Virtual Screening on Cellular Systems

Caleb N. Ellington [*1] **Sohan Addagudi** [*1] **Jiaqi Wang** [*2] **Benjamin Lengerich** [3] **Eric P. Xing** [1 4 5]

Abstract

Virtual screening methods prioritize therapeutic candidates by predicting drug properties and molecular interactions, but are unable to predict complex, system-level responses that cause late-stage failures in drug discovery. Virtual cells have been posed as a solution 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. To address this, we develop **CellVS-Net**, a model for sample-specific estimation of Gaussian graphical models, allowing us to infer how gene-gene networks restructure under a variety of complex multivariate contexts (cell type, drug dose), improving representation of drug effects for early-stage drug discovery. CellVS-Net accurately predicts gene network restructuring for unseen perturbations covering small molecules, large molecules, genetic knockdowns, and overexpressions, providing a new way to compare therapeutic effects from multiple perturbation modalities. To evaluate CellVS-Net's utility and move virtual screening toward understanding drug efficacy, we introduce two benchmarks: **DDR-Bench**, which evaluates a method's ability to predict effective diseases for drugs with unseen target profiles, and **DTR-Bench**, which reconstructs global drug-target interactions for mechanistic understanding. Compared to both molecular and expression representations of drugs, CellVS-Net achieves SOTA on both benchmarks. Overall, CellVS-Net is posed to enable a simulation-based approach to preclinical screening, and associated benchmarks enable hill-climbing on relevant tasks. We provide source code for models and data curation, as well as Kaggle-style submission functions and running public leaderboards.

[1]Carnegie Mellon University, Pittsburgh, PA, USA [2]University of Washington, Seattle, WA, USA [3]University of Wisconsin–Madison, Madison, WI, USA [4]MBZUAI, Abu Dhabi, UAE [5]GenBio AI, USA. Correspondence to: Eric P. Xing <epxing@cs.cmu.edu>.

Preprint. January 31, 2026.

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 during training.

A growing number works on virtual cells aim to extend these efforts by training machine learning models to simulate 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 assay 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.). However, expression – predicted and real – is an intermediate representation of perturbation effect, and does not directly identify the safety and efficacy of a therapeutic. One notable work by Miladinovic et al. (Miladinovic et al., 2025) extends beyond reconstruction-based evaluations to investigate synonymous genetic and pharmacological perturbations, but 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 rele-

vant to drug discovery and (ii) generalize arbitrarily to unseen therapeutics to promote virtual exploration. To address this, we curate the DDR-Bench benchmark, combining the OpenTargets and LINCS databases into a single dataset for predicting disease indications for drugs with novel targets, enabling hill climbing on clinically-relevant applications.

Based on the shortcomings of expression snapshots on this benchmark, we also propose gene-gene networks as a natural framework for modeling how perturbations rewire cellular circuitry. However, most network inference methods rely on partitioned cohorts (Badia-i Mompel et al., 2023; Stone et al., 2021), which fail to capture the continuous and context-dependent rewiring observed in many diseases (Ursu et al., 2022; Hoadley et al., 2018), and plug-in estimators cannot generalize to new conditions. More generally, the increase of dataset complexity, heterogeneity, and size, has motivated the development of methods of "personalized" models across several application areas (Buettner et al., 2015; Fisher et al., 2018; Hart, 2016; Ng et al., 2015). "Personalized" models seek to represent heterogeneous distributions as sample-specific distributions $X_i \sim P_i(X)$, where i indexes a sample X_i and P_i corresponds to the sample-specific distribution. In the most difficult case of sample-specific inference, each P_i is observed only a single time and hence information must be shared across samples.

Toward this aim of sharing information across samples, most personalized models make the simplifying assumption that all P_i belong to the same family; i.e. $X \sim P(X \mid \theta_i)$. Through this lens of personalized modeling, understanding sample heterogeneity is reframed as estimating data distributions with sample-specific parameters. Some methods provide sample-specific estimators without additional information by imposing strong biological priors (**?**) or using a sample-left-out approach (Kuijjer et al., 2019; Saha et al., 2023), but these lack desirable properties such as the ability to generalize to new samples or even test model performance on held-out data. Due to the difficulty of estimating sample-specific parameters, most methods make use of side information (e.g. sample metadata) as a contextual representation of sample-to-sample variation (Kolar et al., 2008; Ahmed & Xing; Parikh et al., 2011; Lengerich et al., 2018; Hastie & Tibshirani, 1993). Given observations X and contextual metadata C, we have

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

where $P(X \mid \theta)$ defines the context-specific model, and $P(\theta \mid C)$ defines a context-specific density of model parameters θ, which we call the context encoder. One of the earliest ways to apply context encoding toward sample-specific parameter inference was the linear varying-coefficient (VC) model (Hastie & Tibshirani, 1993) in which linear regression parameters are predicted from context using a learned linear mapping or kernel density estimator. Extensions of this regime have been widespread (Kolar et al., 2008; Ahmed & Xing; Parikh et al., 2011; Wang et al., 2022), but typically focus on allowing models to vary over only a few continuous covariates (Fan & Zhang, 1999; Hastie & Tibshirani, 1993; Wang et al., 2022), or a small number of groups (Kolar et al., 2008; Parikh et al., 2011; Ahmed & Xing)

Contextualized modeling (Al-Shedivat et al., 2020; Lengerich et al.), combines the adaptability of VC models with the power of modern deep learning architectures by implementing the context encoder as a Dirac delta distribution defined by a deterministic deep neural network f,

$$P(\theta \mid C) = \delta(\theta - f(C)),$$

thus benefiting from a wide range of architectures targeting high-dimensional and complex data types. When contexts are unique to each sample, the inferred models are sample-specific.

$$P(X_i \mid \theta_i) = P(X_i \mid f(C_i))$$

Recently, contextualized modeling was extended beyond linear models to several types of graphical models (Ellington et al., b). However, sample-specific multivariate Gaussian graphical models remain unaddressed, despite their significance. In this work, we develop a convex multivariate Gaussian loss which enables full-rank covariance matrix prediction, and is a drop-in replacement for models which currently optimize for expression using mean squared error losses, which conveniently equate to an isotropic Gaussian loss. 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.

