and connected graphs across modalities, with linear decoders that enable interpretable projection of latent causal structure onto the observed gene and protein feature space. Trained on 809,130 PBMC cells from two CITE-seq cohorts spanning healthy and COVID-19 donors, the model learns four interpretable directed networks: intra-modal gene regulatory and protein interaction graphs, plus cross-modal translation and feedback matrices, constrained by biological priors from CollecTRI and OmniPath. *In silico* interventions confirm causal asymmetry in intra-modal and RNA→Protein blocks, with the learned cross-modal edges recovering central dogma directionality and transcription factor→target gene causal asymmetry. Critically, we zero-shot validate the learned causal structure against held-out Perturb-CITE-seq interventional data, achieving cross-modal RNA→Protein perturbation AUROC of 0.739 and gene-level direction accuracy of 69%, predicting not only which genes are affected by knockouts but whether they are upregulated or downregulated without direct regularization. Cross-referencing top-ranked causal pathways with published literature further confirms recovery of established immunological mechanisms. Our work provides a proof of concept that latent causal modeling paired with linear decoding enables interpretable and efficient causal graph inference in a multimodal setting.

1 INTRODUCTION

CITE-seq (Stoeckius et al., 2017) has proved invaluable in allowing for single-cell multimodal biological inference. However, much of the current work on multimodal representation learning is correlation- or factor-based. Gene regulatory networks (GRNs), protein-protein interactions (PPIs) and corresponding cross-modal networks are inherently *directed* and interconnected: transcription factors directly regulate gene expression, while translated proteins have causal effects on downstream transcription. Capturing this directionality from observational data is critical for mechanistic understanding.

Classical GRN inference methods such as SCENIC (Aibar et al., 2017) operate in observed gene space via correlations. DeepSEM (Shu et al., 2021) applies structural equation models (SEMs) directly to genes, but with thousands of features, the adjacency matrix requires millions of parameters

*Correspondence to: sgoudarzi@berkeley.edu

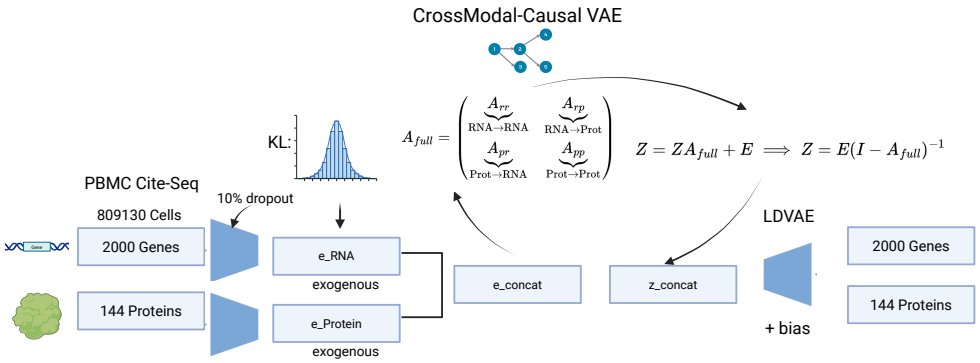


Figure 1: **CrossModal-Causal (CMC) VAE Architecture.** Residual encoders map RNA/Protein to exogenous noise. A Block-Structured SCM learns causal edges between latent factors. Linear decoders project causal graphs to feature space.

and may fail to capture complex multi-gene causal interactions. CausalVAE (Yang et al., 2021) addresses scalability and structural learning by learning causal structure in a latent space, but is limited to single-modality image data.

We propose **CrossModal-CausalVAE**, which learns causal structure in a compressed latent space while explicitly modeling intra-modal (RNA→RNA, Protein→Protein) and cross-modal (RNA↔Protein) relationships. While recent work has emphasized that causal discovery from purely observational data faces fundamental identifiability limitations, and that interventional data is critical for validation (Brouillard et al., 2020), our framework bridges this gap: the model is trained on observational CITE-seq data but validated against held-out interventional Perturb-CITE-seq perturbations. Our contributions are: (1) **formalizing multimodal causal discovery** for CITE-seq as a latent structural causal model (SCM) with block-structured DAG learning; (2) showcasing that **linear decoders with orthogonality** constraints enable exact projection of latent graphs to feature space for high interpretability; (3) **validating causal predictions** against ground-truth perturbational data; and (4) recovering known immunological mechanisms without supervision through the model's top causal edges.

2 Method

2.1 Generative Model

We model multimodal single-cell data with a three-stage generative process. First, modality-specific residual MLP encoders $q_\phi(\boldsymbol{\epsilon}|\mathbf{x})$ map RNA and protein observations to exogenous noise distributions $\boldsymbol{\epsilon}_r, \boldsymbol{\epsilon}_p$ (architecture details in Appendix A). Second, a **Block-Structured Structural Causal Model (SCM)** generates endogenous latent variables $\mathbf{z} = [\mathbf{z}_r; \mathbf{z}_p]$ via:

$$\mathbf{z} = (\mathbf{I} - \mathbf{A}^\top)^{-1}\boldsymbol{\epsilon}, \quad \text{where } \mathbf{A} = \begin{bmatrix} \mathbf{A}_{rr} & \mathbf{W}_{rp} \\ \mathbf{W}_{pr} & \mathbf{A}_{pp} \end{bmatrix} \tag{1}$$

Expanding the matrix inverse yields the per-modality structural equations. Decomposing Eq. 1, we arrive at $\mathbf{z}_r = \underbrace{\mathbf{z}_r\mathbf{A}_{rr}}_{\text{GRN}} + \underbrace{\mathbf{z}_p\mathbf{W}_{pr}}_{\text{Regulation}} + \boldsymbol{\epsilon}_r$ and $\mathbf{z}_p = \underbrace{\mathbf{z}_p\mathbf{A}_{pp}}_{\text{PPI}} + \underbrace{\mathbf{z}_r\mathbf{W}_{rp}}_{\text{Translation}} + \boldsymbol{\epsilon}_p$. Here $\mathbf{A}_{rr}, \mathbf{A}_{pp}$ are DAG-constrained intra-modal adjacency matrices (GRN and PPI), and $\mathbf{W}_{rp}, \mathbf{W}_{pr}$ are unconstrained cross-modal blocks capturing translation and regulatory feedback. Third, linear decoders reconstruct observations: $\hat{\mathbf{x}}_r = \mathbf{D}_r\mathbf{z}_r + \mathbf{b}_r$ and $\hat{\mathbf{x}}_p = \mathbf{D}_p\mathbf{z}_p + \mathbf{b}_p$. We learn the adjacency structure differentiably via the Gumbel-Sigmoid distribution $A_{ij} = \text{GumbelSigmoid}(\Theta_{ij}, \tau) \cdot W_{ij}$ with intra-modal acyclicity blocks enforced via the DAGMA constraint (Bello et al., 2022).

