Cell-Level Virtual Screening

Anonymous Authors[1]

Abstract

Virtual screening methods prioritize therapeutic candidates by predicting molecular properties and interactions. However, molecular models are insufficient to predict higher-order effects that arise in real biological systems. This blind spot leads to many late-stage failures in drug discovery. Virtual cells have been posed as a solution to this problem by predicting gene expression responses to drugs, but they remain weakly validated as screening tools; gene expression is only an intermediate in understanding drug success or failure. Despite burgeoning progress in virtual cells, some basic questions remain. Is expression even a good representation of higher-order drug effects? How can virtual cell methods be applied to prioritize therapeutic candidates? Can they be fairly compared against traditional molecular-level screens? We address these questions in a two-pronged approach. First, we curate two benchmarks that directly compare virtual cells against traditional molecular methods on canonical drug discovery tasks. Drug-Disease Bench evaluates a method's ability to prioritize disease indications for drugs with novel target profiles. Drug-Target Bench evaluates a method's ability to reconstruct drug-target interactions from separate perturbation modalities that act on shared mechanisms, bridging the gap between cell-level methods and classic molecular screens. We identify shortcomings of existing virtual cells on these benchmarks, and propose an alternative representation of cell state: gene networks. Inferring post-perturbation gene networks on-demand for unseen drugs requires methods that generalize beyond traditional plug-in network estimators. We develop a scalable differentiable surrogate loss for multivariate Gaussians, which we apply to train a context encoder that maps perturbation metadata to full gene-gene dependency network parameters. The resulting model, CellVS-Net, achieves SOTA on predicting how gene-gene networks restructure under a variety of complex multivariate experimental conditions, including different cell types, small molecules, large molecules, gene knockdowns, and gene overexpressions. When compared to other molecular and cell-level representations of drugs, we find that CellVS-Net achieves SOTA on both virtual screening benchmarks. Overall, CellVS-Net provides the first demonstration that cell-level virtual screening methods are a viable alternative to molecular screening, and associated benchmarks enable future hill-climbing on clinically relevant tasks. We provide source code for models and data curation, as well as public leaderboards.

Introduction

Despite continuous improvements in virtual screening for molecular interactions, recently achieving near-instantaneous proteome-scale screens (McNutt et al.), virtual screening approaches still remain blind to emergent failures at the level of cellular systems. Moving beyond molecular interactions, large-scale expression and morphology profiling efforts such as LINCS L1000 (Koleti et al.), Tahoe-100M (Zhang et al., 2025), scPerturb (Peidli et al.), JUMP-CP (Chandrasekaran et al.), and Recursion's phenomics platform (Bray et al.) aim to support phenotype-level screens by experimentally measuring cell states under many perturbations and ranking candidates by similarity or reversal of disease signatures. However, these phenotypic screens require running a new assay for each candidate drug and are therefore limited by experimental throughput and design. In contrast, virtual screening aims to predict cell-level responses for unseen perturbations based only on their molecular or target features, enabling in silico ranking of large candidate libraries before any experiment is performed. To be practically useful, such a framework must both capture system-level cellular responses and generalize reliably to drugs, targets, and contexts that were never observed.

A growing number of works on virtual cells aim to extend these efforts by training machine learning models to simu-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

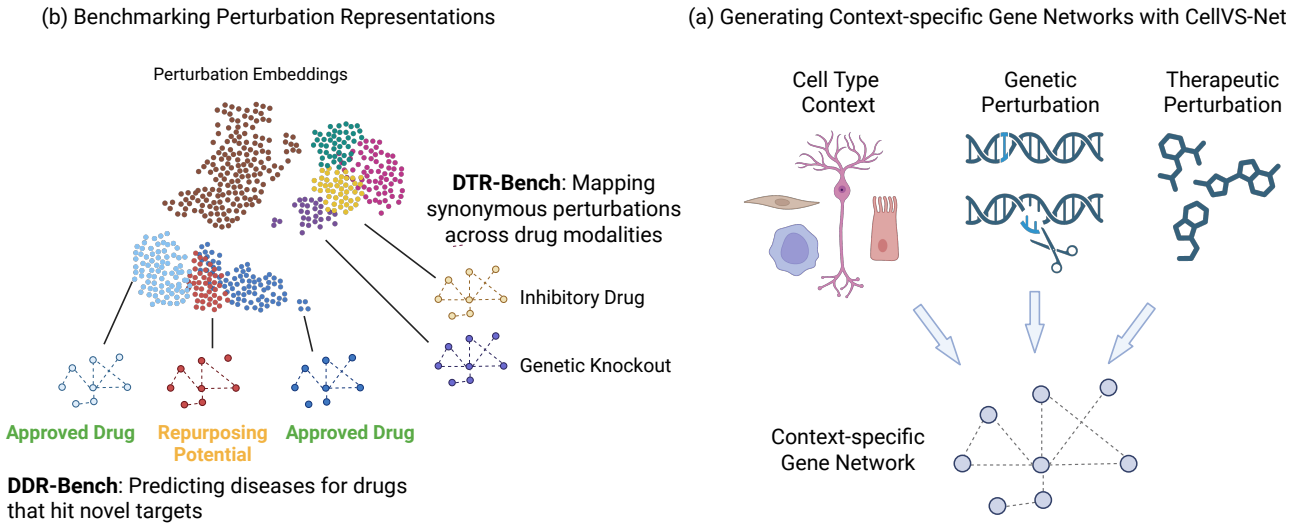

Figure 1. (a) CellVS-Net maps multivariate context (cell type, drug, dose) to context-specific gene networks. (b) We introduce two new benchmarks for evaluating drug representation approaches on clinically-relevant tasks. DDR-Bench predicts effective diseases for drugs with previously unseen target profiles. DTR-Bench maps synonymous perturbations across drug modalities to understand mechanism of action and off-target effects.

late cellular behavior in response to perturbations (Bunne et al.; Song et al.). To the best of our knowledge, all methods that predict cellular response to unseen perturbations are primarily evaluated on their ability to reconstruct a post-perturbation readout (e.g. expression, morphology, IC50) (Theodoris et al., 2023; Cui et al., 2024; Ho et al., 2024; Roohani et al., 2022; Adduri et al., 2025; Yu et al., 2025; He et al., 2025; Ji et al.; Lotfollahi et al., 2023; Fradkin et al.; Bai et al.; Cole et al., 2026). However, expression (predicted and real) is an intermediate representation of perturbation effect, and does not directly identify the safety and efficacy of a therapeutic. Miladinovic et al. (Miladinovic et al., 2025) go beyond reconstruction-based evaluations to investigate synonymous genetic and pharmacological perturbations, but this method is unable to generalize to perturbations beyond its training vocabulary. To enable virtual screening on cellular systems, we believe that next generation methods must (i) provide insights directly relevant to drug discovery and (ii) generalize arbitrarily to unseen therapeutics to promote virtual exploration. To address this, we curate the Drug-Disease Retrieval (DDR-Bench) and Drug-Target Retrieval (DTR-Bench) benchmarks. DDR-Bench evaluates whether a method can identify the correct FDA-approved disease indication for a drug with a novel target profile; a failure case for target-based screening that we hypothesize requires a system-level model of drug response. DTR-Bench extends a traditional drug discovery task (drug-target interaction prediction) to cell-level screening. While molecular methods often predict these interactions directly, cell-level should identify synonymous small molecule and gene knockout perturbations based on the similarity of cell-level responses. We assemble a comprehensive panel of molecular and cell-level representation methods for screening, and identify shortcomings in both categories.

Based on these shortcomings, we also investigate alternatives for cell-level drug effect representation. In particular, post-perturbation gene networks would capture how perturbations rewire cellular circuitry, providing a richer description of perturbation effects than expression snapshots. However, most network inference methods rely on plug-in estimators with large cohorts (Badia-i Mompel et al., 2023; Stone et al., 2021), which fail to capture the continuous and context-dependent rewiring and cannot generalize to new perturbation conditions, a key requirement for virtual screening.

To enable context-dependent prediction of network rewiring, we develop a scalable differentiable surrogate for the multivariate Gaussian negative log-likelihood that decomposes covariance estimation into pairwise regression problems. This avoids direct optimization of the expensive $\log|\Sigma|$ and Σ^{-1} terms in Gaussian likelihood, and serves as a drop-in replacement for the isotropic-Gaussian (mean squared error) losses commonly used in cellular response models. We apply this new objective to train CellVS-Net, a model which maps multivariate cellular and therapeutic contexts (cell type, drug, dose) to predict a post-perturbation gene network represented as a Gaussian graphical model. Rather than fitting a separate network for each drug–cell–dose combination, CellVS-Net learns a single context encoder that maps this context to the parameters of a gene–gene dependency model. We test CellVS-Net by using this loss to predict sample-specific networks on-demand for unseen cell lines and perturbations including small molecules,

large molecules (biologics), genetic knockdowns, and over-expressions.

Our contributions are as follows:

1. We provide the first evaluation of virtual cell methods' ability to prioritize drugs in practical virtual screening settings, assembling a comprehensive panel of molecular and cell-level methods for fair cross-modal comparison. We develop DDR-Bench, which evaluates whether a method can identify correct disease indications for drugs with novel target profiles, and DTR-Bench, which extends drug-target interaction modeling to cell-level methods, bridging the conceptual gap between molecule-level and cell-level screens.

2. We identify shortcomings of virtual cell methods on these benchmarks, motivating new representations of post-perturbation cell state.

3. We posit perturbed gene networks as an improved representation of cell state and develop a scalable differentiable surrogate loss for multivariate Gaussians to train CellVS-Net, a context-adaptive amortized estimator for generating gene-gene networks on-demand.

4. CellVS-Net accurately maps multivariate contexts including cell types, small molecules, large molecules, gene knockdowns, and gene over-expressions to sample-specific gene-gene dependency networks.

5. CellVS-Net achieves SOTA on predicting networks for held-out perturbations. When applied as a screening representation, CellVS-Net also achieves SOTA on the zero-shot DDR and DTR benchmarks.

Methods

Benchmarks

To evaluate and develop cell-level virtual screening methods, we first curate a comprehensive database containing drug, target, expression, and disease approval data by combining the OpenTargets and LINCS databases. OpenTargets provides structured databases of drug-disease pairs based on Phase-IV FDA approval data, drug-target pairs based on known mechanisms of action, and target-disease associations from various lines of evidence. LINCS provides post-perturbation gene expression data for a variety of perturbation types including small molecule, large molecule, genetic knockdown, and genetic overexpression for multiple cell lines. By merging these sources, we produce a large database of (perturbation, target, disease, expression) for multiple perturbation types, which can be reused for evaluation of molecule-based, target-based, and cell-based screening methods.