Methods

Multivariate Gaussian loss We seek a context-specific density of a multivariate Gaussian $P(\mu, \Sigma \mid C)$ such that

$$P(X \mid C) = \int_\theta d\theta N(X \mid \mu, \Sigma) P(\mu, \Sigma \mid C)$$

is maximized, where $N(X \mid \mu, \Sigma)$ is the probability of gene expression $X \in \mathbb{R}^p$ under the Gaussian graphical model with parameters $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$, and C is

sample context which can contain both multivariate and real features. Using the delta distribution and context encoder, this becomes

$$P(X_i \mid \mu_i, \Sigma_i) = P(X_i \mid f(C_i)).$$

The key challenge is to define a differentiable loss function ℓ that is proportional to the negative log probability of our contextualized multivariate Gaussian and is convex in its parameters, such that this loss function can be used in end-to-end gradient-based optimization in combination with deep learning architectures.

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

Optimizing the covariance matrix and its inverse presents a challenging non-convex problem. To address this, we instead optimize a convenient linear parameterization of the covariance matrix based on the definition of Pearson's correlation coefficient ρ^2.

$$\rho^2 = \frac{\Sigma_{ij}^2}{\Sigma_{ii}^2\Sigma_{jj}^2} = \frac{\Sigma_{ij}}{\Sigma_{ii}^2}\frac{\Sigma_{ij}}{\Sigma_{jj}^2} = \beta_{ij}\beta_{ji}$$

Where β_{ij} is the covariance $\text{Cov}(X_i, X_j)$ over the variance of the outcome $\text{Var}(X_j)$, and also the well-known ordinary least squares (OLS) solution to regressing X_i onto X_j.

$$\widehat{\beta}_{ij} = \underset{\beta}{\text{argmin}} \|X_i - \beta X_j\|_2^2$$

The convenient marginalization properties of Gaussians allow that each i, j pair can be solved independently, which can then be compiled piecewise into a full correlation matrix. To impute the full multivariate Gaussian parameters also requires the feature variance $\sigma \in \mathbb{R}_+^p$

$$\Sigma_{ij} = \text{sign}(\beta_{ij})\sigma_i^2\sigma_j^2\sqrt{\beta_{ij}\beta_{ij}},$$

which can be easily estimated with the isotropic Gaussian loss along with μ

$$-\log p(X_i \mid \mu_i, \sigma_i) \propto \frac{1}{2\sigma^2}(X_i - \mu_i)^2 + \frac{1}{2}\log(2\pi\sigma^2),$$

In matrix form, the full multivariate Gaussian loss for $X \in \mathbb{R}^p$, $\mu \in \mathbb{R}^p$, $\sigma^2 \in \mathbb{R}_+^p$, $\beta \in \mathbb{R}^{p\times p}$ is

$$\ell(X, \mu, \sigma, \beta) = \frac{1}{p^2}\sum_{i,j}((X_j - \mu_j) - (X_i - \mu_i)\beta_{ij})^2 + \frac{1}{2}\sum_{i=1}^{d}\log(2\pi\sigma_i^2) + \frac{1}{2}\sum_{i=1}^{d}\frac{(X_i-\mu_i)^2}{\sigma_i^2},$$

and its contextualized variant is

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

for context encoders $\mu(C)$, $\sigma(C)$, and $\beta(C)$. This differentiable objective can be used in end-to-end optimization to solve for the context encoders and context-specific parameters simultaneously, producing full-rank context-specific Gaussian graphical models. For identifiability, a two-step estimation procedure is required, first optimizing for $\mu(C)$ and then for $\sigma(C)$ and $\beta(C)$, but in practice we see similar results when optimizing them jointly. It is noteworthy that this loss can serve as a drop-in replacement for mean-squared error loss, commonly used in cell modeling. In this work, we commonly refer to the pairwise regression error term as mean squared error (MSE), which we use on its own to measure the network's goodness of fit.

CellVS-Net Drug efficacy, toxicity, and mechanism of action, however, 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 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. We test CellVS-Net as a lightweight late-fusion head on top of various frozen foundtion models for cell and perturbation modeling. Implementation details on the architecture and optimizers used are discussed in the Appendix.

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 reused for evaluation of molecule-based, target-based, and cell-based screening methods.

Drug-Disease Retrieval First, we apply the combined OpenTargets-LINCS dataset to create the Drug-Disease-Retrieval Benchmark (DDR-Bench). DDR-Bench evaluates a method's ability to identify potential therapeutic applications for a drug with an unseen target profile. Specifically, given a drug with a held-out target profile, can another drug be retrieved from a reference set which has the same FDA-approved disease indication, despite having different targets? This screening task addresses a failure case of target-based screening, since simple target matching is not a viable approach when a drug is introduced with a unique target profile. Intuitively, we believe that succeeding on this screening task would require a system-level model of drug response.