Proposition 1 (Latent-to-Feature Projection)**.** *Let the SCM be $\mathbf{z} = \mathbf{A}\mathbf{z} + \boldsymbol{\varepsilon}$ with linear decoding $\mathbf{x} = \mathbf{D}\mathbf{z} + \mathbf{b}$, where $\mathbf{D} \in \mathbb{R}^{n\times d}$, $n > d$. If $\mathbf{D}^\top\mathbf{D} \approx \mathbf{I}_d$ (enforced via orthogonality regularization), then the matrix of direct causal effects in feature space is $\mathbf{G} = \mathbf{D}\mathbf{A}\mathbf{D}^\top$.*

Proof of Proposition 1. A localized intervention $\mathrm{do}(\mathbf{z} \leftarrow \mathbf{z} + \delta\mathbf{z})$ propagates as $\Delta\mathbf{z} = \mathbf{A}\,\delta\mathbf{z}$, inducing $\Delta\mathbf{x} = \mathbf{D}\mathbf{A}\,\delta\mathbf{z}$ in observed space. Since $\mathbf{D}$ is full column rank, its Moore–Penrose pseudoinverse is $\mathbf{D}^{+} = (\mathbf{D}^{\top}\mathbf{D})^{-1}\mathbf{D}^{\top} \approx \mathbf{D}^{\top}$ under the orthogonality constraint. Substituting $\delta\mathbf{z} \approx \mathbf{D}^{\top}\delta\mathbf{x}$ gives $\Delta\mathbf{x} \approx \mathbf{D}\mathbf{A}\mathbf{D}^{\top}\,\delta\mathbf{x}$. The bias $\mathbf{b}$ cancels since $\Delta\mathbf{x}$ depends only on $\Delta\mathbf{z}$. □

2.2 Optimization

The full objective combines the multimodal ELBO with structural and biological constraints:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{rec}} + \beta\mathcal{L}_{\text{KL}}}_{\text{VAE}} + \underbrace{\lambda_{\text{DAG}}\mathcal{L}_{\text{DAGMA}} + \lambda_{\text{stable}}\mathcal{L}_{\rho} + \lambda_{\text{sparse}}\|\mathbf{A}\|_1}_{\text{Structure}} + \underbrace{\lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}} + \lambda_{\text{dec}}\|\mathbf{D}\|_1}_{\text{Decoder}} + \underbrace{\lambda_{\text{prior}}\mathcal{L}_{\text{prior}}}_{\text{Biological Prior}} \tag{2}$$

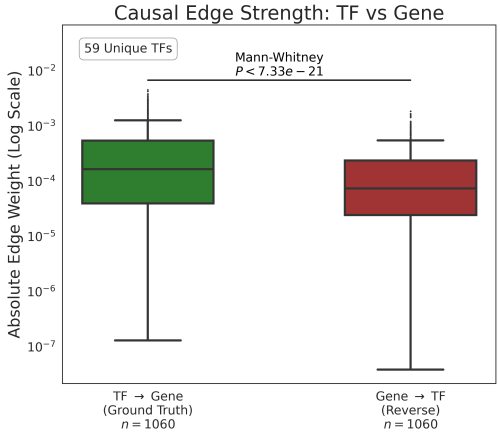
Here $\mathcal{L}_{\text{rec}}$ is MSE on z-scored data, $\mathcal{L}_{\text{KL}}$ is the VAE KL divergence, $\mathcal{L}_{\text{DAGMA}}$ enforces acyclicity (Bello et al., 2022), $\mathcal{L}_{\text{ortho}} = \|\mathbf{D}^{\top}\mathbf{D} - \mathbf{I}\|_F^2$ enforces orthogonality (Prop. 1), and $\mathcal{L}_{\text{prior}}$ is a margin ranking loss between edges present and absent in CollecTRI/OmniPath databases.

3 Experiments and Results

We train on a combined cohort of 809,130 PBMC CITE-seq cells (138 donors) pooled from Hao et al. (2021) (healthy PBMC) and Stephenson et al. (2021) (healthy + COVID-19), using 2,000 HVGs and 144 surface proteins, with latent dimensions $k_r = 128$, $k_p = 64$ (dataset, preprocessing, and hyperparameter grid search details in Appendices I and B).

Causal SCM Latents vs. Standard VAE. To isolate the contribution of the causal mechanism, we compare against a standard multimodal linear-decoder VAE where the SCM is disabled throughout training and evaluation ($\mathbf{z} = \boldsymbol{\epsilon}$, structural losses zeroed). For causal validation, we select the top 50 edges per block from the *full model's* learned adjacency and run interventions across the normalized exogenous errors on both models using the same edges, ensuring a fair paired comparison. We report the percentage of edges exceeding $2\times$ asymmetry and the geometric mean ratio. Results are shown in Figure 2.

Figure 2: CrossModal-CausalVAE vs. standard VAE (SCM disabled). (Left) Absolute projected edge weights for ground-truth TF→target pairs from CollecTRI vs. the reverse target→TF edges, across 1,060 pairs spanning 59 TFs. (Right, top) Latent space quality: $\mathbf{Z}$ concentrates cell-type identity while $\boldsymbol{\epsilon}$ approaches unstructured noise. (Right, bottom) Causal asymmetry on the same 50 top edges per block, evaluated on both models.



	Silhouette (↑)		Wasserstein (↓)	
Latent Space	**RNA**	**Prot**	**RNA**	**Prot**
CMC VAE $\mathbf{Z}$	**0.118**	**0.117**	2.307	3.405
CMC VAE $\boldsymbol{\epsilon}$	0.100	−0.012	**1.085**	**0.550**

	% Edges > 2× Asymmetry			
Causal Asym.	**R→R**	**R→P**	**P→P**	**P→R**
Standard VAE	18%	70%	34%	4%
CMC-VAE	**52%**	**90%**	**48%**	0%

Our SCM-projected DAG recovers asymmetry of absolute weights of TF → Gene well (Mann-Whitney $p < 7.3 \times 10^{-21}$). The SCM formulation disentangles biological signal from noise: $\mathbf{Z}$ achieves higher Silhouette scores than $\boldsymbol{\epsilon}$, confirming that the SCM concentrates causal identity into $\mathbf{Z}$, allowing the latent DAG $\mathbf{A}$ to learn while pushing $\boldsymbol{\epsilon}$ toward unstructured noise.