Representing Drug Effects Both benchmarks reduce each perturbation to a fixed-length vector and compare drugs or targets by Euclidean distance. This is deliberately modality-agnostic: any vector representation (e.g. expert-derived molecular features, foundation-model embeddings, or our predicted networks) plugs into the same retrieval and edge-classification protocols, no sample splitting is required, and the protocol scales as new drugs receive approval, mirroring billion-scale semantic search (McNutt et al.). This provides a stringent zero-shot evaluation setting in which performance differences are interpreted entirely in terms of how well each representation captures therapeutically meaningful cell-level effects.

Drug-Disease Retrieval DDR-Bench evaluates whether a method can identify the correct FDA-approved disease indication for a drug with an unseen target profile; a failure case for target-based screening that we hypothesize requires a system-level model of drug response. Let $\mathcal{D}$ denote the set of curated drugs and let each drug $q \in \mathcal{D}$ be represented by a fixed-length vector $\phi_q \in \mathbb{R}^d$ produced by the method under evaluation (e.g., a CellVS-Net network, a fingerprint, or an expression embedding); the embedding dimension d is method-specific. For a query drug $q \in \mathcal{D}$, the reference set $\mathcal{R} = \mathcal{D} \setminus \{q\}$ consists of all other drugs with different target profiles. We rank the references in ascending order of Euclidean distance $d(q, q') = \|\phi_q - \phi_{q'}\|_2$ for $q' \in \mathcal{R}$. Let $y^\star(q)$ denote the FDA-approved disease indication of q, drawn from a finite set of indications $\mathcal{Y}$, and let $y_{(1)}, \ldots, y_{(|\mathcal{R}|)} \in \mathcal{Y}$ denote the indications of the ranked reference drugs (so $y_{(r)}$ is the indication of the r-th closest reference drug). We define

$$\text{Hits@}k(q) \;=\; \mathbb{1}\left[y^\star(q) \in \{y_{(1)}, \ldots, y_{(k)}\}\right],$$

where $\mathbb{1}[\cdot]$ is the indicator function returning 1 when the condition holds and 0 otherwise, and report the mean over queries for $k \in \{1, 5, 10, 25\}$ (Figure 2). Dataset construction details are deferred to Appendix E.

Drug-Target Retrieval DTR-Bench reconstructs known drug-target interactions from cell-level perturbation signatures, bridging molecular and cell-level screens. Let $\mathcal{D}$ denote the set of small-molecule drugs (LINCS chemical perturbations) and $\mathcal{T}$ the set of protein targets (LINCS shRNA knockdowns of those targets); each drug $i \in \mathcal{D}$ and each target $j \in \mathcal{T}$ is mapped to a vector $\phi_i, \phi_j \in \mathbb{R}^d$ in a shared representation space (the embedding dimension d is method-specific). Let $d_{ij} = \|\phi_i - \phi_j\|_2$ denote their Euclidean distance and $y_{ij} \in \{0, 1\}$ the ground-truth interaction label, with $y_{ij} = 1$ iff drug i binds target j in the curated drug-target graph. We sweep a threshold $\tau \in \mathbb{R}_+$ on d_{ij} over all $|\mathcal{D}| \times |\mathcal{T}|$ pairs, predicting $\hat{y}_{ij}(\tau) = \mathbb{1}[d_{ij} \le \tau]$, and report AUROC and AUPRC for $\{\hat{y}_{ij}(\tau)\}$ versus $\{y_{ij}\}$ as τ varies. We additionally perform bidirectional Hits@k retrieval: for

(a) Drug-target-disease graph. *(b)* Hold out same-target drugs. *(c)* Rank by L2 distance. *(d)* Hit if any top-k matches $y^\star(q)$.

Figure 2. DDR-Bench evaluation. (a) Drugs, targets, and diseases form a tripartite graph. (b) For each query drug q, all reference drugs sharing q's target signature are masked. (c) The remaining reference set is ranked by L2 distance to q in the chosen representation space. (d) Each retrieved drug $d_{(i)}$ is paired with its FDA-approved disease; a top-k neighbor is a hit when its disease equals the query's FDA-approved disease $y^\star(q)$, and a miss otherwise. Hits@k is 1 when at least one of the top-k neighbors is a hit.

(a) True bipartite graph. *(b)* Embed drugs and targets. *(c)* Threshold induces edges. *(d)* Compare to true graph.

Figure 3. DTR-Bench evaluation. (a) Ground-truth drug-target interactions form a bipartite graph. (b) Each method maps drugs and targets into a shared representation space. (c) A distance threshold τ on $d_{ij} = \|\phi_i - \phi_j\|_2$ induces a predicted bipartite graph; AUROC/AUPRC are obtained by sweeping τ. (d) Comparing predicted to true edges yields true positives (solid green), false positives (solid red), and missed true edges (dashed grey); Hits@k is computed by ranking targets per drug and vice versa.

a drug query $i \in \mathcal{D}$, we rank all targets $j \in \mathcal{T}$ in ascending order of d_{ij} and define

$$\text{Hits@}k(i) \;=\; \mathbb{1}[\exists\, j \in \mathcal{T} : y_{ij} = 1 \;\wedge\; j \in \text{top-}k]\,,$$

with the symmetric definition for target queries (ranking drugs $i \in \mathcal{D}$ for a target query $j \in \mathcal{T}$). We report Hits@k for $k \in \{1, 5, 10, 50\}$. Figure 3 illustrates the protocol. Different drug modalities are measured with different assays, so expression-based representations are susceptible to batch effects; before evaluation we PCA-decompose the combined drug and target representations and remove up to the first 3 principal components, reporting the best of the 3 attempts. Dataset construction details are deferred to Appendix E.

Comprehensive Panel of Screening Methods

Drug-only For a cell and target-agnostic baseline, we include a molecular fingerprint baseline (Capecchi et al.). Circular fingerprints remain a workhorse for classical virtual screening pipelines and ligand-based similarity search. Including this baseline grounds our evaluation against a mature, purely structure-based representation that does not see any cellular readout or target information.

Drug-target Interactions We include SPRINT (McNutt et al.), a recent proteome-scale ligand-protein binding model with strong performance on drug-target interaction prediction. For each drug, we use the SPRINT-predicted binding profile against a fixed reference proteome as its vector representation. This baseline is directly optimized for drug-target interaction prediction and grounds our evaluation against a strong target-aware molecular method that uses no cellular readout.

Gene Expression Most virtual cell methods predict post-perturbation gene expression from compound structure. We construct an expression-prediction baseline that takes a ChemBERTa embedding of a drug's SMILES and regresses to either the LINCS L1000 landmark expression vector or its 50-component PCA loadings (Appendix D). The predicted expression or its PCA compression then serves as the drug's representation, simulating the deployment-time output of an expression-only virtual cell.

Cell Embedding We also include two expression embedding baselines to capture latent cell states. Embedding-based methods can outperform other methods on these tasks by removing redundant expression features and distilling cell states into semantically meaningful low-dimensional latent features. For a simple strong baseline, we compare PCA-compressed expression, which provides semantically meaningful low-dimensional features through simple linear compression. We also assess a foundation model (FM) embedding baseline, where we train a model to predict FM embeddings of the post-perturbation gene expression.

Oracles In a realistic virtual screening scenario, expression measurements would *not* be available: the goal is precisely to avoid running large numbers of physical experiments. As an "oracle", we also compare with post-perturbation gene expression. We can can conceptually treat observed expression as the output of an idealized virtual cell that is a perfect generator of the true transcriptional response. Any method that predicts expression is ultimately trying to approximate this oracle. Using expression as a baseline therefore serves two purposes: (i) it provides an

optimistic upper bound for expression-based objectives. If a task is hard even with the true expression snapshot, no virtual cell that only regresses expression can be expected to perform substantially better; and (ii) it lets us ask whether structured representations, such as networks, can *surpass* the utility of raw expression snapshots for downstream retrieval, despite being estimated from the same underlying data. We also consider "oracle" embeddings (both PCA and FM) by directly applying these embedding methods to the oracle expression.

Improving Cell-level Screening Methods with Gene Networks

We identify shortcomings with both molecular and cell-level methods on DDR-Bench and DTR-Bench. To go beyond expression snapshot prediction and embedding, we propose representing drug effects using post-perturbation gene networks. This requires the development of a scalable method for amortized estimation, which can produce gene networks on-demadn for unseen perturbations.

Notation and setup Each biological sample n is a p-dimensional gene-expression vector $X_n \in \mathbb{R}^p$ ($X_{n,i}$ is the expression of gene i), paired with a context vector $C_n \in \mathbb{R}^m$ that encodes its experimental conditions (cell-type identity, perturbation modality, perturbation payload content, dose, and measurement time). We model samples as draws from sample-specific distributions $X_n \sim P(X \mid \theta_n)$, where θ_n contains all distribution parameters (mean, variance, dependence structure).

To share information across samples, we treat parameters as a function of context,

$$P(X, C) \propto \int_\theta P(X \mid \theta)\, P(\theta \mid C)\, d\theta.$$

Following the contextualized modeling framework (Al-Shedivat et al., 2020; Lengerich et al.; Hastie & Tibshirani, 1993), the context encoder is a deterministic deep network $f : \mathbb{R}^m \to \Theta$ giving $P(\theta \mid C) = \delta(\theta - f(C))$ (with δ the Dirac delta) and $P(X_n \mid \theta_n) = P(X_n \mid f(C_n))$. This regime has been extended to several graphical model classes (Ellington et al., b), but sample-specific multivariate Gaussian graphical models remain unaddressed.