To construct the benchmark, we filter the combined OpenTargets-LINCS dataset to only consider small molecules which are represented in LINCS chemical perturbations. Then, we filter to a set of diseases which have at least 2 Phase-IV FDA-approved drugs, each with unique target signatures (known to bind to a unique set of proteins) to produce the DDR-Bench dataset (Table 8). For evaluation, we hold out one target signature at a time and query the remaining reference set to evaluate a method's ability to retrieve the correct disease indication when target matching is not an option. If the query returns a drug with an identical disease indication in the top k retrievals, we consider that a positive "Hit". We run this retrieval task for $k \in \{1, 5, 10, 25\}$, and report Hits@k.

We chose a retrieval-based formulation of this task for four reasons: (i) it is modality-agnostic and only requires ranking the reference set against the query, allowing us to uniformly benchmark molecule-based and cell-based methods, (ii) it does not require sample splitting for e.g. supervised prediction, using the entire small set of disease labels for evaluation, (iii) the reference set can be continuously updated with zero overhead as new drugs are approved, and (iv) as a byproduct, the evaluation creates a search utility similar to those used for semantic search in natural language processing, which can be rapidly scaled to billions of drug entries with little computational overhead (McNutt et al.). By default, our evaluation accommodates any vector representation of a drug (e.g. fingerprint, expression) and ranks the reference set in terms of L2 distance to the query drug.

Drug-Target Retrieval To construct DTR-Bench, we filter the combined OpenTargets-LINCS dataset to drug-target pairs where both the drug is represented in LINCS chemical perturbations and the target is represented in LINCS genetic knockdown experiments. We apply quality control filters to both modalities based on replicate correlation and self-ranking performance to ensure high-confidence perturbation profiles. In total, we obtain 559 drug-target pairs spanning 332 unique drugs and 194 unique targets (Table 7).

We perform two sets of evaluations, again leveraging vector-based representations of drug and genetic target perturbations. First, we threshold pairwise distances between drugs and targets to construct a bipartite drug-target interaction graph and compare this to a true graph of drug-target interactions. We report AUROC and AUPRC for edge classification by scanning over distance thresholds. Second, we perform a retrieval task in both directions: drug-to-target and target-to-drug. For each query, we rank all candidates in the opposite therapeutic modality using L2 distance between these vectors and the query drug or target vector. We compute Hits@k for $k \in \{1, 5, 10, 50\}$ to evaluate ranking quality. This retrieval formulation maintains the same modality-agnostic benefits as DDR-Bench while specifically evaluating cross-modal mechanism-of-action prediction. We demonstrate retrieval as a search utility for drug-target mapping in Figure 2.

Different drug modalities require different assays to measure, making expression-based models and baselines susceptible to batch effects. To to correct for these effects prior to evaluation, we perform PCA on the combined drug and target representations immediately before evaluation and remove up to the first 3 principal components. We report the best performance from these 3 attempts.

Representing Drug Effects

Across DDR-Bench and DTR-Bench, all methods reduce each perturbation to a fixed-length vector representation and compare drugs or targets using Euclidean distance in that shared space. This formulation is deliberately modality-agnostic: any representation that maps a perturbation to a vector can be plugged into the same retrieval and edge-classification protocols. In the results, we therefore interpret performance differences entirely in terms of how well each representation captures therapeutically meaningful cell-level effects.

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.

Second, we consider post-perturbation gene expression as an "oracle" virtual cell. In a realistic virtual screening scenario,

these expression measurements would *not* be available: the goal is precisely to avoid running large numbers of physical experiments. However for evaluation, we 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.

Third, we include a foundation model (FM) embedding baseline, where we apply a pretrained gene expression foundation model to the observed post-perturbation expression and use its latent embeddings as perturbation representations. We include this baseline to simulate another approach to using an oracle virtual cell, using the latent embedding features instead of generated expression. Embedding-based can outperform other methods on these tasks by removing redundant expression features and distilling cell states into semantically meaningful low-dimensional latent features. We further compare this to PCA-compressed expression, which provides semantically meaningful low-dimensional features through simple linear compression.

Results

Estimating Perturbation-specific Gene Networks

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. To stress-test this approach, we compare a contextualized network estimator to population and group-specific networks in a minimal regime which is explicitly biased against the contextualized estimator. 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 1. 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	0.978	0.980	**0.681**
Group-specific	51.576	**0.662**	1.38e6
CellVS-Net	**0.669**	0.665	0.730
+ dose, time	0.6433	0.638	0.767

Table 1 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 2. 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	0.9721
Group-specific	2.12e6
CellVS-Net onehot	0.572
CellVS-Net + dose, time, celltype	**0.541**

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 handcrafted 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

Table 3. 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.

Model Type	Context Encoder	Chemical	shRNA	Over Expression	Ligand
Population	None	1.0594	0.9740	0.7769	0.9315
CellVS-Net Molecule	Morgan Fingerprint (Onengut-Gumuscu et al., 2015)	0.5787	—	—	—
	ChemBERTa (Singh et al., 2025)	0.5728	—	—	—
CellVS-Net Target	AIDO.Structure (Zhang et al., 2024)	**0.5389**	0.6817	**0.6678**	0.5774
	AIDO.Protein (Sun et al., 2024)	0.5432	0.6819	0.6767	**0.5772**
	AIDO.Cell (Ho et al., 2024)	0.5687	**0.6815**	0.7681	0.6016
	AIDO.DNA (Ellington et al., 2024)	0.5757	0.6831	0.7880	0.6012
	Gene PCA	0.5824	0.6833	0.7389	0.6077

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 2). 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 1, 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.