The SCM learns directionally asymmetric structure, with the strongest asymmetry in the RNA→Protein block, followed by intra-modal blocks (i.e, GRN/PPI). Consistent with the central

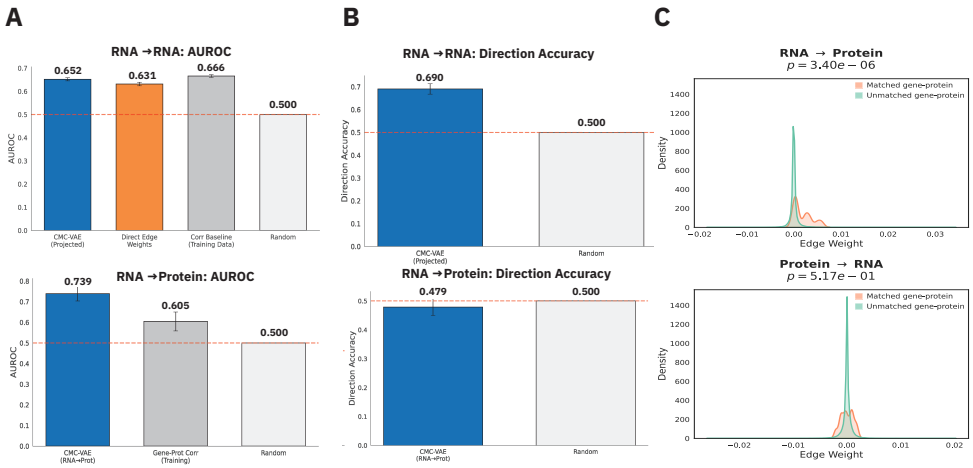


Figure 3: **Zero-shot perturbation validation and central dogma recovery.** (A) Held-out Perturb-CITE-seq AUROC: CMC-VAE matches the correlation baseline on RNA→RNA and outperforms it on RNA→Protein (0.739 vs. 0.605). (B) Direction accuracy: 69% on RNA→RNA, near chance on RNA→Protein, reflecting the smaller protein panel and lower per-feature signal. (C) Matched gene-protein edge weights are elevated above background in RNA→Protein ($p = 3.4 \times 10^{-6}$) but not Protein→RNA ($p = 0.52$), recovering central dogma directionality.

dogma, translation (RNA→Protein) is the dominant directional flow, while protein-to-transcriptome feedback is comparatively weak (see Appendix F). Importantly, cross-modal blocks have neither DAG penalties nor biological priors to enforce this asymmetry. In the same set of edges, the standard VAE achieves much lower overall asymmetry compared to CMC-VAE (Fig. 5 in Appendix E).

We validate against held-out Perturb-CITE-seq interventional data (Frangieh et al., 2021) via zero-shot *in silico* gene knockouts: each knockout is projected into exogenous noise space ($\delta\boldsymbol{\epsilon} \approx \mathbf{D}_r^\top \delta\mathbf{x}_r$), propagated through the SCM via $\mathbf{z}_{\text{pert}} = (\mathbf{I} - \mathbf{A}^\top)^{-1}(\boldsymbol{\epsilon} + \delta\boldsymbol{\epsilon})$, and decoded to predict genome-wide and proteome-wide effects (full methodology in Appendix J). CMC-VAE achieves RNA→RNA AUROC of 0.652 (37 perturbations), comparable to the training-data correlation baseline (0.666), but uniquely recovers the *sign* of perturbation effects with 69% direction accuracy, a capability correlation-based methods cannot provide. On cross-modal prediction, CMC-VAE attains RNA→Protein AUROC of 0.739 (36 perturbations), substantially exceeding the gene-protein correlation baseline of 0.605 (Figure 3A, B). Furthermore, in the RNA→Protein block, edge weights for matched gene-protein pairs (e.g., *CD4* mRNA → CD4 protein) are significantly elevated above unmatched background ($p = 3.4 \times 10^{-6}$), whereas no difference is seen in the Protein→RNA block ($p = 0.52$; Figure 3C), recovering central dogma directionality without supervision.

Biological Validation. Cross-referencing top inferred causal edges with published literature recovers known biology (Table 4 in Appendix H), including ZAP-70 immunological-synapse translocation (RNA) driving the CD28-dependent Vav1 pathway (Protein) in the RNA→Protein block, consistent with the canonical TCR→ZAP-70→Vav1 signaling cascade in T-cell activation, and CD22-mediated BCR regulation driving cross-presentation of soluble antigens via endosomes in the Protein→Protein block, consistent with the textbook mechanism of BCR–antigen internalization for MHC-II processing.

4 Conclusion

Identifying biological causality, especially through observational data, is notoriously difficult. Our model CrossModal-CausalVAE demonstrates that a latent-based block-structured SCM within a multimodal VAE can recover directed, interpretable regulatory networks from paired single-cell transcriptomic and proteomic data. Causal asymmetry analysis on top edges proves asymmetry of parent-to-child while demonstrating central dogma without regularization on the inter-modal graphs. Validation against held-out Perturb-CITE-seq interventions confirms that the causal structure learned by our latent-based model predicts real perturbational effects.

REFERENCES

Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, et al. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11): 1083–1086, 2017.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via m-matrices and a log-determinant acyclicity characterization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8226–8239, 2022.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21865–21877, 2020.

Chris J Frangieh, Johannes C Melms, Pratiksha I Thakore, et al. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nature Genetics*, 53 (3):332–341, 2021.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck III, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M Fleming, Bertrand Yeung, Angela J Rogers, Juliana M McElrath, Catherine A Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

Sophia Müller-Dott, Eirini Tsirvouli, Miguel Vazquez, Ricardo O Ramirez Flores, Pau Badia-i Mompel, Robin Fallegger, Dénes Türei, Astrid Lægreid, and Julio Saez-Rodriguez. Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Research*, 51(20):10934–10949, 2023.

Hantao Shu, Jingtian Zhou, Qiuyu Lian, Han Li, Dan Zhao, Jianyang Zeng, and Jianzhu Ma. Modeling gene regulatory networks using neural network architectures. *Nature Computational Science*, 1(7):491–501, 2021.

Emily Stephenson, Gary Reynolds, Rachel A Botting, Fernando J Calero-Nieto, Michael D Morgan, Zewen Kelvin Tuong, Karsten Bach, Waradon Sungnak, Kaylee B Worlock, Masahiro Yoshida, Natsuhiko Kumasaka, Krzysztof Kania, Justin Engelbert, Bayanne Olabi, Jarmila Stremenova Spegarova, Nicola K Wilson, Nicole Mende, Laura Jardine, Louis C S Gardner, Issac Goh, Dave Horsfall, Jim McGrath, Simone Webb, Michael W Mather, Rik G H Lindeboom, Emma Dann, Ni Huang, Krzysztof Polanski, Elena Prigmore, Florian Gothe, Jonathan Scott, Rebecca P Payne, Kenneth F Baker, Aidan T Hanrath, Ina C D Schim van der Loeff, Andrew S Barr, Amada Sanchez-Gonzalez, Laura Bergamaschi, Federica Mescia, Josephine L Barnes, Eleanor Kilich, Abigail de Wilton, Anika Saigal, et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine*, 27(5):904–916, 2021.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, 2017.