Multivariate Gaussian loss We seek a contextualized multivariate Gaussian, $X \mid C \sim \mathcal{N}(\mu(C), \Sigma(C))$, with mean vector $\mu(C) \in \mathbb{R}^p$ and covariance matrix $\Sigma(C) \in \mathbb{R}^{p\times p}$, $\Sigma(C) \succ 0$, both produced by neural encoders, and optimized end-to-end with stochastic gradient descent. Up to constants, the exact negative log-likelihood for one sample is

$$\ell(X, \mu, \Sigma) \;\propto\; \log|\Sigma| + (X-\mu)^\top \Sigma^{-1}(X-\mu), \quad (1)$$

where $\log|\Sigma|$ denotes the log-determinant of Σ. Optimizing this objective requires recomputing Σ^{-1}, $\log|\Sigma|$, and a positive-definite constraint at every gradient step, which prohibitive to run on every sample batch even for small gene panels ($p \sim 50$). We instead optimize a composite-likelihood surrogate built from the regression form of Pearson's correlation between genes i and j,

$$\rho_{ij}^2 \;=\; \frac{\Sigma_{ij}^2}{\Sigma_{ii}\,\Sigma_{jj}} \;=\; \beta_{ij}\,\beta_{ji}, \quad (2)$$

$$\beta_{ij} \;=\; \frac{\mathrm{Cov}(X_i, X_j)}{\mathrm{Var}(X_i)} \;=\; \operatorname*{argmin}_{\beta\in\mathbb{R}} \mathbb{E}\,\|X_j - \beta X_i\|_2^2, \quad (3)$$

where Σ_{ij} is the (i,j) entry of Σ, $\beta_{ij} \in \mathbb{R}$ is the ordinary least-squares (OLS) coefficient for regressing gene X_j on gene X_i, and $\beta \in \mathbb{R}^{p\times p}$ collects all pairwise coefficients. Let $\sigma_i = \sqrt{\Sigma_{ii}}$ denote the marginal standard deviation of gene i, with $\sigma \in \mathbb{R}_+^p$ the vector of all marginal standard deviations and $\mu \in \mathbb{R}^p$ the marginal mean vector. The marginalization properties of Gaussians let each (i,j) pair be fit independently and reassembled via

$$\Sigma_{ij} \;=\; \mathrm{sign}(\beta_{ij})\,\sigma_i\,\sigma_j\,\sqrt{\beta_{ij}\,\beta_{ji}}, \quad (4)$$

where μ and σ are estimated under the per-gene isotropic Gaussian likelihood $-\log p(X_i \mid \mu_i, \sigma_i) \;\propto\; \frac{1}{2}\log(2\pi\sigma_i^2) + (X_i-\mu_i)^2/(2\sigma_i^2)$. The full contextualized objective for one sample with context C is

$$\begin{aligned}\ell(X, \mu(C), \sigma(C), \beta(C)) = \\ \tfrac{1}{p^2}\sum_{i,j=1}^{p}\big((X_j-\mu(C)_j) - (X_i-\mu(C)_i)\,\beta(C)_{ij}\big)^2 \\ + \tfrac{1}{2}\sum_{i=1}^{p}\log\big(2\pi\,\sigma(C)_i^2\big) + \tfrac{1}{2}\sum_{i=1}^{p}\frac{(X_i-\mu(C)_i)^2}{\sigma(C)_i^2}, \quad (5)\end{aligned}$$

which is differentiable, jointly fits the context encoders $\mu(C), \sigma(C), \beta(C)$, and avoids any explicit matrix inversion or determinant. The estimated parameters yield a context-specific Gaussian graphical model when the induced correlation matrix is positive definite, and project onto the positive-definite cone otherwise. This surrogate is a pseudo-likelihood rather than the exact NLL, but inherits the interpretability of Gaussian covariance and serves as a drop-in replacement for the mean-squared-error losses used in current cell models; we report the pairwise regression term alone as MSE.

CellVS-Net Drug efficacy, toxicity, and mechanism of action emerge from coordinated shifts in gene–gene dependencies rather than from marginal expression alone. We therefore seek a representation that describes the system-level effect of a perturbation and generalizes to unseen drugs, targets, and cell types. We apply this new loss to create

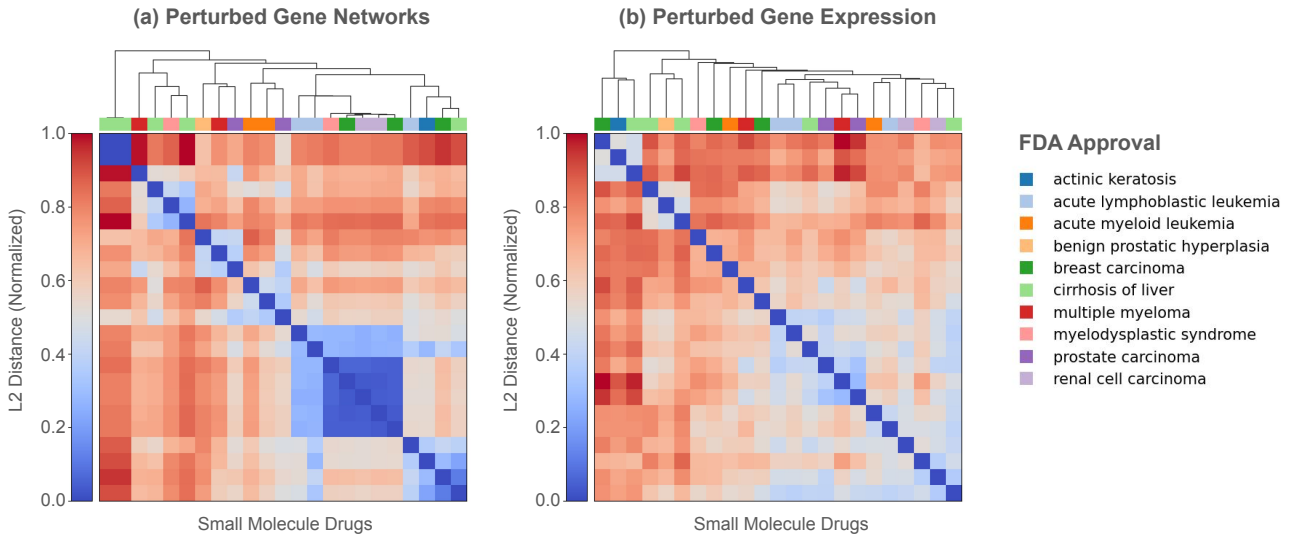

Figure 4. Organization of drugs based on (a) perturbed gene networks and (b) perturbed gene expression representations. Drugs are annotated with their FDA-approved disease indications. All samples are taken from the PC3 cell type.

Table 1. Pairwise regression loss (MSE) of inferred networks on a context-held-out split for each perturbation type. All rows are evaluated on the intersection of held-out perturbations for fair comparison. Per-modality choice of biochemical encoder upstream of the context encoder, and the full encoder ablation, are reported in Appendix B and Table 9.

Model	Chemical	shRNA	Over Expression	Ligand
Population	1.0594	0.9788	0.7769	0.9315
CellVS-Net Molecule	0.6003	—	—	—
CellVS-Net Target	**0.5633**	**0.6740**	**0.6801**	**0.5873**

CellVS-Net, a deep learning model that learns to generate Gaussian graphical models representing gene-gene networks on-demand for unseen drugs, targets, and cell types, capturing context-specific network restructuring. With CellVS-Net, each perturbation is represented not by its transcriptomic snapshot but by the inferred structure of gene–gene relationships that best explains the observed data under that context. The context encoder accommodates many different biochemical representations of a perturbation; modality-specific encoders for small molecules, protein targets, and genetic perturbations are detailed in Appendix B. Implementation details on the architecture and optimizers used are discussed in the Appendix.

We use CellVS-Net's predicted gene-gene network as a structured cell-level representation. Each drug is represented by the upper-triangular squared correlations $\beta_{ij}(C)\beta_{ji}(C)$ concatenated with the predicted mean shift $\mu(C)$, capturing coordinated rewiring of gene-gene dependencies rather than marginal expression alone. This is the natural cell-level alternative to expression snapshots within our framework.

Results

Generating Perturbation-specific Gene Networks

In order to apply post-perturbation gene networks to screening, we first consider the problem of estimating perturbation-specific gene networks. Traditional plug-in estimators fit an independent network for each cell line or perturbation. This approach overfits severely in low-sample regimes and cannot produce estimates for unseen conditions. Population models avoid overfitting, but are a high-bias model which collapses all samples from heterogeneous contexts into a single model. These failure modes mirror real virtual screening settings, where most contexts have few measurements and many perturbations are never directly observed. We require a network estimator which is adaptable to completely unseen contexts and perturbations. CellVS-Net addresses this by learning a smooth mapping from context features to network parameters.

In experiments across perturbation modalities, CellVS-Net reduces network MSE by 47% on chemical perturbations and by 12–37% on the genetic and ligand modalities relative to the population baseline (Table 1). Target-encoded CellVS-Net dominates molecule-only encoding even on chemical

Table 2. Evaluating methods of representing small molecule drugs in terms of their ability to predict FDA approvals. We compile a dataset of diseases, targets, and small molecules, where each disease has multiple approved small molecule drugs targeting different genes or sets of genes. We hold out drugs with identical target profiles, and use each held-out drug to query the remaining drugs, returning the k nearest neighbors in terms of Euclidean distance with $k \in \{1, 5, 10, 25\}$. We report a hit if any of the returned drugs have an FDA approval for the same disease as the held-out drug. Paired bootstrap p-values for Hits@5 versus the random baseline are shown in Table 5; 95% confidence intervals on every cell are listed in the appendix Table 10.

Representation	Method	Disease Hits @1	@5	@10	@25
Perturbed gene network	CellVS-Net	**0.1250**	**0.4464**	**0.6250**	**0.8571**
Perturbed gene expression	Predicted expression	0.0536	0.2143	0.4286	0.7321
	Predicted expression PC loadings	0.0714	0.2857	0.4821	0.7321
	Predicted AIDO.Cell embeddings	0.0714	0.2679	0.5357	0.7679
Molecular interaction	SPRINT (McNutt et al.)	0.0179	0.2500	0.4464	0.7679
Molecule-only	Fingerprint (Capecchi et al.)	0.0357	0.1786	0.3571	0.6071
Random	Random	0.0179	0.1429	0.2857	0.7500
Oracle gene expression	Observed expression	0.1071	0.2500	0.4286	0.8214
	PCA expression	0.0714	0.2143	0.3929	0.8036
	AIDO.Cell embedding (Ho et al., 2024)	0.0357	0.3214	0.5179	0.8393

perturbations, indicating that target identity carries information about cell-level effect that molecular structure alone does not. The full encoder ablation, including pretrained representations across cell, genome, protein-sequence, and protein-structure modalities, is reported in Appendix Table 9.

Training on Disjoint Modalities Improves Overall Performance

Table 3. Network prediction MSE using ChemBERTa representations for chemical perturbations and the best-performing representations for other perturbation types (from Table 1) compared against CellVS-Net trained on all perturbations together with those same representations.

Model	Chem	shRNA	OE	Ligand
CellVS-Net (separate)	0.6003	0.6740	0.6801	0.5873
CellVS-Net (joint)	0.5907	0.5524	0.3410	0.5419

The four perturbation modalities in LINCS L1000 (chemical, shRNA, overexpression, ligand) act on shared cellular machinery but only observed one at a time and are typically modeled in isolation. A single CellVS-Net context encoder accepts heterogeneous context features and supports joint training across all four, allowing dependency structure learned from one modality to inform predictions for the others. We train one joint encoder on the union of all perturbation types using the best-performing context representation for each modality from Table 1 and compare against the per-modality encoders (Table 3). We find that joint training across contexts, even for disjoint context modalities, reduces network MSE on every modality. We see the largest gains on the data-poorer genetic perturbations: shRNA improves from 0.674 to 0.552 (-18%) and overexpression from 0.680 to 0.341 (-50%). Chemical and ligand modalities, which already have larger training sets, see smaller but consistent improvements (-2% and -8%). This pattern is consistent with positive transfer from data-rich to data-scarce modalities through a shared dependency-network output space.