Generating Gene Networks On-demand for Unseen Therapies

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, croup-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 3).

CellVS-Net strongly outperforms the context-agnostic baseline by learning to map cell type and perturbation contexts to gene network rewiring. When considering small-molecule perturbations, CellVS-Net generalizes effectively to held-out molecules. Learning a model of how small molecules affect cellular systems reduces error by 35%. Representing molecules with their known target protein improves performance even further, reducing error by 41%. All other perturbation types are represented only by their genetic target (shRNA, over expression) or large molecule protein (ligand). We include a non-pretrained context representation in both cases to evaluate the importance of pretrained representations for generalization. Molecules see minimal improvement from pretraining while molecules see substantial improvement in most cases. Representing drugs with their known protein targets also improves over molecule-only approaches. In general, CellVS-Net performs best with target representations from AIDO.StructureEncoder (Zhang et al., 2024). For a full overview of context representations, see Methods.

Disease Retrieval: Predicting Disease Indications for Drugs with Novel Targets Cell-level virtual screening approaches should induce a reliable similarity among perturbations with similar cell-level effects, even if they hit different targets. To evaluate this, we gather a dataset of small molecule drugs from the OpenTargets platform that have different molecular target profiles, but are approved for

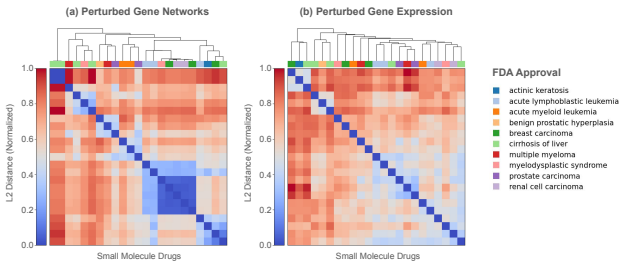

Figure 1. Organization of drugs based on (a) perturbed gene networks and (b) perturbed gene expression representations. Drugs are annotated with their FDA-approved disease indications. All samples are taken from the PC3 cell type.

Table 4. 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.

		Disease Hits			
Representation	Method	@1	@5	@10	@25
Perturbed gene network	CellVS-Net	**0.1250**	**0.4464**	**0.5893**	**0.8571**
Perturbed gene expression	Predicted expression	0.0714	0.1786	0.3571	0.6964
	Predicted expression PC loadings	0.0536	0.1250	0.3214	0.6786
	Predicted AIDO.Cell embeddings	0.0536	0.3214	0.4286	0.7143
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	0.0357	0.1071	0.2500	0.6429
Oracle gene expression	Observed expression	0.0893	0.2679	0.3929	0.7500
	PCA expression	0.0893	0.2679	0.4107	0.8214
	AIDO.Cell embedding (Ho et al., 2024)	0.0536	0.3393	0.5714	0.8393

a common disease. This dataset establishes a ground truth for drugs that have similar therapeutic effects, but which are considered unrelated by target-centric virtual screening approaches. We find that CellVS-Net is more representative of cell-level therapeutic effects than other methods (Table 4).

We reduce each perturbation to a fixed-length vector representation and compare drugs or targets using Euclidean distance in that shared space. This formulation is deliberately modality-agnostic: any representation that maps a perturbation to a vector can be plugged in. In the results, we therefore interpret performance differences entirely in terms of how well each representation captures therapeutically meaningful cell-level effects. We show an example of how perturbations organize under different representations in Figure 1.

In addition to CellVS-Net, we apply several molecular and cellular baselines. We use post-perturbation gene expression as an "oracle" virtual cell. In a realistic virtual screening scenario, these expression measurements would not be available: the goal is precisely to avoid running large numbers of physical experiments. However for evaluation, we 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 – 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 include PCA transformations and AIDO.Cell embeddings of this expression data.

Drug-Target Retrieval: Matching Synonymous Perturbations Across Modalities. Cell-level virtual screening should induce a reliable similarity among perturbations with similar cell-level effects, even if they come from different classes of drugs. Specifically, we ask two questions (i) Can known drug-target interactions be reconstructed from cell-level representations of small-molecule drug perturbations and genetic target perturbations? (ii) Given a small molecule drug, can we retrieve its protein target based on similarity of their perturbation effects, and vice versa? This bidirectional retrieval task enables investigation of therapeutic mechanisms by connecting drug-induced expression changes to genetic knockdown phenotypes, providing a system-level validation of proposed drug-target relationships. Simultaneously, this task connects cell-level virtual screening to the traditional virtual screening task of drug-target interaction prediction. We validate this cross-modal drug-target lookup approach based on known gene targets for some well-characterized small molecules as well as genetic perturbations such as knock downs and hairpin RNAs (Table 5). We include SPRINT (McNutt et al.), a SOTA molecular screening method that was directly optimized for drug-target interaction prediction. CellVS-Net outperforms all baselines on global drug-target interaction graph reconstruction, including both oracle and prediction baselines.