Dénes Türei, Tamás Korcsmáros, and Julio Saez-Rodriguez. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, 13(12):966–967, 2016.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021.

A MODEL ARCHITECTURE

Table 1: Architecture and training hyperparameters.

Component	Specification
RNA Encoder	MLP: 2000→512→256→128 (LeakyReLU, BN, Dropout 0.1)
Protein Encoder	MLP: 144→128→64→64 (LeakyReLU, BN, Dropout 0.1)
Latent RNA (k_r) / Protein (k_p)	128 / 64
Logit Initialization	$\mathcal{N}(0, 0.05)$
Weight Initialization	Xavier uniform
RNA Decoder	Linear: 128→2000 (no bias) + learnable bias
Protein Decoder	Linear: 64→144 (no bias) + learnable bias
Input Dropout (RNA, training only)	10%
SCM Activation	Smooth α-ramp: 0→1 over epochs 20–50

Each encoder $q_\phi(\boldsymbol{\epsilon}|\mathbf{x})$ uses a residual MLP with batch normalization:

$$\mathbf{h} = \text{LeakyReLU}(\text{BN}(\mathbf{W}_1\mathbf{x})), \quad \boldsymbol{\mu}_\epsilon, \boldsymbol{\sigma}_\epsilon = \text{Linear}(\mathbf{h} + \text{LeakyReLU}(\text{BN}(\mathbf{W}_2\mathbf{h}))) \tag{3}$$

The structural equations decompose into intra-modal and cross-modal interactions:

$$\mathbf{z}_r = \mathbf{A}_{rr}^\top \mathbf{z}_r + \mathbf{W}_{pr}^\top \mathbf{z}_p + \boldsymbol{\epsilon}_r, \quad \mathbf{z}_p = \mathbf{A}_{pp}^\top \mathbf{z}_p + \mathbf{W}_{rp}^\top \mathbf{z}_r + \boldsymbol{\epsilon}_p \tag{4}$$

B TRAINING DETAILS AND SCHEDULE

Training uses a phased schedule. During warmup (epochs 0–20), the SCM is bypassed ($\mathbf{z} = \boldsymbol{\epsilon}$), allowing the VAE to learn valid representations. From epoch 20, the SCM activates and structural constraints are annealed.

Table 2: Loss weight schedule across training phases.

Parameter	Warmup (Ep. 0–20)	Annealing (Ep. 20–100)	Final
λ_{rec}	100.0	100.0	100.0
$\beta_{\text{RNA}}, \beta_{\text{Prot}}$ (KL)	10^{-4}	$10^{-4} \to 4.0$	4.0
α (SCM interpolation)	0.0	$0 \to 1.0$ (30-epoch ramp)	1.0
λ_{DAG} (Acyclicity)	0.0	$0 \to 5.0$	5.0
λ_{prior} (CollecTRI/OmniPath)	0.0	$0 \to 500.0$	500.0
λ_{stable} (Spectral Radius)	1.0	1.0	1.0
λ_{sparse} (intra/inter)	0.01	0.01	0.01
λ_{ortho}	5.0	5.0	5.0
$\lambda_{\text{dec,RNA}}$ (feature-norm L1)	1000	1000	1000
$\lambda_{\text{dec,Prot}}$ (feature-norm L1)	18	18	18
Gumbel τ	1.0	$1.0 \to 0.5$ (switch at ep. 30)	0.5

Loss components: Reconstruction uses MSE on z-scored data: $\mathcal{L}_{\text{rec}} = \|\mathbf{x}_r - \hat{\mathbf{x}}_r\|_2^2 + \|\mathbf{x}_p - \hat{\mathbf{x}}_p\|_2^2$. Acyclicity uses DAGMA: $h(\mathbf{A}) = \text{ReLU}(-\log\det(s\mathbf{I} - \mathbf{A} \odot \mathbf{A}) + d\log s)$ with $s = \rho(\mathbf{A} \odot \mathbf{A}) + 1$. Stability: $\mathcal{L}_{\text{stable}} = \text{ReLU}(\rho(\mathbf{A}_{\text{full}}) - 0.9)$. Orthogonality: $\mathcal{L}_{\text{ortho}} = \|\mathbf{D}^\top\mathbf{D} - \mathbf{I}\|_F^2/d^2$ (normalized). Decoder sparsity uses per-modality feature-normalized L1: $\|\mathbf{D}\|_1/n_{\text{features}}$. Prior: MarginRankingLoss between projected edge weights for known vs. random background edges, with margin $m = 0.1$. The SCM is activated via a smooth interpolation: $\mathbf{z} = (1-\alpha)\boldsymbol{\epsilon} + \alpha \cdot (\mathbf{I} - \mathbf{A}^\top)^{-1}\boldsymbol{\epsilon}$, where α ramps from 0 to 1 over 30 epochs after warmup, avoiding the abrupt activation used in earlier versions.

B.1 HYPERPARAMETER GRID SEARCH

The configuration in Table 2 was selected from a grid search over four key hyperparameters, with all other values held fixed. Model selection was performed jointly on held-out validation reconstruction loss and zero-shot perturbation AUROC on a held-out subset of perturbations. Table 3 lists the swept and fixed ranges.

Table 3: Hyperparameter grid search and fixed parameters.