Drug-Disease Retrieval: Predicting Disease Indications for Drugs with Novel Targets

A useful cell-level screen should cluster drugs with similar therapeutic effects even when they hit different targets. We assemble small-molecule drugs from OpenTargets with disjoint target profiles but a shared FDA-approved disease, providing a ground truth that target-centric methods cannot recover by construction. CellVS-Net networks place same-disease drugs closer than expression, foundation-model embeddings, target-binding, or fingerprint baselines (Table 2; Figure 4). Critically, CellVS-Net is the only representation whose Hits@5 is significantly different from random after multiple-testing correction (Table 5); it also surpasses the oracle expression upper bound, indicating that gene-network restructuring carries therapeutic signal that raw expression snapshots do not. The clustermap in Figure 4 visualizes this: same-disease drugs form more cohesive blocks under network distance, whereas expression-based distances mix across indications.

Table 4. Recovering known drug-target relationships using different perturbation representations on DTR-Bench. AUROC and AUPRC are calculated using ground-truth and predicted bipartite drug-target graphs via distance thresholding. Expression-based representations are derived from LINCS L1000; PCA applies a 50-component PCA to the full dataset; AIDO.Cell embeds each sample. One-sided paired bootstrap p-values versus the random baseline are reported alongside each metric (10,000 resamples); 95% confidence intervals are listed in appendix Table 11.

	AUROC	p-value	AUPRC	p-value
CellVS-Net predicted networks	**0.540**	**0.0013**	**0.012**	**0.0028**
Predicted expression	0.482	0.639	0.009	0.430
Predicted PCA expression	0.476	0.758	0.009	0.268
Predicted AIDO.Cell embeddings	0.501	0.261	0.009	0.121
SPRINT (McNutt et al.)	0.445	0.993	0.008	0.874
Random	0.489	—	0.008	—
Oracle observed expression	0.513	0.082	0.009	0.132
Oracle PCA expression	0.521	0.033	0.010	0.065
Oracle AIDO.Cell (Ho et al., 2024)	0.521	0.027	0.010	0.022

Drug-Target Retrieval: Matching Synonymous Perturbations Across Modalities.

DTR-Bench asks whether known drug-target interactions can be reconstructed from cell-level signatures, bridging cell-level screens to traditional drug-target interaction prediction. We pair LINCS chemical perturbations with shRNA knockdowns of their known targets and compare to SPRINT (McNutt et al.), a SOTA target-binding model that does not see cellular readout. CellVS-Net outperforms all oracle and predicted baselines on global drug-target graph reconstruction, and is the only method whose AUROC and AUPRC are significantly above random (Table 4). SPRINT, which is directly optimized for drug-target binding from molecular structure, performs at chance on this protocol. This gap motivates representations that act on shared cellular machinery rather than on molecular structure alone. Per-query lookup results and a web tool for browsing shRNA - chemical embeddings are in Appendix Table 12 and Appendix Figure 5.

Table 5. Paired bootstrap p-values (10,000 resamples) for Hits@5 versus random performance on DDR-Bench. Row ordering matches Table 2. CellVS-Net is the only method significantly better than random after multiple testing correction.

Representation	Method	p-value
Perturbed gene network	CellVS-Net	$<10^{-4}$
Perturbed gene expression	Predicted expression	0.131
	Predicted PC loadings	0.017
	Predicted AIDO.Cell	0.044
Molecular interaction	SPRINT	0.106
Molecule-only	Fingerprint (Morgan)	0.360
Oracle gene expression	Observed expression	0.078
	PCA expression	0.168
	AIDO.Cell embedding	0.013

Discussion

In this work, we aim to move late-stage drug-discovery failures into earlier in-silico screening stages through the development of CellVS-Net, a computational tool trained on cell line perturbation data which accurately represents drug purposes and effects. CellVS-Net introduces a scalable differentiable surrogate for the multivariate Gaussian likelihood that decomposes covariance estimation into pairwise regressions, with desirable statistical properties for estimation in long-tail drug screening applications (Table 13). The result generalizes smoothly to unseen drugs, doses, and cell types, while improving with richer context representations, reducing MSE by 12–47% across modalities (Tables 1, 14). Beyond CellVS-Net itself, the surrogate is a drop-in replacement for the isotropic-Gaussian (mean squared error) loss that dominates current cell-modeling pipelines, and adds context-specific gene-gene dependence as an output without changing the encoder architecture.

In this study, we also provide the first quantitative definition of cell-level virtual screening through the formulation of the DDR-Bench and DTR-Bench benchmarks. Both benchmarks are method-agnostic, evaluating drug, drug-target, and cell modeling approaches across several data modalities on clinically grounded endpoints: recovering disease indications for drugs with unseen targets (Table 2) and reconstructing drug-target relationships from the effects of different perturbation modalities (Table 4). On DDR-Bench and DTR-Bench, CellVS-Net improves over molecular and expression baselines, including the oracle expression upper bound that represents an ideal virtual cell. The fact that gene networks surpass this oracle indicates that structured representations of gene-gene dependence carry therapeutically relevant signal that is not recoverable from raw expression snapshots, even in principle. This finding redirects the virtual-cell research agenda away from reconstruction accuracy alone and toward downstream therapeutic utility as a primary evaluation criterion.

References

Adduri, A. K., Gautam, D., Bevilacqua, B., Imran, A., Shah, R., Naghipourfar, M., Teyssier, N., Ilango, R., Nagaraj, S., Dong, M., Ricci-Tam, C., Carpenter, C., Subramanyam, V., Winters, A., Tirukkovular, S., Sullivan, J., Plosky, B. S., Eraslan, B., Youngblut, N. D., Leskovec, J., Gilbert, L. A., Konermann, S., Hsu, P. D., Dobin, A., Burke, D. P., Goodarzi, H., and Roohani, Y. H. Predicting cellular responses to perturbation across diverse contexts with state, June 2025. URL https://www.biorxiv.org/content/10.1101/2025.06.26.661135v1. bioRxiv.

Al-Shedivat, M., Dubey, A., and Xing, E. Contextual Explanation Networks. *J. Mach. Learn. Res.*, 21(194):1–44, 2020. ISSN 1532-4435. URL http://jmlr.org/papers/v21/18-856.html.

Badia-i Mompel, P., Wessels, L., Müller-Dott, S., Trimbour, R., Ramirez Flores, R. O., Argelaguet, R., and Saez-Rodriguez, J. Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews Genetics*, pp. 1–16, June 2023. ISSN 1471-0064. doi: 10.1038/s41576-023-00618-5. URL https://www.nature.com/articles/s41576-023-00618-5. Publisher: Nature Publishing Group.

Bai, D., Ellington, C. N., Mo, S., Song, L., and Xing, E. P. AttentionPert: accurately modeling multiplexed genetic perturbations with multi-scale effects. 40:i453–i461. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae244. URL https://doi.org/10.1093/bioinformatics/btae244.

Bray, M.-A., Singh, S., Han, H., Davis, C. T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S. M., Gibson, C. C., and Carpenter, A. E. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. 11(9):1757–1774. ISSN 1750-2799. doi: 10.1038/nprot.2016.105.

Bunne, C., Roohani, Y., Rosen, Y., Gupta, A., Zhang, X., Roed, M., Alexandrov, T., AlQuraishi, M., Brennan, P., Burkhardt, D. B., Califano, A., Cool, J., Dernburg, A. F., Ewing, K., Fox, E. B., Haury, M., Herr, A. E., Horvitz, E., Hsu, P. D., Jain, V., Johnson, G. R., Kalil, T., Kelley, D. R., Kelley, S. O., Kreshuk, A., Mitchison, T., Otte, S., Shendure, J., Sofroniew, N. J., Theis, F., Theodoris, C. V., Upadhyayula, S., Valer, M., Wang, B., Xing, E., Yeung-Levy, S., Zitnik, M., Karaletsos, T., Regev, A., Lundberg, E., Leskovec, J., and Quake, S. R. How to build the virtual cell with artificial intelligence: Priorities and opportunities. 187(25):7045–7063. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2024.11.015. URL https://www.cell.com/cell/abstract/S0092-8674(24)01332-1.

Capecchi, A., Probst, D., and Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. 12(1):43. ISSN 1758-2946. doi: 10.1186/s13321-020-00445-4. URL https://doi.org/10.1186/s13321-020-00445-4.

Chandrasekaran, S. N., Ackerman, J., Alix, E., Ando, D. M., Arevalo, J., Bennion, M., Boisseau, N., Borowa, A., Boyd, J. D., Brino, L., Byrne, P. J., Ceulemans, H., Ch'ng, C., Cimini, B. A., Clevert, D.-A., Deflaux, N., Doench, J. G., Dorval, T., Doyonnas, R., Dragone, V., Engkvist, O., Faloon, P. W., Fritchman, B., Fuchs, F., Garg, S., Gilbert, T. J., Glazer, D., Gnutt, D., Goodale, A., Grignard, J., Guenther, J., Han, Y., Hanifehlou, Z., Hariharan, S., Hernandez, D., Horman, S. R., Hormel, G., Huntley, M., Icke, I., Iida, M., Jacob, C. B., Jaensch, S., Khetan, J., Kost-Alimova, M., Krawiec, T., Kuhn, D., Lardeau, C.-H., Lembke, A., Lin, F., Little, K. D., Lofstrom, K. R., Lotfi, S., Logan, D. J., Luo, Y., Madoux, F., Zapata, P. A. M., Marion, B. A., Martin, G., McCarthy, N. J., Mervin, L., Miller, L., Mohamed, H., Monteverde, T., Mouchet, E., Nicke, B., Ogier, A., Ong, A.-L., Osterland, M., Otrocka, M., Peeters, P. J., Pilling, J., Prechtl, S., Qian, C., Rataj, K., Root, D. E., Sakata, S. K., Scrace, S., Shimizu, H., Simon, D., Sommer, P., Spruiell, C., Sumia, I., Swalley, S. E., Terauchi, H., Thibaudeau, A., Unruh, A., Waeter, J. V. d., Dyck, M. V., Staden, C. v., Warchoł, M., Weisbart, E., Weiss, A., Wiest-Daessle, N., Williams, G., Yu, S., Zapiec, B., Żyła, M., Singh, S., and Carpenter, A. E. JUMP cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. URL https://www.biorxiv.org/content/10.1101/2023.03.23.534023v2. Pages: 2023.03.23.534023 Section: New Results.