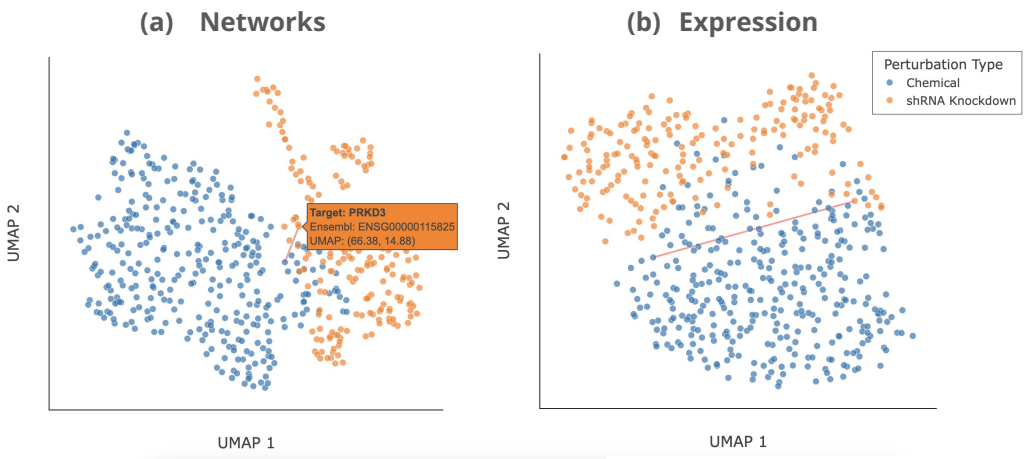

Figure 2. UMAP projection of (a) CellVS-Net networks and (b) expression snapshots for chemical and shRNA perturbations. We provide an interactive web tool on GitHub to explore this embedding space and highlight known interactions for a given drug or target. UMAP visualization provides a heuristic for showing relative distances between expression and networks 2D. We show an example highlighting one known drug-target pair in red: PRKD3 (shRNA knockdown) and Midostaurin (drug).

Lookup between inhibitory small molecules and genetic knockdowns provides a way to understand molecular mechanisms based on similar cell-level effects (Table 9). To provide easy access to CellVS-Net and explore results, we provide a web tool for exploring mappings between shRNA and chemical perturbations (Figure 2).

Table 5. 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. 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.

	AUROC	AUPRC
CellVS-Net predicted networks	**0.524**	**0.012**
Predicted expression	0.4659	0.0086
Predicted PCA expression	0.4657	0.0082
Predicted AIDO.Cell embeddings	0.4928	0.0093
SPRINT (McNutt et al.)	0.444	0.008
Random	0.470	0.008
Oracle AIDO.Cell (Ho et al., 2024)	0.521	0.010
Oracle PCA expression	0.522	0.010
Oracle observed expression	0.514	0.009

Discussion

90% of clinical drug development fails (Sun et al., 2022). 70-80% of all failures in this stage are attributed to lack of clinical efficacy (40-50%) or unmanageable toxicity 30%. Drugs that make it to human trials pass through a gauntlet of molecular modeling, cell screening, and animal studies.

In this work, we aim to move later cell screening-stage drug failures into earlier-stage virtual 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 itself introduces a new convex loss for multivariate Gaussians, and has desirable statistical properties for estimation in long-tail drug screening applications (Table 1). The result generalizes smoothly to unseen drugs, doses, and cell types, while improving with richer context representations, reducing MSE by 40-50% (Tables 2, 3).

In this study, we also define cell-level virtual screening through the formulation of the DDR-Bench and DTR-Bench benchmarks. Both of these benchmarks are method-agnostic, evaluating both drug, drug-target, and cell modeling methods across several data modalities on clinically grounded endpoints: recovering disease indications for drugs with unseen targets (Table 4) and reconstructing drug–target relationships from the effects of different perturbation modalities (Table 5). On DDR-Bench and DTR-Bench, CellVS-Net improves over molecular and expression baselines, including oracles representing ideal virtual cells.

More broadly, virtual screening on cellular systems offers a path to scale virtual screening to predict clinical stage failure modes. Unlike traditional virtual screening tools that only utilize molecular data, CellVS-Net is well-positioned to utilize a diversity of observational data under individual patient contexts if and when this becomes available. Overall, virtual screening on cellular systems promises to enable large-scale in silico exploration of small molecules, large molecules, and genetic perturbations through unifying cell-level representations. To support this trajectory and enable hill-climbing over time, we maintain a public leaderboard.

Impact Statement

"This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here."

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.

Ahmed, A. and Xing, E. P. Recovering time-varying networks of dependencies in social and biological studies. 106(29):11878–11883. doi: 10.1073/pnas.0901910106. URL https://www.pnas.org/doi/10.1073/pnas.0901910106.

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.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33 (2):155–160, February 2015. ISSN 1546-1696. doi: 10.1038/nbt.3102. URL https://www.nature.com/articles/nbt.3102. Number: 2 Publisher: Nature Publishing Group.

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.

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.

Fan, J. and Zhang, W. Statistical estimation in varying coefficient models. *The Annals of Statistics*, 27(5):1491–1518, October 1999. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1017939139. URL https://projecteuclid.org/journals/annals-of-statistics/volume-27/issue-5/Statistical-estimation-in-varying-coefficient-models/10.1214/aos/1017939139.full. Publisher: Institute of Mathematical Statistics.

Fisher, A. J., Medaglia, J. D., and Jeronimus, B. F. Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, 115(27):E6106–E6115, July 2018. doi: 10.1073/pnas.1711978115. URL https://www.pnas.org/doi/10.1073/pnas.1711978115. Publisher: Proceedings of the National Academy of Sciences.

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.