Parameter	**Values Swept**	**Selected**	**Role**
β (KL weight)	{1.0, 2.0, 3.0, 4.0, 5.0}	**4.0**	Latent information richness
$\lambda_{\text{dec,RNA}}$	{500, 1000, 2000, 4000}	**1000**	Feature-norm L1 on RNA decoder
λ_{ortho}	{5.0, 15.0}	**5.0**	Decoder column orthogonality
$\lambda_{\text{prior}}^{\text{final}}$	{500, 1000}	**500**	Biological prior strength (annealed)
Fixed across all runs			
$\lambda_{\text{dec,Prot}}$	18		Feature-norm L1 on protein decoder
λ_{rec}	100.0		Reconstruction weight
$\lambda_{\text{DAG}}^{\text{final}}$	5.0		DAGMA acyclicity weight
λ_{sparse} (intra/inter)	0.01		L1 sparsity on SCM blocks
Latent dims (k_r, k_p)	(128, 64)		Latent dimensions
Batch size	256		—
Learning rate	10^{-3} (AdamW)		—
Warmup epochs	20		VAE-only warmup (SCM bypassed)
α-ramp epochs	30		Smooth SCM activation schedule
Gumbel τ	$1.0 \to 0.5$		Annealed across training
Decoder sparsity mode	feature-norm		Per-feature L1 normalization
`prior_norm`	False		No prior weight normalization

C ALGORITHM

Algorithm 1 CrossModal-CausalVAE Training

Require: Data $\{(\mathbf{x}_r^{(i)}, \mathbf{x}_p^{(i)})\}_{i=1}^N$, Priors $\mathbf{M}_{\text{GRN}}, \mathbf{M}_{\text{PPI}}$
Initialize encoders, SCM parameters $\{\boldsymbol{\Theta}, \mathbf{W}\}$, decoders $\mathbf{D}_r, \mathbf{D}_p$
for epoch $t = 1$ to T **do**
  Anneal τ; compute $\alpha(t)$, $\beta(t)$, $\lambda_{\text{DAG}}(t)$, $\lambda_{\text{prior}}(t)$ via schedule (Table 2)
  for minibatch $(\mathbf{X}_r, \mathbf{X}_p)$ **do**
    Encode: $\boldsymbol{\mu}_r, \boldsymbol{\sigma}_r \leftarrow \text{Enc}_r(\mathbf{X}_r)$; $\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p \leftarrow \text{Enc}_p(\mathbf{X}_p)$
    Sample: $\boldsymbol{\epsilon} \leftarrow [\boldsymbol{\mu}_r + \boldsymbol{\sigma}_r \odot \boldsymbol{\eta}_r; \boldsymbol{\mu}_p + \boldsymbol{\sigma}_p \odot \boldsymbol{\eta}_p]$
    if $t \leq 20$ (warmup: α=0, λ_{DAG}=0, λ_{prior}=0) **then**
      $\mathbf{z} \leftarrow \boldsymbol{\epsilon}$ {SCM bypassed; train VAE only}
    else
      $\mathbf{A} \leftarrow \text{AssembleBlocks}(\text{GumbelSigmoid}(\boldsymbol{\Theta}, \tau))$
      $\mathbf{z} \leftarrow (1-\alpha)\boldsymbol{\epsilon} + \alpha(\mathbf{I} - \mathbf{A}^\top)^{-1}\boldsymbol{\epsilon}$ {SCM activates; DAG/prior losses ramp}
    end if
    Decode: $\hat{\mathbf{X}}_r \leftarrow \mathbf{D}_r\mathbf{z}_r + \mathbf{b}_r$; $\hat{\mathbf{X}}_p \leftarrow \mathbf{D}_p\mathbf{z}_p + \mathbf{b}_p$
    Compute $\mathcal{L}_{\text{total}}$ (Eq. 2), backpropagate, update
  end for
end for
Output: $\mathbf{G}_{\text{GRN}} = \mathbf{D}_r\mathbf{A}_{rr}\mathbf{D}_r^\top$, $\mathbf{G}_{\text{PPI}} = \mathbf{D}_p\mathbf{A}_{pp}\mathbf{D}_p^\top$

D LOSS FUNCTION CURVES

Figure 4 shows the training and validation curves for all nine loss components across the full 100-epoch schedule. Three dynamics are notable.

Phase transition at epoch 20. The reconstruction, KL-divergence, DAGMA, sparsity, and prior losses all exhibit a sharp discontinuity at epoch 20, when the SCM is activated and the structural regularizers begin annealing in. Reconstruction briefly spikes as the model adapts from $\mathbf{z} = \boldsymbol{\epsilon}$ to $\mathbf{z} = (\mathbf{I} - \mathbf{A}^\top)^{-1}\boldsymbol{\epsilon}$, but recovers within $\sim$5 epochs and continues decreasing for the remainder of training, indicating that the SCM does not impair the model's ability to fit the data.

KL collapse is structural, not pathological. The RNA KL divergence spikes near epoch 20 and then collapses toward zero as training progresses, while the protein KL decays smoothly from $\sim$9 to $\sim$1.5. This is the expected behavior under the SCM formulation: once $\mathbf{A}$ absorbs causal structure, the encoder is free to push $\boldsymbol{\epsilon}$ toward the standard-normal prior, and the residual KL reflects only the noise component of the latent. Crucially, the per-dimension standard deviation of $\boldsymbol{\epsilon}$ remains close to 1.0 throughout (Figure 2, top-right table: Wasserstein distance to $\mathcal{N}(0, 1)$ of 1.085 for RNA and 0.550 for protein), confirming that $\boldsymbol{\epsilon}$ is approaching genuine noise rather than collapsing to zero.

Structural penalties anneal cleanly. The DAGMA acyclicity loss spikes at epoch 20 (when the DAG penalty turns on) and then decays smoothly toward zero, indicating successful enforcement of the acyclicity constraint on $\mathbf{A}_{rr}$ and $\mathbf{A}_{pp}$. The biological prior loss similarly spikes briefly and then drops to near-zero, indicating that known CollecTRI/OmniPath edges quickly come to dominate over background edges in projected weight magnitude. Decoder sparsity and weight orthogonality losses are largely stable across training, with orthogonality remaining below 0.018 throughout, justifying the $\mathbf{D}^\top\mathbf{D} \approx \mathbf{I}$ assumption used in Proposition 1. Train and validation curves track each other closely across all nine panels, indicating no overfitting.