Cole, E., Huizing, G.-J., Addagudi, S., Ho, N., Hasanaj, E., Kuijs, M., Johnstone, T., Carilli, M., Davi, A., Ellington, C., Feinauer, C., Li, P., Menegaux, R., Mohammadi, S., Shao, Y., Zhang, J., Lundberg, E., Song, L., Bar-Joseph, Z., and Xing, E. P. Foundation models improve perturbation response prediction. *bioRxiv*, 2026. doi: 10.64898/2026.02.18.706454. URL https://www.biorxiv.org/content/10.64898/2026.02.18.706454v1.full.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21:1470–1480, February 2024. doi: 10.1038/s41592-024-02201-0. URL https://www.nature.com/articles/s41592-024-02201-0.

Ellington, C. N., Lengerich, B. J., Lo, W., Alvarez, A., Rubbi, A., Kellis, M., and Xing, E. P. Contextualized: Heterogeneous modeling toolbox. 9 (97):6469, a. ISSN 2475-9066. doi: 10.21105/

joss.06469. URL https://joss.theoj.org/papers/10.21105/joss.06469.

Ellington, C. N., Lengerich, B. J., Watkins, T. B. K., Yang, J., Adduri, A. K., Mahbub, S., Xiao, H., Kellis, M., and Xing, E. P. Learning to estimate sample-specific transcriptional networks for 7,000 tumors. 122(21):e2411930122, b. doi: 10.1073/pnas.2411930122. URL https://www.pnas.org/doi/10.1073/pnas.2411930122.

Ellington, C. N., Sun, N., Ho, N., Tao, T., Mahbub, S., Li, D., Zhuang, Y., Wang, H., Song, L., and Xing, E. P. Accurate and general dna representations emerge from genome foundation models at scale. *bioRxiv*, 2024. doi: 10.1101/2024.12.01.625444. URL https://doi.org/10.1101/2024.12.01.625444.

Fradkin, P., Azadi, P., Suri, K., Wenkel, F., Bashashati, A., Sypetkowski, M., and Beaini, D. How molecules impact cells: Unlocking contrastive PhenoMolecular retrieval. URL http://arxiv.org/abs/2409.08302.

Hasanaj, E., Cole, E., Mohammadi, S., Addagudi, S., Zhang, X., Song, L., and Xing, E. P. Multimodal benchmarking of foundation model representations for cellular perturbation response prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.26.661186. URL https://www.biorxiv.org/content/10.1101/2025.06.26.661186.

Hastie, T. and Tibshirani, R. Varying-Coefficient Models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1993.tb01939.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1993.tb01939.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1993.tb01939.x.

He, S., Zhu, Y., Tavakol, D. N., Ye, H., Lao, Y.-H., Zhu, Z., Xu, C., Chauhan, S., Garty, G., Tomer, R., Vunjak-Novakovic, G., Zou, J., Azizi, E., and Leong, K. W. Squidiff: predicting cellular development and responses to perturbations using a diffusion model. *Nature Methods*, 2025. doi: 10.1038/s41592-025-02877-y. URL https://www.nature.com/articles/s41592-025-02877-y.

Ho, N., Ellington, C. N., Hou, J., Addagudi, S., Mo, S., Tao, T., Li, D., Zhuang, Y., Wang, H., Cheng, X., Song, L., and Xing, E. P. Scaling dense representations for single cell with transcriptome-scale context. URL https://www.biorxiv.org/content/10.1101/2024.11.28.625303v1. Pages: 2024.11.28.625303 Section: New Results.

Ho, N., Ellington, C. N., Hou, J., Addagudi, S., Mo, S., Tao, T., Li, D., Zhuang, Y., Wang, H., Cheng, X., Song, L., and Xing, E. P. Scaling dense representations for single cell with transcriptome-scale context, November 2024. URL https://www.biorxiv.org/content/10.1101/2024.11.28.625303v1. bioRxiv.

Ji, Y., Tejada-Lapuerta, A., Schmacke, N. A., Zheng, Z., Zhang, X., Khan, S., Rothenaigner, I., Tschuck, J., Hadian, K., Hornung, V., and Theis, F. J. Scalable and universal prediction of cellular phenotypes enables in silico experiments. URL https://www.biorxiv.org/content/10.1101/2024.08.12.607533v3. ISSN: 2692-8205 Pages: 2024.08.12.607533 Section: New Results.

Koleti, A., Terryn, R., Stathias, V., Chung, C., Cooper, D. J., Turner, J. P., Vidović, D., Forlin, M., Kelley, T. T., D'Urso, A., Allen, B. K., Torre, D., Jagodnik, K. M., Wang, L., Jenkins, S. L., Mader, C., Niu, W., Fazel, M., Mahi, N., Pilarczyk, M., Clark, N., Shamsaei, B., Meller, J., Vasiliauskas, J., Reichard, J., Medvedovic, M., Ma'ayan, A., Pillai, A., and Schürer, S. C. Data portal for the library of integrated network-based cellular signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. 46:D558–D566. ISSN 0305-1048. doi: 10.1093/nar/gkx1063. URL https://doi.org/10.1093/nar/gkx1063.

Lengerich, B., Ellington, C. N., Rubbi, A., Kellis, M., and Xing, E. P. Contextualized machine learning. URL http://arxiv.org/abs/2310.11340.

Lotfollahi, M., Klimovskaia Susmelj, A., De Donno, C., Hetzel, L., Ji, Y., Ibarra, I. L., Srivatsan, S. R., Naghipourfar, M., Daza, R. M., Martin, B., Shendure, J., McFaline-Figueroa, J. L., Boyeau, P., Wolf, F. A., Yakubova, N., Günnemann, S., Trapnell, C., Lopez-Paz, D., and Theis, F. J. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, June 2023. ISSN 1744-4292. doi: 10.15252/msb.202211517. URL https://www.embopress.org/doi/full/10.15252/msb.202211517. Publisher: John Wiley & Sons, Ltd.

McNutt, A. T., Adduri, A. K., Ellington, C. N., Dayao, M. T., Xing, E. P., Mohimani, H., and Koes, D. R. Scaling structure aware virtual screening to billions of molecules with SPRINT. URL http://arxiv.org/abs/2411.15418.

Miladinovic, D., Höppe, T., Chevalley, M., Georgiou, A., Stuart, L., Mehrjou, A., Bantscheff, M., Schölkopf, B., and Schwab, P. In silico biological discovery with large perturbation models. *Nature Computational Science*, October 2025. doi: 10.1038/s43588-025-00870-1. URL https://www.nature.com/articles/s43588-025-00870-1.

Norman, T. M., Horlbeck, M. A., Replogle, J. M., Ge, A. Y., Xu, A., Jost, M., Gilbert, L. A., and Weissman, J. S. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science (New York, N.Y.)*, 365(6455):786–793, August 2019. ISSN 1095-9203. doi: 10.1126/science.aax4438.

Peidli, S., Green, T. D., Shen, C., Gross, T., Min, J., Garda, S., Yuan, B., Schumacher, L. J., Taylor-King, J. P., Marks, D. S., Luna, A., Blüthgen, N., and Sander, C. scPerturb: harmonized single-cell perturbation data. 21(3):531–540. ISSN 1548-7105. doi: 10.1038/s41592-023-02144-y. URL https://www.nature.com/articles/s41592-023-02144-y.

Roohani, Y., Huang, K., and Leskovec, J. GEARS: Predicting transcriptional outcomes of novel multi-gene perturbations, July 2022. URL https://www.biorxiv.org/content/10.1101/2022.07.12.499735v1. Pages: 2022.07.12.499735 Section: New Results.

Singh, R., Barsainyan, A. A., Irfan, R., Amorin, C. J., He, S., Davis, T., Thiagarajan, A., Sankaran, S., Chithrananda, S., Ahmad, W., Jones, D., McLoughlin, K., Kim, H., Bhutani, A., Sathyanarayana, S. V., Viswanathan, V., Allen, J. E., and Ramsundar, B. ChemBERTa-3: An Open Source Training Framework for Chemical Foundation Models. *ChemRxiv*, 2025. doi: 10.26434/chemrxiv-2025-4glrl-v2. URL https://doi.org/10.26434/chemrxiv-2025-4glrl-v2.

Song, L., Segal, E., and Xing, E. Toward AI-driven digital organism: Multiscale foundation models for predicting, simulating and programming biology at all levels. URL http://arxiv.org/abs/2412.06993.

Stone, M., McCalla, S. G., Siahpirani, A. F., Periyasamy, V., Shin, J., and Roy, S. Identifying strengths and weaknesses of methods for computational network inference from single cell RNA-seq data. Publication Title: bioRxiv, June 2021. URL https://www.biorxiv.org/content/10.1101/2021.06.01.446671v1.

Sun, N., Zou, S., Tao, T., Mahbub, S., Li, D., Zhuang, Y., Wang, H., Cheng, X., Song, L., and Xing, E. P. Mixture of experts enable efficient and effective protein understanding and design. *bioRxiv*, 2024. doi: 10.1101/2024.11.29.625425. URL https://doi.org/10.1101/2024.11.29.625425.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., and Ellinor, P. T. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, June 2023. doi: 10.1038/s41586-023-06139-9. URL https://www.nature.com/articles/s41586-023-06139-9.

Yu, H., Qian, W., Song, Y., and Welch, J. D. Perturbnet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular Systems Biology*, 2025. doi: 10.1038/s44320-025-00131-3. URL https://www.embopress.org/doi/full/10.1038/s44320-025-00131-3.

Zhang, J., Meynard-Piganeau, B., Gong, J., Cheng, X., Luo, Y., Ly, H., Song, L., and Xing, E. Balancing locality and reconstruction in protein structure tokenizer. *bioRxiv*, 2024. doi: 10.1101/2024.12.02.626366. URL https://doi.org/10.1101/2024.12.02.626366.

Zhang, J., Ubas, A. A., de Borja, R., Svensson, V., Thomas, N., Thakar, N., Lai, I., Winters, A., Khan, U., Jones, M. G., Tran, V., Pangallo, J., Papalexi, E., Sapre, A., Nguyen, H., Sanderson, O., Nigos, M., Kaplan, O., Schroeder, S., Hariadi, B., Marrujo, S., Curca, C., Salvino, A., Gallareta Olivares, G., Koehler, R., Geiss, G., Rosenberg, A., Roco, C., Merico, D., Alidoust, N., Goodarzi, H., and Yu, J. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling, February 2025. URL https://www.biorxiv.org/content/10.1101/2025.02.20.639398v1. bioRxiv.

A. Training

For each perturbation type in the LINCS L1000 dataset (small molecule, shRNA, overexpression, ligand) we apply quality control filters based on replicate correlation and self-ranking performance to ensure high-confidence perturbation profiles, then hold-out 20% of perturbations at random. We construct a context vector C for each sample from metadata including perturbation type, target gene (for genetic perturbations), dose, timepoint, and control expression for the corresponding cell line. Expression measurements are compressed to 50 metagenes using principal component analysis, inferred from the train set. All contexts and expression samples are feature-normalized according to train-set mean and standard deviation prior to fitting. To train the model, we apply the Contextualized modeling Python library (Ellington et al., a). We test several methods for representing perturbations to improve generalization to unseen conditions, described in the section below.