Hart, S. Precision Education Initiative: Moving Towards Personalized Education. *Mind, brain and education : the official journal of the International Mind, Brain, and Education Society*, 10(4):209–211, December 2016. ISSN 1751-2271. doi: 10.1111/mbe.12109. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5476312/.

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.

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V., Akbani, R., Bowlby, R., Wong, C. K., Wiznerowicz, M., Sanchez-Vega, F., Robertson, A. G., Schneider, B. G., Lawrence, M. S., Noushmehr, H., Malta, T. M., Stuart, J. M., Benz, C. C., and Laird, P. W. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, 173(2):

291–304.e6, April 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.03.022. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5957518/.

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.

Kolar, M., Song, L., Ahmed, A., and Xing, E. P. Estimating time-varying networks. ISBN: 0812.5087 Publication Title: arXiv [stat.ML], December 2008. URL http://arxiv.org/abs/0812.5087.

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.

Kuijjer, M. L., Tung, M. G., Yuan, G., Quackenbush, J., and Glass, K. Estimating Sample-Specific Regulatory Networks. *iScience*, 14:226–240, April 2019. ISSN 2589-0042. doi: 10.1016/j.isci.2019.03.021. URL http://dx.doi.org/10.1016/j.isci.2019.03.021.

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

Lengerich, B. J., Aragam, B., and Xing, E. P. Personalized regression enables sample-specific pan-cancer analysis. *Bioinformatics (Oxford, England)*, 34(13):i178–i186, July 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty250.

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.

Ng, K., Sun, J., Hu, J., and Wang, F. Personalized Predictive Modeling and Risk Factor Identification using Patient Similarity. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2015:132–136, 2015. ISSN 2153-4063.

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.

Onengut-Gumuscu, S., Chen, W.-M., Burren, O., Cooper, N. J., Quinlan, A. R., Mychaleckyj, J. C., Farber, E., Bonnie, J. K., Szpak, M., Schofield, E., Achuthan, P., Guo, H., Fortune, M. D., Stevens, H., Walker, N. M., Ward, L. D., Kundaje, A., Kellis, M., Daly, M. J., Barrett, J. C., Cooper, J. D., Deloukas, P., Todd, J. A., Wallace, C., Concannon, P., and Rich, S. S. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nature Genetics*, 47(4):381–386, April 2015. ISSN 1546-1718. doi: 10.1038/ng.3245. URL https://www.nature.com/articles/ng.3245. Number: 4 Publisher: Nature Publishing Group.

Parikh, A. P., Wu, W., Curtis, R. E., and Xing, E. P. TREEGL: reverse engineering tree-evolving gene networks underlying developing biological lineages. *Bioinformatics*, 27(13):i196–204, July 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr239. URL http://dx.doi.org/10.1093/bioinformatics/btr239.

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.

Saha, E., Fanfani, V., Mandros, P., Ben-Guebila, M., Fischer, J., Hoff-Shutta, K., Glass, K., DeMeo, D. L., Lopes-Ramos, C., and Quackenbush, J. Bayesian Optimized sample-specific Networks Obtained By Omics data (BONOBO). preprint, Genomics, November 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.11.16.567119.

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, D., Gao, W., Hu, H., and Zhou, S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica. B*, 12(7):3049–3062, July 2022. ISSN 2211-3835. doi: 10.1016/j.apsb.2022.02.002. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9293739/.

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.

Ursu, O., Neal, J. T., Shea, E., Thakore, P. I., Jerby-Arnon, L., Nguyen, L., Dionne, D., Diaz, C., Bauman, J., Mosaad, M. M., Fagre, C., Lo, A., McSharry, M., Giacomelli, A. O., Ly, S. H., Rozenblatt-Rosen, O., Hahn, W. C., Aguirre, A. J., Berger, A. H., Regev, A., and Boehm, J. S. Massively parallel phenotyping of coding variants in cancer with Perturb-seq. *Nature Biotechnology*, 40(6):896–905, June 2022. ISSN 1546-1696. doi: 10.1038/s41587-021-01160-7. URL https://www.nature.com/articles/s41587-021-01160-7. Number: 6 Publisher: Nature Publishing Group.

Wang, Z., Kaseb, A. O., Amin, H. M., Hassan, M. M., Wang, W., and Morris, J. S. Bayesian Edge Regression in Undirected Graphical Models to Characterize Interpatient Heterogeneity in Cancer. *Journal of the American Statistical Association*, 117(538):533–546, 2022. ISSN 0162-1459. doi: 10.1080/01621459.2021.2000866.

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.