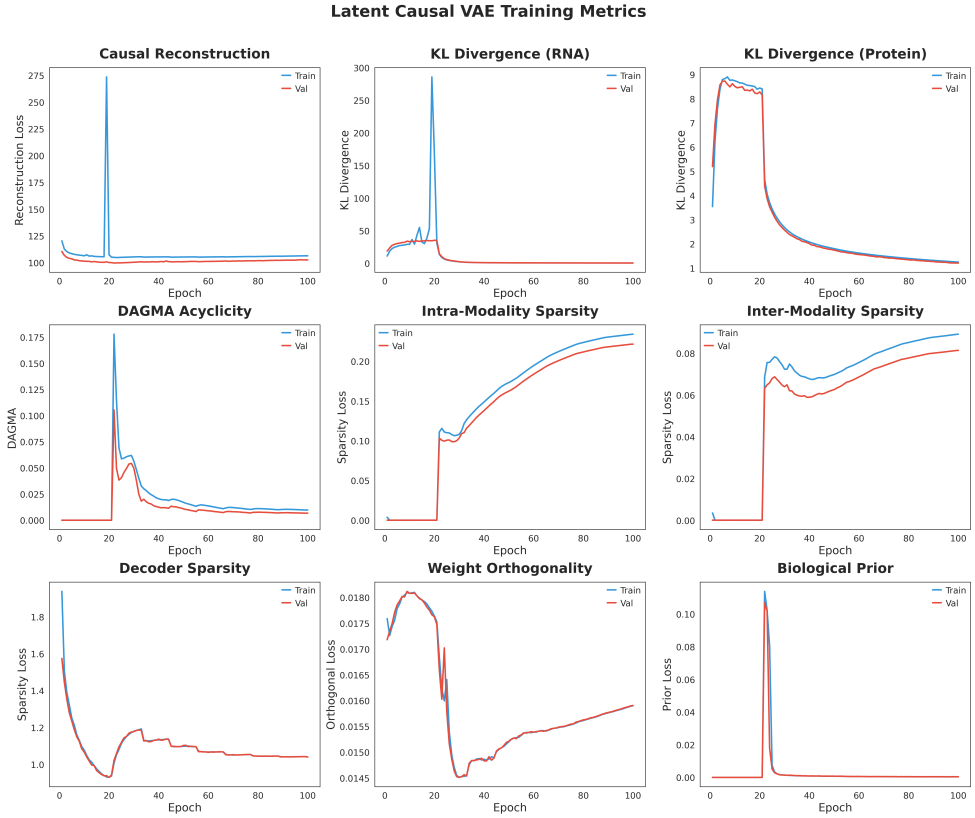
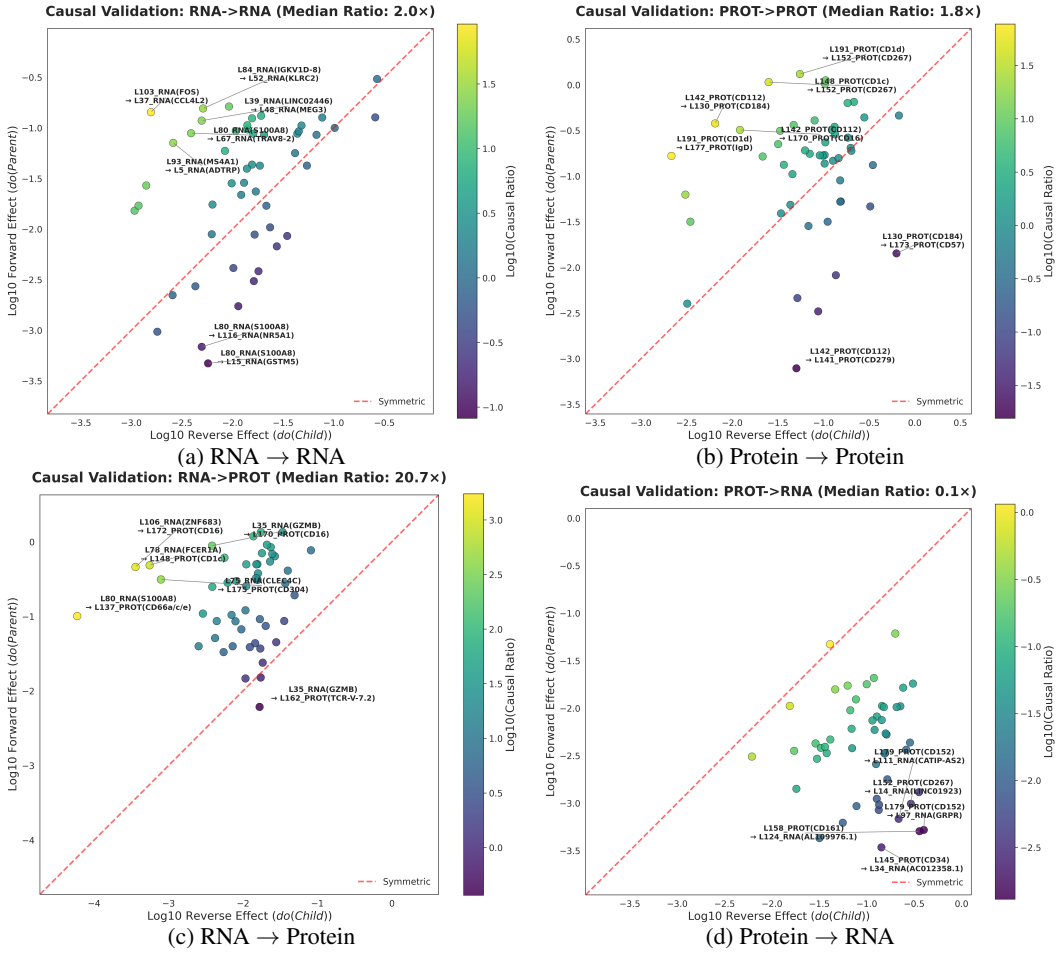


Figure 4: **Loss function curves across 100 epochs.** All nine training and validation loss components shown; train (blue) and validation (red) tracks. The vertical discontinuity near epoch 20 marks SCM activation and the onset of structural regularizer annealing.

E CAUSAL VALIDATION PLOTS



(a) RNA → RNA (b) Protein → Protein

(c) RNA → Protein (d) Protein → RNA

Figure 5: **Causal Validation.** Log-log plots of forward vs. reverse intervention effects. Points off the diagonal indicate causal asymmetry. Color denotes Causal Ratio magnitude.

F PROTEIN→RNA ASYMMETRY AND THE CENTRAL DOGMA

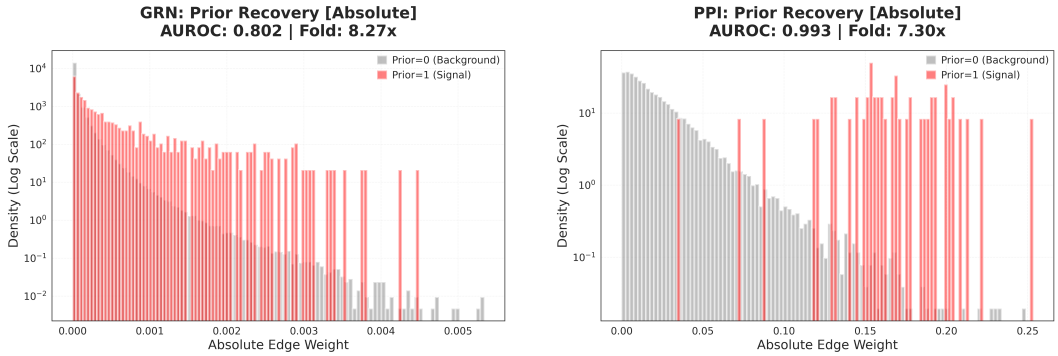
The Protein→RNA block exhibits negligible forward asymmetry, in stark contrast to the RNA→Protein block. Rather than indicating model failure, we propose that this asymmetry pattern is consistent with the central dogma of molecular biology: the primary directional flow of genetic information proceeds from DNA→RNA→Protein, with protein-to-transcriptome feedback representing a secondary, weaker regulatory mode.