Compute resources. Each CellVS-Net training run reported in Tables 14 and 9 takes approximately 1 h on a single NVIDIA H100 80 GB GPU. Pre-computing perturbation embeddings runs once per representation: AIDO.Cell embeddings for all samples take roughly 24 h, while each of the other gene-level embeddings used as context in Table 9 (AIDO.DNA, AIDO.Protein, AIDO.StructureTokenizer, ChemBERTa, PCA) takes roughly 1 h, and SPRINT embeddings used in Tables 2 and 12 take roughly 1 h. End-to-end data processing from raw LINCS L1000 and OpenTargets sources takes about 2 h on CPU with $\sim$200 GB of memory; all remaining benchmark scripts (table generation, retrieval, bootstrap intervals) complete in under 5 min of CPU time.

B. Perturbation Representations

We employ multiple representation strategies for perturbations in our trained networks, motivated by recent efforts in benchmarking multimodal foundation models for cellular perturbation prediction (Hasanaj et al., 2025). For small-molecule perturbations, we use SMILES-based molecular representations, while for all perturbation types, we also explore target-based representations derived from gene-level embeddings.

SMILES-based networks. For small-molecule perturbations, we compare two chemical featurization strategies. First, we compute Morgan fingerprints (Capecchi et al.), a substructure representation that encodes local atomic environments and has proven effective in traditional cheminformatics pipelines. Second, we apply ChemBERTa-100M-MLM (Singh et al., 2025), a transformer-based molecular foundation model trained on large SMILES corpora, which provides contextualized embeddings that better capture semantic and structural relationships among compounds. These two representations provide complementary baselines for evaluating molecular embedding quality and their effect on drug–target inference.

Target-based networks. For perturbations with gene targets, we integrate embeddings from multiple biological foundation models spanning expression, genomic sequence, and protein structure modalities.

`AIDO.Cell (expression-based)`. We use AIDO.Cell 100M (Ho et al.), a full-transcriptome single-cell foundation model trained across diverse cellular contexts. Gene embeddings are computed using K562 control cells from Norman et al (Norman et al., 2019).

`AIDO.DNA (sequence-based)`. We extract sequence-level gene representations using the AIDO.DNA model (Ellington et al., 2024). For each gene, we define a 4 kbp window centered at the transcription start site (TSS), run model inference to obtain nucleotide embeddings, and apply mean pooling across the sequence to generate a single fixed-length embedding vector per gene.

`AIDO.Protein (structure-informed)`. To capture protein-level information, we utilize AIDO.ProteinIF-16B (Sun et al., 2024), a large-scale model trained jointly on sequence and inferred structure representations. Residue-level embeddings are mean-pooled to yield protein-level embeddings, and for genes encoding multiple isoforms, we average across all available proteins.

`AIDO.StructureTokenizer (geometry-based)`. We further incorporate 3D structural information using the AIDO.StructureTokenizer model (Zhang et al., 2024), which tokenizes protein backbone geometry and side-chain orientations to produce structure-aware embeddings. For genes with multiple resolved structures, we mean-pool over all available embeddings.

`PCA (non-FM)`. As described in previous benchmarking studies (Hasanaj et al., 2025), we derive baseline gene embeddings by applying PCA to control-condition expression profiles. For each gene, we collect its unperturbed expression values across all control samples and project this vector into a PCA space learned over the full control expression matrix

(compressing variation across samples). This was once again computed with K562 control cells from Norman.

C. CellVS-Net Molecule Trained On All Available Drugs

Model Type	Context Encoder	Chemical
Population	None	0.9807
CellVS-Net Molecule	Morgan Fingerprint	0.5433
	ChemBERTa-100M-MLM	0.5284

Table 6. Mean squared error (MSE) of inferred networks across held out chemical perturbations. This evaluation uses the same test set as Table 3, but the training set includes all available drugs with corresponding SMILES strings, rather than only drugs with known targets.

D. Prediction of Molecular Representations

We trained supervised regression models to predict perturbation-induced molecular representations from chemical structure–derived embeddings.

Input Features. For all prediction tasks, the input representation $X \in \mathbb{R}^d$ consisted of precomputed chemBERTa embeddings associated with the compound corresponding to each perturbation instance. These embeddings are fixed-length continuous vectors derived from SMILES strings and are independent of cellular context.

Predictor Model. We used a multi-output ridge regression model to map chemical embeddings to molecular representations. Given an input matrix $X \in \mathbb{R}^{n\times d}$ and target matrix $Y \in \mathbb{R}^{n\times p}$, the model solves

$$\min_W \ \|Y - XW\|_2^2 + \alpha\|W\|_2^2,$$

where $W \in \mathbb{R}^{d\times p}$ is the regression weight matrix and $\alpha = 1.0$ is the regularization parameter. A separate model was trained for each type of molecular representation. Models were fit using only perturbation instances in the training split and then used to generate predictions for all instances.

Predicting PCA Metagenes. Gene expression profiles were first standardized and projected into a low-dimensional space using principal component analysis (PCA). The top $K = 50$ principal components were retained and treated as metagene features.

Predicting Gene Expression. In the expression prediction setting, the supervision target Y consisted of the landmark gene expression vector for each perturbation instance.

Predicting AIDO Cell 3M Embeddings. For representation learning with foundation-model embeddings, the supervision target was the 128-dimensional AIDO Cell 3M embedding associated with each perturbation instance.

E. Benchmark Curation

DDR-Bench construction. We filter the combined OpenTargets-LINCS dataset to small molecules represented in LINCS chemical perturbations, then restrict to diseases with at least 2 Phase-IV FDA-approved drugs whose target signatures (sorted lists of Ensembl protein IDs known to be bound) are distinct. For evaluation, one target signature is held out at a time and used to query the remaining drugs. Coverage by disease is reported in Table 8.

DTR-Bench construction. We filter the combined OpenTargets-LINCS dataset to drug-target pairs in which the drug appears as a LINCS chemical perturbation and the target appears as a LINCS shRNA knockdown. We apply quality control filters to both modalities based on replicate correlation and self-ranking performance to retain only high-confidence perturbation profiles. Summary statistics are reported in Table 7.

Total pairs	Unique drugs	Unique targets	Avg. drugs/target	Avg. targets/drug
559	332	194	2.88 ± 4.72	1.68 ± 2.32

Table 7. DTR-Bench summary statistics.

Disease ID	Disease Name	Targets	Drugs
EFO_0000305	breast carcinoma	5	7
EFO_0001422	cirrhosis of liver	5	5
EFO_0000220	acute lymphoblastic leukemia	4	4
EFO_0000222	acute myeloid leukemia	4	4
EFO_0000284	benign prostatic hyperplasia	3	4
EFO_0001378	multiple myeloma	3	4
EFO_0002496	actinic keratosis	3	3
EFO_0000681	renal cell carcinoma	3	3
EFO_0000198	myelodysplastic syndrome	2	3
EFO_0001663	prostate carcinoma	2	3
EFO_1001469	Mantle cell lymphoma	2	2
MONDO_0015760	T-cell non-Hodgkin lymphoma	2	2
EFO_0004193	basal cell carcinoma	2	2
EFO_1001012	leptomeningeal metastasis	2	2
EFO_0004289	lymphoid leukemia	2	2
EFO_1001051	mycosis fungoides	2	2
EFO_0003060	non-small cell lung carcinoma	2	2
EFO_1000045	pancreatic neuroendocrine tumor	2	2

Table 8. DDR-Bench coverage by disease. For each disease, we report the number of distinct target signatures with at least one drug and the total number of distinct drugs mapped to those signatures. Target signatures are represented as a sorted list of Ensembl ids.

F. Extended Results

F.1. Encoder Ablation for CellVS-Net

Table 9. Pairwise regression loss (MSE) of inferred networks on a context-held-out split for various perturbation types. All CellVS-Net variants and the population baseline are evaluated on the intersection of all held-out perturbations for fair comparison. Best per column in bold.

Model Type	Context Encoder	Chemical	shRNA	Over Expression	Ligand
Population	None	1.0594	0.9788	0.7769	0.9315
CellVS-Net Molecule	Morgan Fingerprint (Capecchi et al.)	0.5786	—	—	—
	ChemBERTa (Singh et al., 2025)	0.6003	—	—	—
CellVS-Net Target	AIDO.Structure (Zhang et al., 2024)	0.5795	0.6741	0.6907	**0.5873**
	AIDO.Protein (Sun et al., 2024)	**0.5633**	**0.6740**	**0.6801**	0.5890
	AIDO.Cell (Ho et al., 2024)	0.6005	0.6754	0.7403	0.6414
	AIDO.DNA (Ellington et al., 2024)	0.6015	0.6817	0.7645	0.6808
	Gene PCA	0.6105	0.6777	0.7070	0.6122

Context representations impose a prior on the similarity of downstream network estimation tasks for CellVS-Net. Good representations can greatly improve accuracy and generalization, even in the presence of noise features and non-linear effects in this modeling regime (Lengerich et al.; Ellington et al., b). We try several representations for small molecule, large molecule, and genetic perturbations, aiming to produce a highly generalizable perturbation-specific network generator. We compare these context-adaptive models against a context-agnostic population estimator. Unlike previous experiments, group-specific modeling and one-hot contexts are not applicable in this regime, as unseen contexts cannot be mapped onto the original groups or feature set. We evaluate models in terms of the pairwise regression loss on held-out perturbations with expression measurements (Table 9).

CellVS-Net strongly outperforms the context-agnostic baseline by learning to map cell type and perturbation contexts to gene network rewiring. For small-molecule perturbations, CellVS-Net generalizes effectively to held-out molecules: representing molecules with structural fingerprints reduces error by 45%, and representing them by their known protein targets improves further to 47%. For shRNA, over-expression, and ligand perturbations, contexts are represented by their target gene or ligand protein. We include a non-pretrained context representation (Gene PCA) in each case to evaluate the importance of pretrained representations for generalization. Pretrained protein and structure representations consistently outperform PCA across genetic and ligand modalities, with AIDO.Protein achieving the lowest MSE on chemical, shRNA, and over-expression perturbations and AIDO.Structure on ligand perturbations.

Table 10. Evaluating methods of representing small molecule drugs in terms of their ability to predict FDA approvals. We compile a dataset of diseases, targets, and small molecules, where each disease has multiple approved small molecule drugs targeting different genes or sets of genes. We hold out drugs with identical target profiles, and use each held-out drug to query the remaining drugs, returning the k nearest neighbors in terms of Euclidean distance with $k \in \{1, 5, 10, 25\}$. We report a hit if any of the returned drugs have an FDA approval for the same disease as the held-out drug. Confidence intervals (95%) are computed via bootstrap (10,000 resamples) where available.