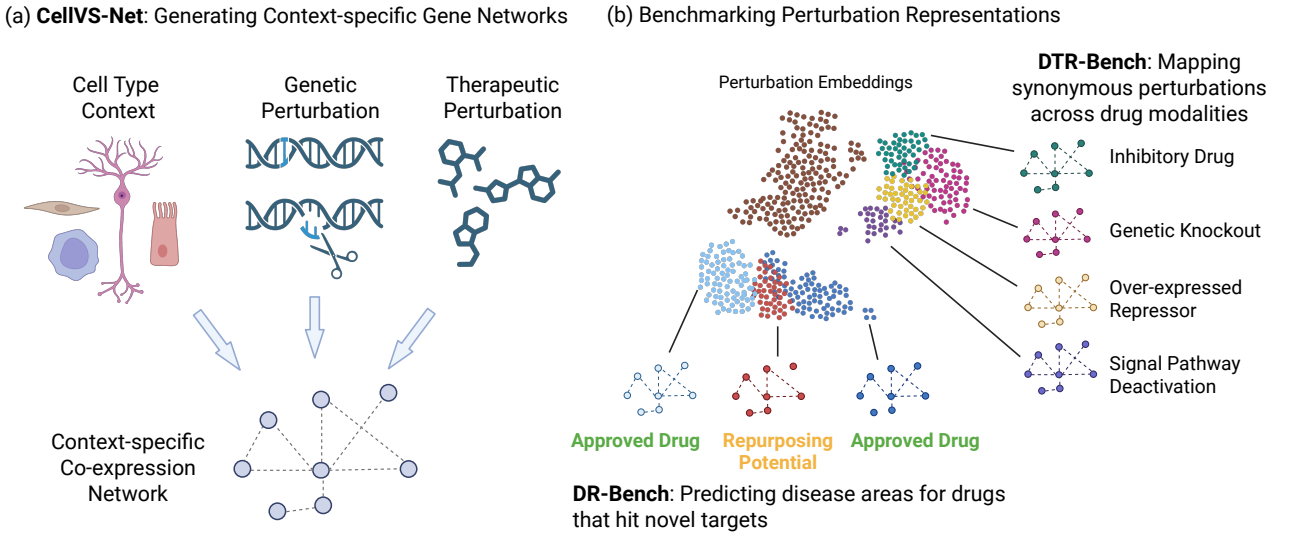

Figure 3. (a) CellVS-Net maps multivariate context (cell type, drug, dose) to context-specific gene networks. (b) We introduce two new benchmarks for evaluating drug representation approaches on clinically-relevant tasks. DDR-Bench predicts effective diseases for drugs with previously unseen target profiles. DTR-Bench maps synonymous perturbations across drug modalities to understand mechanism of action and off-target effects.

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.

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 a precomputed ChemBERTa embedding 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

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

			Drug→Target Hits				Target→Drug Hits			
	AUROC	AUPRC	@1	@5	@10	@50	@1	@5	@10	@50
CellVS-Net	**0.524**	**0.012**	**0.066**	0.096	0.120	**0.340**	0.016	0.046	0.062	0.268
AIDO.Cell (Ho et al., 2024)	0.521	0.010	0.015	0.048	0.063	0.313	0.021	0.093	0.124	0.345
PCA expression	0.522	0.010	0.012	0.042	0.072	**0.340**	**0.052**	**0.113**	**0.139**	**0.371**
Expression	0.514	0.009	0.015	0.033	0.075	0.331	0.021	0.057	0.108	0.294
Random	0.470	0.008	0.003	0.015	0.060	0.334	0.010	0.041	0.046	0.247
SPRINT (McNutt et al.)	0.444	0.008	0.033	**0.105**	**0.154**	0.307	0.005	0.052	0.077	0.278
Pred PCA from gene embs	0.4657	0.0082	0.0120	0.0633	0.1145	0.3825	0.0155	0.0258	0.0515	0.2010
Pred AIDO from gene embs	0.4928	0.0093	0.0211	0.0422	0.0813	0.3012	0.0103	0.0619	0.1031	0.2887
Pred expr from gene embs	0.4659	0.0086	0.0090	0.0512	0.0873	0.3825	0.020	0.041	0.0619	0.1856

Table 9. 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.