In the learned SCM, the cross-modal blocks $\mathbf{W}_{rp}$ (RNA→Protein) and $\mathbf{W}_{pr}$ (Protein→RNA) are unconstrained with neither DAG penalties nor biological priors to enforce this directional asymmetry. The model independently learns that interventions on RNA latents propagate strongly to protein latents (translation), while interventions on protein latents propagate weakly to RNA latents (feedback regulation).

The standard VAE baseline shows a qualitatively similar pattern (70% RNA→Protein, 4% Protein→RNA), suggesting this asymmetry reflects genuine biological signal in the data rather than an artifact of the SCM formulation. However, the CMC-VAE substantially amplifies the intra-modal asymmetry (52% vs. 18% for RNA→RNA, and 48% vs. 34% for Protein→Protein), confirming that the SCM mechanism contributes causal structure beyond what a standard encoder-decoder captures.

G BIOLOGICAL PRIOR RECOVERY

We evaluate how well the inferred GRN and PPI networks recover known interactions from the biological priors used during training. For the GRN, we measure the distribution of projected edge weights $|\mathbf{G}_{\mathrm{GRN}}|$ for edges present vs. absent in the CollecTRI transcription factor–target interaction mask (Müller-Dott et al., 2023). For the PPI, we compare against OmniPath protein–protein interactions (Türei et al., 2016). During training, priors guide structure via a margin ranking loss that encourages known edges to have higher projected weights than random background edges.



(a) GRN: CollecTRI edges vs. background

(b) PPI: OmniPath edges vs. background

Figure 6: **Prior Recovery.** Distribution of projected edge weights for edges present vs. absent in biological prior databases. Known interactions exhibit higher projected weights, confirming that the prior guidance is reflected in the learned graph structure, with the GRN having lower absolute weights due to its larger size.

H EXTENDED BIOLOGICAL VALIDATION

Table 4 lists top causal edges from each of the four SCM blocks, annotated via Enrichr pathway enrichment of decoder weight loadings and selected by interventional ratio ($E_{\mathrm{fwd}}/E_{\mathrm{rev}}$). Edges marked ($\star$) are discussed in the main text.

Table 4: Top causal edges across all four SCM blocks, annotated via Enrichr pathway enrichment of decoder weight loadings and ranked by interventional ratio ($E_{\mathrm{fwd}}/E_{\mathrm{rev}}$). Reactome pathway identifiers are retained for provenance. Edges marked ($\star$) are discussed in the main text.

Block	Edge (Source → Target)	Ratio
RNA → RNA		
	Transcription (R-HSA-9031628) → Interleukin Signaling (R-HSA-6783783)	91.8
	CD22-mediated BCR Regulation (R-HSA-5690714) → Lymphoid–Non-Lymphoid Interactions (R-HSA-198933)	30.4
	Antigen-activated BCR / Second Messengers (R-HSA-983695) → Cytokine Signaling (R-HSA-1266695)	27.5
	Immune System (R-HSA-168256) → Signaling by Interleukins (R-HSA-449147)	22.8
RNA → Protein		
	Immune System (R-HSA-168256) → Chemokine Receptors Bind Chemokines (R-HSA-380108, Prot)	1731.2
	Lymphoid–Non-Lymphoid Interactions (R-HSA-198933) → Non-Canonical NF-κB Pathway (R-HSA-5668541, Prot)	1280.4
$\star$	Translocation of ZAP-70 to Imm. Synapse (R-HSA-202430) → CD28-dep. Vav1 Pathway (R-HSA-389359, Prot)	868.8
	IL-5 / GM-CSF Signaling (R-HSA-512988) → Chemokine Receptors Bind Chemokines (R-HSA-380108, Prot)	395.6
	Peptide Chain Elongation (R-HSA-156902) → Butyrophilin (BTN) Family Interactions (R-HSA-8851680, Prot)	36.3
Protein → Protein		
	CD28-dep. Vav1 Pathway (R-HSA-389359, Prot) → Lymphoid–Non-Lymphoid Interactions (R-HSA-198933, Prot)	77.3
$\star$	CD22-mediated BCR Regulation (R-HSA-5690714, Prot) → Cross-presentation via Endosomes (R-HSA-1236978, Prot)	58.3
	CD22-mediated BCR Regulation (R-HSA-5690714, Prot) → Chemokine Receptors Bind Chemokines (R-HSA-380108, Prot)	26.4
	CD22-mediated BCR Regulation (R-HSA-5690714, Prot) → CD103/CD235ab	10.6
Protein → RNA		
	Nuclear Events Downstream of ALK Signaling (R-HSA-9701898, Prot) → Peptide Chain Elongation (R-HSA-156902)	1.15
	CD163-mediated Anti-Inflammatory Response (R-HSA-9662834, Prot) → Hemostasis (R-HSA-109582)	0.68

The two starred edges highlight canonical immunological cascades recovered without supervision. The ZAP-70 → Vav1 axis mirrors the well-characterized T-cell receptor signaling pathway, in which

ZAP-70, upon translocation to the immunological synapse, phosphorylates LAT and SLP-76 to recruit Vav1 as a guanine-nucleotide exchange factor for Rac/Rho GTPases, driving actin-cytoskeleton reorganization and synapse formation. The CD22-mediated BCR regulation → endosomal cross-presentation edge reflects the textbook mechanism by which BCR–antigen complexes are internalized into endosomes for processing and MHC-II-mediated presentation to helper T cells, with CD22 serving as an inhibitory co-receptor that tunes the internalization threshold.

Consistent with the central-dogma interpretation in Appendix F, the Protein→RNA block shows no edges exceeding $2\times$ asymmetry. The top Protein→RNA edge (ALK-signaling nuclear events → peptide chain elongation, ratio ≈ 1.15) is nominally consistent with known translational regulation downstream of receptor tyrosine kinase signaling, but its magnitude is negligible relative to forward translation edges in the RNA→Protein block ($\sim$1000$\times$ weaker), confirming that protein→transcriptome feedback is a minor mode in the learned model.

I Dataset Details

We train on a combined cohort pooled from two large-scale public PBMC CITE-seq datasets spanning healthy and COVID-19 conditions, totaling 809,130 cells from 138 donors (Table 5).

Table 5: Training cohort composition.