		Disease Hits			
Representation	Method	@1	@5	@10	@25
Perturbed gene network	CellVS-Net	**0.1250 [0.0536, 0.2143]**	**0.4464 [0.3214, 0.5714]**	**0.6250 [0.5000, 0.7500]**	**0.8571 [0.7679**
Perturbed gene expression	Predicted expression	0.0536 [0.00, 0.13]	0.2143 [0.11, 0.32]	0.4286 [0.31, 0.55]	0.7321 [0.61
	Predicted PC	0.0714 [0.02, 0.14]	0.2857 [0.18, 0.41]	0.4821 [0.36, 0.61]	0.7321 [0.61
	Predicted AIDO.Cell	0.0714 [0.02, 0.14]	0.2679 [0.16, 0.39]	0.5357 [0.41, 0.66]	0.7679 [0.64
Molecular interaction	SPRINT	0.0179 [0.000, 0.054]	0.2500 [0.143, 0.375]	0.4464 [0.321, 0.571]	0.7679 [0.661
Molecule-only	Fingerprint	0.0357 [0.000, 0.089]	0.1786 [0.089, 0.286]	0.3571 [0.232, 0.482]	0.6071 [0.482
Random	Random	0.0179 [0.0000, 0.0536]	0.1429 [0.0536, 0.2321]	0.2857 [0.1786, 0.4107]	0.7500 [0.6250
Oracle gene expression	Observed expression	0.1071 [0.0357, 0.1964]	0.2500 [0.1429, 0.3571]	0.4286 [0.3036, 0.5536]	0.8214 [0.7143
	PCA expression	0.0714 [0.018, 0.143]	0.2143 [0.107, 0.321]	0.3929 [0.268, 0.518]	0.8036 [0.696
	AIDO.Cell embedding	0.0357 [0.000, 0.089]	0.3214 [0.196, 0.446]	0.5179 [0.393, 0.643]	0.8393 [0.732

Table 11. Recovering known drug–target relationships using different perturbation representations. AUROC and AUPRC are calculated using ground-truth and predicted bipartite drug–target graphs, using distance thresholding to induce predictions. Performance is evaluated on DTR-Bench with bootstrap confidence intervals (10,000 resamples; 332×194 pairs).

	AUROC (95% CI)	AUPRC (95% CI)
CellVS-Net predicted networks	**0.5399 [0.5165, 0.5631]**	**0.0123 [0.0093, 0.0176]**
Predicted expression	0.4822 [0.4572, 0.5070]	0.0085 [0.0075, 0.0096]
Predicted PCA expression	0.4763 [0.4509, 0.5015]	0.0089 [0.0078, 0.0101]
Predicted AIDO.Cell embeddings	0.5005 [0.4752, 0.5256]	0.0093 [0.0082, 0.0106]
SPRINT (McNutt et al.)	0.4447 [0.4206, 0.4691]	0.0076 [0.0068, 0.0086]
Random	0.4886 [0.4637, 0.5126]	0.0084 [0.0075, 0.0094]
Oracle AIDO.Cell (Ho et al., 2024)	0.5208 [0.4977, 0.5445]	0.0104 [0.0087, 0.0130]
Oracle PCA expression	0.5207 [0.4967, 0.5453]	0.0096 [0.0084, 0.0111]
Oracle observed expression	0.5130 [0.4890, 0.5377]	0.0092 [0.0082, 0.0105]

Method	AUROC	AUPRC	Drug→Target Hits @1	Drug→Target Hits @5	Drug→Target Hits @10	Drug→Target Hits @50	Target→Drug Hits @1	Target→Drug Hits @5	Target→Drug Hits @10	Target→Drug Hits @50
CellVS-Net predicted networks	**0.5399** (0.5165-0.5631)	**0.0123** (0.0093-0.0176)	0.0000 (0.0000-0.0000)	0.0272 (0.0121-0.0453)	0.0332 (0.0151-0.0544)	0.3230 (0.2719-0.3716)	0.0261 (0.0052-0.0518)	0.0468 (0.0207-0.0777)	0.0675 (0.0363-0.1036)	0.3108 (0.2435-0.3782)
Predicted expression	0.4822 (0.4756-0.5251)	0.0085 (0.0079-0.0101)	0.0150 (0.0030-0.0242)	0.0480 (0.0121-0.0453)	0.0781 (0.0453-0.1027)	0.3463 (0.2779-0.3807)	0.0155 (0.0000-0.0361)	0.0464 (0.0206-0.0773)	0.0875 (0.0206-0.0825)	0.2269 (0.1701-0.2887)
Predicted PCA expression	0.4763 (0.4633-0.5120)	0.0089 (0.0078-0.0102)	0.0179 (0.0000-0.0211)	0.0510 (0.0302-0.0755)	0.0781 (0.0514-0.1088)	0.3281 (0.2628-0.3656)	0.0154 (0.0000-0.0361)	0.0362 (0.0103-0.0567)	0.0670 (0.0412-0.1186)	0.2010 (0.1959-0.3196)
Predicted AIDO.Cell embeddings	0.5005 (0.4861-0.5344)	0.0093 (0.0081-0.0108)	0.0061 (0.0000-0.0211)	0.0393 (0.0121-0.0514)	0.0756 (0.0393-0.0906)	0.2567 (0.2236-0.3172)	0.0261 (0.0000-0.0258)	0.0777 (0.0361-0.1031)	0.1395 (0.0773-0.1649)	0.3351 (0.2371-0.3660)
SPRINT (McNutt et al.)	0.4447 (0.4206-0.4691)	0.0076 (0.0068-0.0086)	**0.0332** (0.0151-0.0544)	**0.1025** (0.0695-0.1360)	**0.1507** (0.1118-0.1903)	0.3048 (0.2568-0.3565)	0.0051 (0.0000-0.0155)	0.0514 (0.0206-0.0876)	0.0773 (0.0412-0.1186)	0.2780 (0.2165-0.3402)
Random	0.4886 (0.4637-0.5126)	0.0084 (0.0075-0.0094)	0.0030 (0.0000-0.0090)	0.0331 (0.0151-0.0542)	0.0630 (0.0392-0.0904)	0.3253 (0.2771-0.3765)	0.0052 (0.0000-0.0155)	0.0155 (0.0000-0.0361)	0.0414 (0.0155-0.0722)	0.2734 (0.2113-0.3351)
Oracle AIDO.Cell (Ho et al., 2024)	0.5208 (0.4977-0.5445)	0.0104 (0.0087-0.0130)	0.0151 (0.0030-0.0301)	0.0484 (0.0271-0.0723)	0.0634 (0.0392-0.0904)	0.3135 (0.2651-0.3645)	0.0206 (0.0052-0.0412)	0.0927 (0.0515-0.1340)	0.1236 (0.0773-0.1701)	0.3451 (0.2784-0.4124)
Oracle PCA expression	0.5207 (0.4967-0.5453)	0.0096 (0.0084-0.0111)	0.0120 (0.0030-0.0242)	0.0422 (0.0211-0.0665)	0.0723 (0.0453-0.1027)	**0.3384** (0.2870-0.3897)	**0.0516** (0.0206-0.0876)	**0.1136** (0.0722-0.1598)	**0.1392** (0.0928-0.1907)	**0.3718** (0.3041-0.4383)
Oracle observed expression	0.5130 (0.4890-0.5377)	0.0092 (0.0082-0.0105)	0.0151 (0.0030-0.0302)	0.0332 (0.0151-0.0544)	0.0757 (0.0483-0.1057)	0.3293 (0.2779-0.3807)	0.0207 (0.0052-0.0412)	0.0569 (0.0258-0.0928)	0.1084 (0.0670-0.1546)	0.2943 (0.2320-0.3608)

Table 12. Recovering known drug–target relationships using different perturbation representations. AUROC and AUPRC are calculated using ground-truth and predicted bipartite drug–target graphs, using distance thresholding to induce predictions. Query-level recall rates (Hits@k) are reported for both drug→target and target→drug retrieval tasks as Drug Hits and Target Hits respectively. Expression-based representations were derived from LINCS L1000 small molecule and shRNA data. PCA applies a 50-component PCA to this full dataset. AIDO.Cell embeds each sample.

Here, contexts are one-hot encoded, containing no prior knowledge of cell line similarity, intentionally disadvantaging contextualized networks, which must learn to share information and extrapolate between modeling tasks from scratch. We also strip away perturbations and focus only on control measurements for each cell line to isolate the role of context sharing.

Table 13. Mean-squared error (MSE) of inferred transcriptional networks on a sample-held-out split for control measurements from all cell lines. CellVS-Net and group-specific models use one-hot encoded celltype contexts. Full Test contains all held-out samples. $n_c > 3$ assesses conditions with more than 3 observations, while $n_c \leq 3$ assesses conditions with less than 3 observations.

	Full Test	$n_c > 3$	$n_c \leq 3$
Population	0.9859	0.9860	0.9184
Group-specific	0.6680	0.6668	3.9306
CellVS-Net	0.6169	0.6169	**0.7011**
+ dose, time	**0.6065**	**0.6064**	0.8912

Table 13 shows that CellVS-Net achieves the best performance on the full dataset by mitigating the failure modes of the population and condition-specific baselines. Population models suffer from high bias, underfitting due to their inability to model cell line-specific effects, while cell line-specific models dramatically overfit on conditions with few samples ($n_c \leq 3$), with MSE exploding in low-sample regimes. In contrast, CellVS-Net automatically interpolates between a population-like default when data are scarce and cell line-specific behavior when sufficient data are available, yielding stable performance across data regimes that more closely resemble the long-tail distribution of a virtual screening atlas.

Table 14. MSE of inferred networks on a sample-held-out split for perturbed expression measurements. Perturbation contexts are one-hot encoded, while different encoding schemes are used for cell line contexts.

Model Variant	Mean Squared Error
Population	0.9735
Group-specific	4.795e04
CellVS-Net onehot	0.6304
CellVS-Net + dose, time, celltype	**0.6122**

Next, we evaluate the impact of richer context features that are essential for extrapolating to unseen conditions. Continuous covariates such as dose and time, or high-dimensional summaries of cell state, are difficult to incorporate into discrete group-based models, which typically require hand-crafted bins or separate models per group. In a virtual screening setting, however, new compounds will often be proposed at doses and timepoints that do not exactly match those in the training data, and any useful model must interpolate smoothly across these axes.