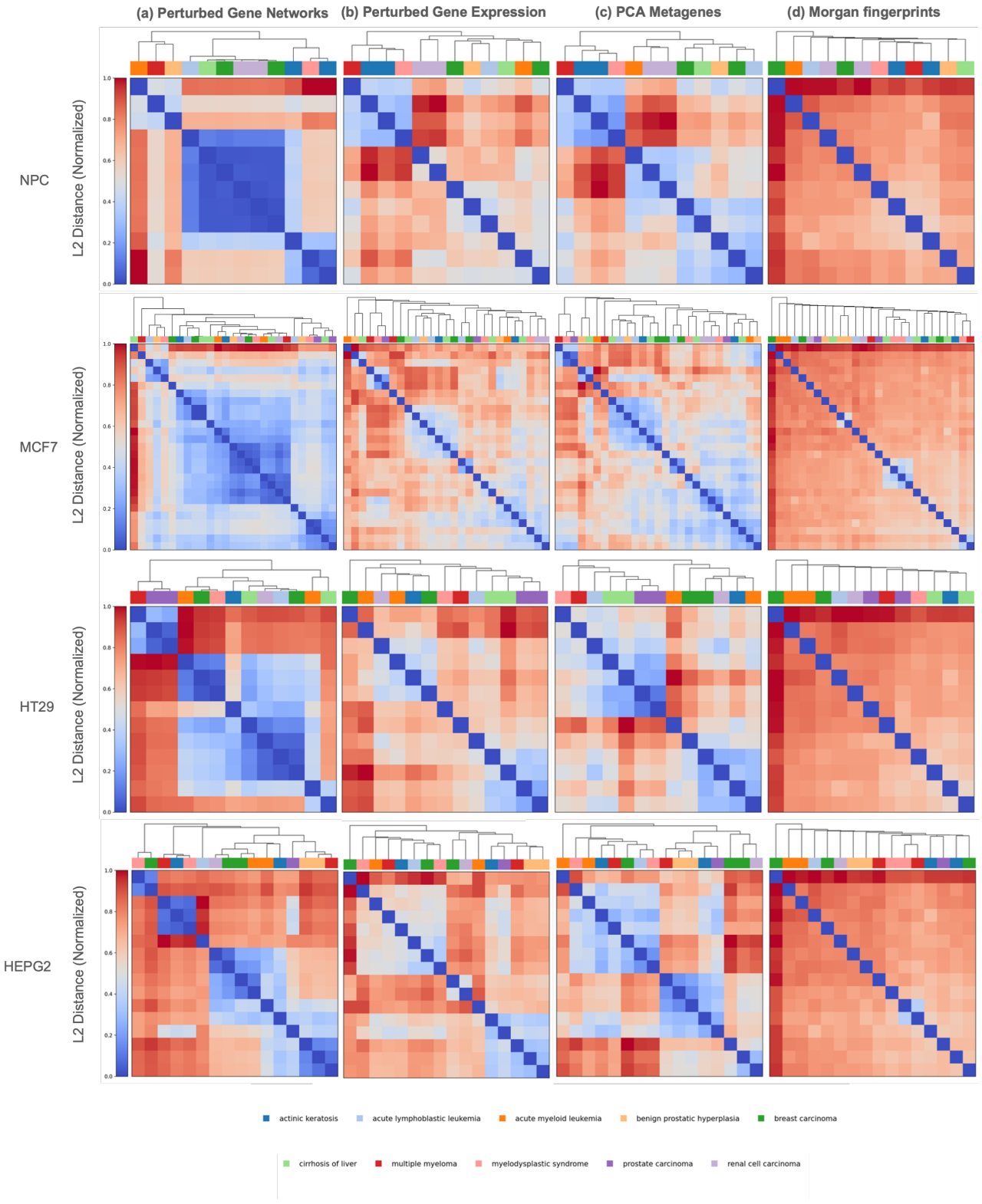

Figure 4. Organization of drugs based on four representations across cell types.

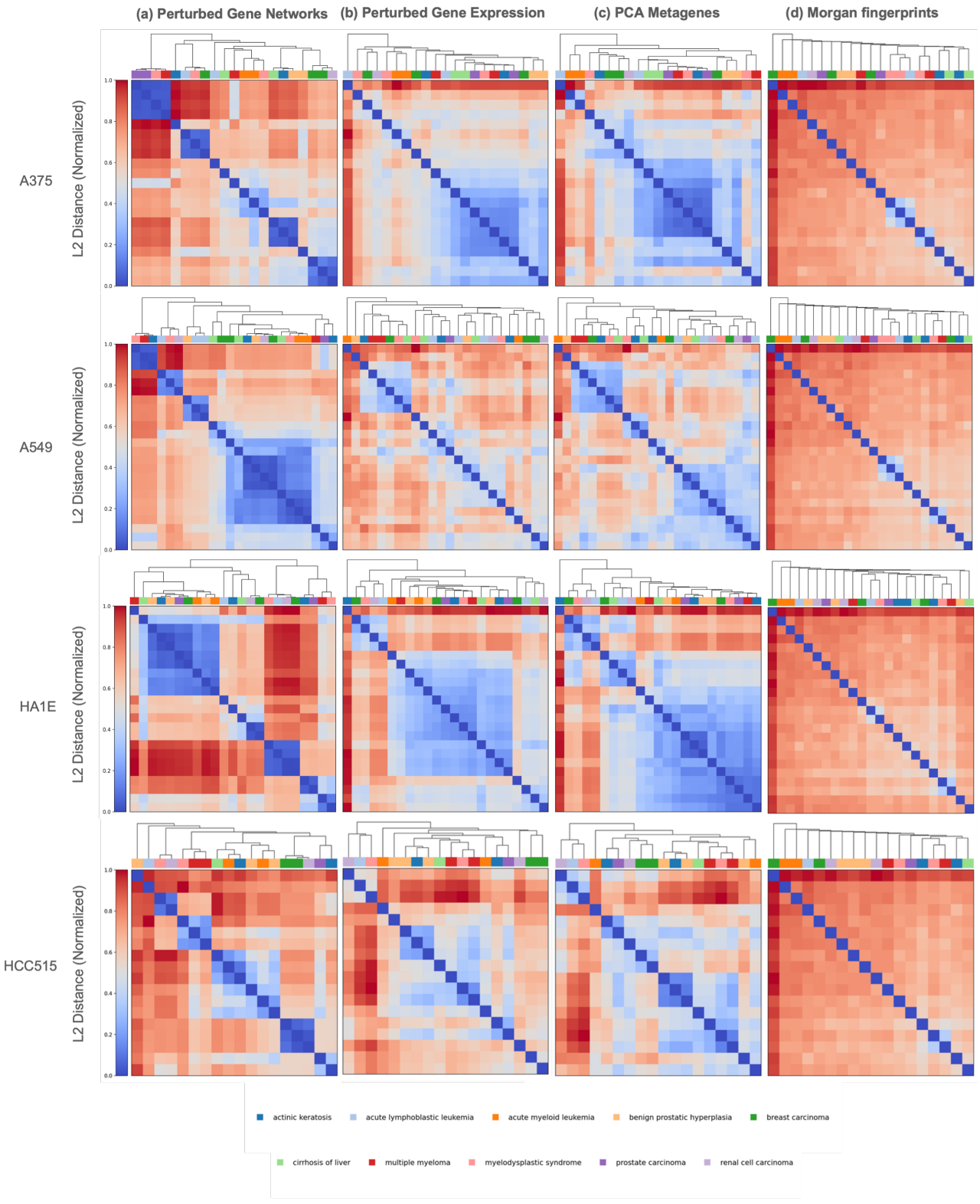

Figure 5. Organization of drugs based on four representations across cell types.