Dataset	**Cells**	**Donors**	**Condition**	**Platform**
Hao et al. (2021)	161,764	8	Healthy PBMC	10x 5′ CITE-seq
Stephenson et al. (2021)	647,366	130	COVID-19 + Healthy	10x 5′ CITE-seq
Combined total	**809,130**	**138**	Mixed	—

Table 6: Modality-specific preprocessing and feature overlap strategy.

Modality	**# Features**	**Preprocessing**	**Overlap Rule**
RNA (transcriptome)	2,000 HVGs	Library norm (10^4) → $\log(x{+}1)$ → HVG selection → z-score	Common genes across datasets
Protein (surface)	144 proteins	CLR transform via `muon` → z-score	Mapped via AB→CD nomenclature

Split and priors. Cells are split 80/10/10 (train/val/test) at random with seed 42. Donor IDs (138 levels) are used as batch labels for batch correction. Biological priors consist of 1,604 TF→target interactions from CollecTRI via OmniPath for the GRN (Müller-Dott et al., 2023), and 144 directed protein-protein interactions from OmniPath for the PPI (Türei et al., 2016); priors enter training via a margin ranking loss (Appendix B).

J Perturbation Validation Methodology

We zero-shot validate learned causal structure against held-out Perturb-CITE-seq interventions (Frangieh et al., 2021), which are *not* used at any point during training. The model must predict the effects of single-gene CRISPR knockouts it has never seen.

Perturbation data preprocessing. sgRNA labels are parsed to target gene symbols, with control cells identified by substring matching (`NO_SITE`, `NON-GENE`, `NON_TARGET`, `SCRAMBLE`). Perturbations with fewer than 30 cells are discarded. Genes are aligned to the training RNA feature space by symbol; surface proteins are aligned after stripping conjugation suffixes (e.g., `_TotalSeqB/C`, `_isotype`) and mapping AB→CD nomenclature to handle antibody naming mismatches between cohorts.

Ground-truth differential expression. For each perturbation g with perturbed cells $\mathcal{P}_g$ and control cells $\mathcal{C}$, we compute per-feature log fold-change $\Delta_j = \bar{\mathrm{x}}_j^{\mathcal{P}_g} - \bar{\mathrm{x}}_j^{\mathcal{C}}$, two-sided Mann-Whitney U-test

p-values, and a Cohen's-d-like effect size

$$e_j \;=\; \frac{|\Delta_j|}{\sqrt{\big(s_j^{\mathcal{P}_g\,2} + s_j^{\mathcal{C}\,2}\big)/2}}. \tag{5}$$

Features with $p_j < 0.05$ are called differentially expressed (DE).

***In silico* intervention.** For a knockout of gene g, we first encode up to 20,000 control cells through the learned encoders to obtain their exogenous noise posterior means $(\boldsymbol{\mu}_{\epsilon_r}, \boldsymbol{\mu}_{\epsilon_p})$. The intervention is represented in observed RNA space as $\delta\mathbf{x}_r \in \mathbb{R}^{n_r}$ with

$$(\delta\mathbf{x}_r)_j \;=\; \begin{cases} -\hat{x}_{r,g} & j = g \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

i.e., we zero out the reconstructed expression of the target gene. By Proposition 1, this observed-space intervention maps into latent exogenous space via $\delta\boldsymbol{\epsilon}_r \approx \mathbf{D}_r^\top \delta\mathbf{x}_r$, with $\delta\boldsymbol{\epsilon}_p = \mathbf{0}$. Importantly, we intervene on the *exogenous noise* rather than on the endogenous latent $\mathbf{z}$ directly: this preserves the full downstream DAG propagation path, whereas directly clamping $\mathbf{z}$ would zero out the corresponding column of $(\mathbf{I} - \mathbf{A}^\top)^{-1}$ and prevent the intervention from propagating through the learned causal structure. The perturbed latent state is then

$$\mathbf{z}_{\text{pert}} \;=\; (\mathbf{I} - \mathbf{A}^\top)^{-1}\big(\boldsymbol{\epsilon} + \delta\boldsymbol{\epsilon}\big), \qquad \boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_r; \boldsymbol{\epsilon}_p], \quad \delta\boldsymbol{\epsilon} = [\delta\boldsymbol{\epsilon}_r; \mathbf{0}], \tag{7}$$

and the predicted per-cell perturbation effect is

$$\Delta\hat{\mathbf{x}} \;=\; \mathbf{D}\big(\mathbf{z}_{\text{pert}} - \mathbf{z}_{\text{ctrl}}\big). \tag{8}$$

Predictions are averaged across the control cell population to obtain a population-level perturbation signature $\overline{\Delta\hat{\mathbf{x}}}$.

Evaluation metrics. Given predicted effects $\{\overline{\Delta\hat{x}}_j\}$ and ground-truth $\{\Delta_j, e_j, p_j\}$ for held-out features j (excluding the target gene itself), we report, per perturbation:

- **AUROC**: ranking DE features ($p_j < 0.05$) against non-DE features using $|\overline{\Delta\hat{x}}_j|$ as the score (computed only when ≥ 3 DE features exist, for RNA; ≥ 1 for protein, reflecting the smaller protein panel).
- **Spearman ρ (effect)**: rank correlation between $|\overline{\Delta\hat{x}}_j|$ and e_j.
- **Spearman ρ (signed)**: rank correlation between $\overline{\Delta\hat{x}}_j$ and Δ_j.
- **Direction accuracy**: on the set $\mathcal{S}$ of significantly DE features, $\frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \mathbb{1}\big[\text{sign}(\overline{\Delta\hat{x}}_j) = \text{sign}(\Delta_j)\big]$. This metric is unique to causal models: correlation baselines are unsigned and cannot predict direction.

We aggregate across perturbations by taking the mean; error bars in Figure 3A denote SEM.

Baselines.

1. *Direct edge weights*: predicted effect = $|\mathbf{G}_{\text{GRN}}[g,:]|$, the model's projected feature-space adjacency without SCM propagation. This tests whether latent-space propagation adds value beyond reading off the inferred graph.
2. *Correlation baseline (training)*: $|\text{Pearson}(x_g, x_j)|$ computed on the full training cohort. Because CMC-VAE is trained on this same cohort and evaluated zero-shot on Perturb-CITE-seq, this is the fair correlational comparator.
3. *Gene-protein correlation baseline (training)*: the cross-modal analogue $|\text{Pearson}(x_{r,g}, x_{p,j})|$ from training data, used for RNA→Protein evaluation.

We deliberately do not use correlations computed from Perturb-CITE-seq control cells as the primary baseline, as that would give the baseline access to in-distribution information that CMC-VAE lacks.