To study this, we move from control-only networks to prediction of post-perturbation networks and incrementally augment the input features of the context encoder (Table 14). We represent small-molecule identities with one-hot encodings and vary the representation of the cell-type context from a one-hot label to embeddings of the unperturbed transcriptomic profile. Post-perturbation prediction is more challenging than the control-only setup in Table 13, yet CellVS-Net again avoids extreme over- and under-fitting. Replacing one-hot cell-type indicators with control expression and augmenting with dose and time substantially improves generalization for predicting post-perturbation networks. These results support the view that rich, continuous context encodings are necessary for CellVS-Net to achieve the smooth extrapolation across doses, timepoints, and cell types that virtual screening requires.

Model	@1	@5	@10	@25
CellVS-Net (separate)	0.1250	0.4464	0.6250	0.8571
Joint CellVS-Net (all types)	0.0714	0.2857	0.5179	0.7857

Table 15. DDR-Bench drug-disease retrieval performance (from Table 2).

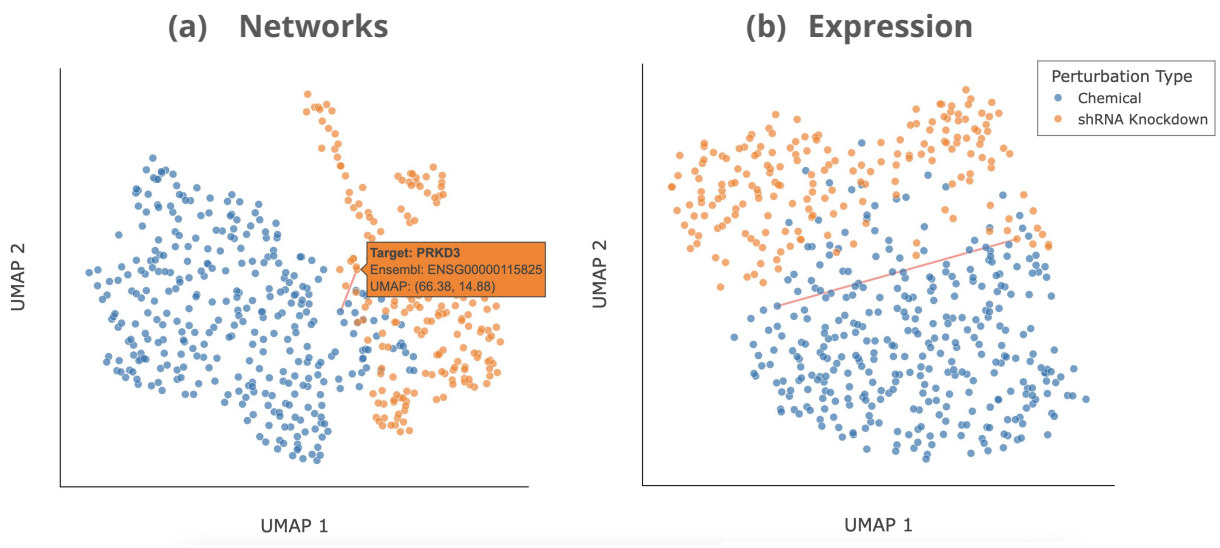

Figure 5. UMAP projection of (a) CellVS-Net networks and (b) expression snapshots for chemical and shRNA perturbations. We provide an interactive web tool on GitHub to explore this embedding space and highlight known interactions for a given drug or target. The example in red highlights one known drug–target pair: PRKD3 (shRNA knockdown) and Midostaurin (drug).

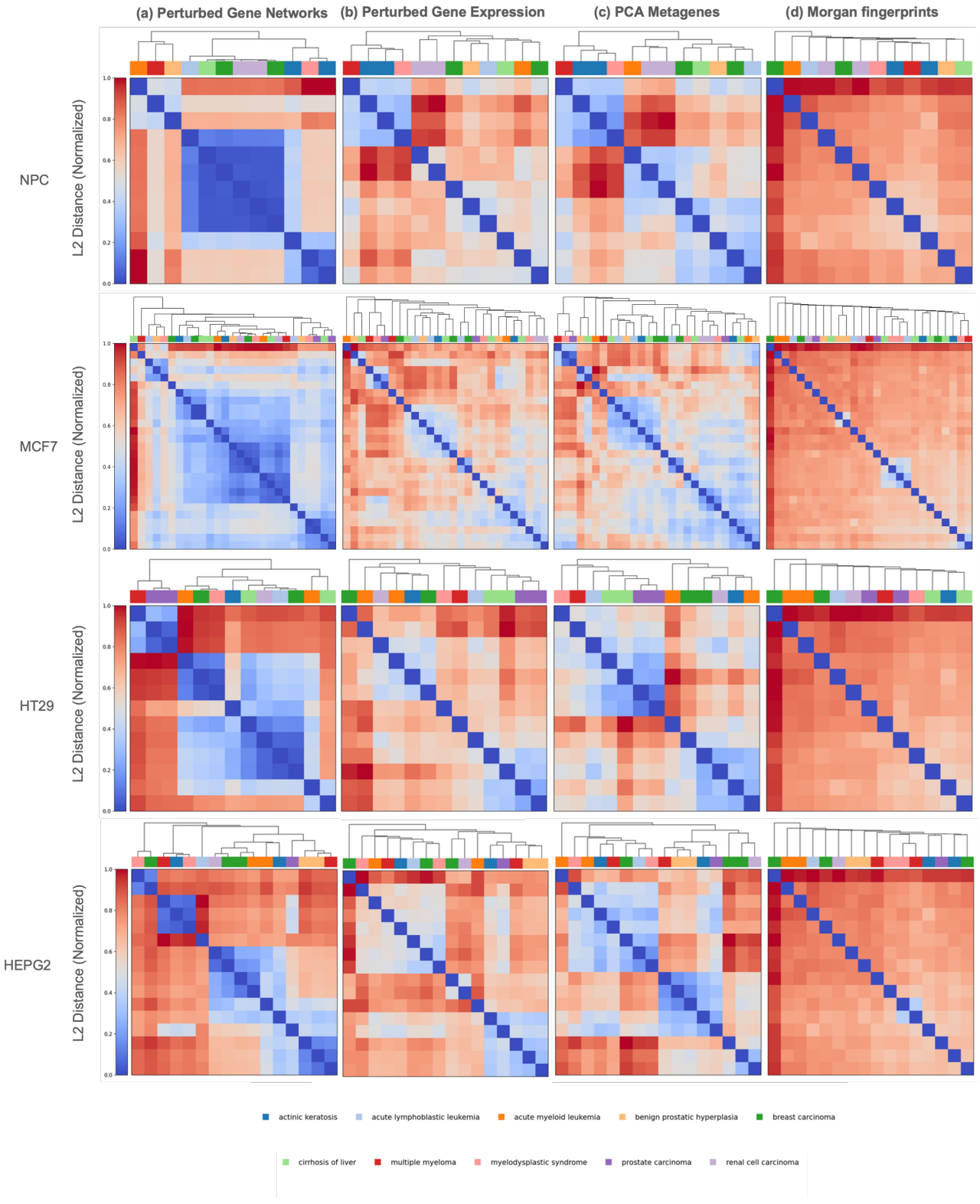

Figure 6. Organization of drugs based on four representations across cell types.

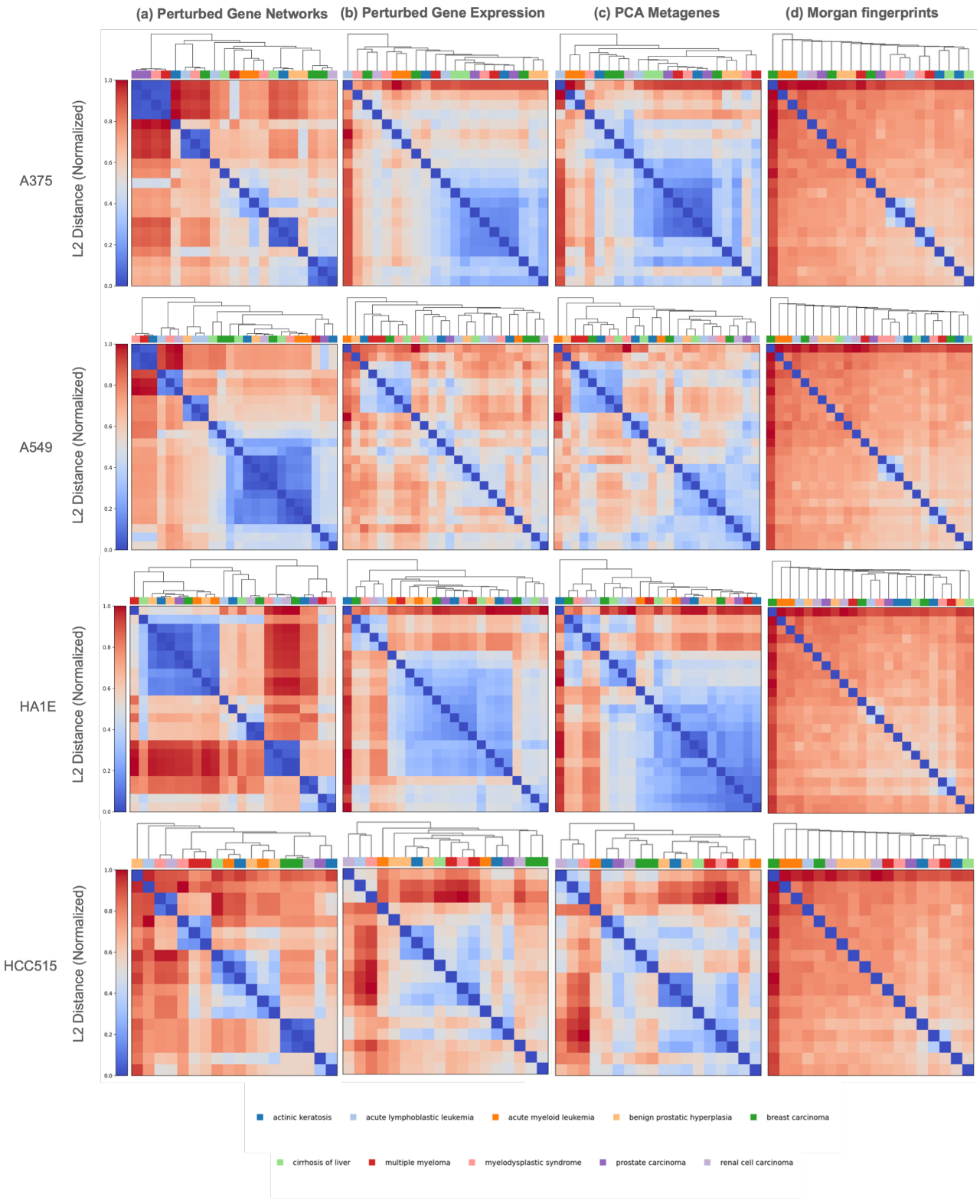

Figure 7. Organization of drugs based on four representations across cell types.