Virtual Screening on Cellular Systems

Caleb N. Ellington[*]
Carnegie Mellon University
cellingt@cs.cmu.edu

Sohan Addagudi[*]
Carnegie Mellon University
saddagud@andrew.cmu.edu

Jiaqi Wang[*]
University of Washington
jiaqiw18@uw.edu

Benjamin Lengerich
University of Wisconsin, Madison
lengerich@wisc.edu

Eric P. Xing
CMU, MBZUAI, GenBio AI
epxing@cs.cmu.edu

Abstract

Virtual screening methods prioritize therapeutic candidates by predicting drug properties and molecular interactions, but often fail to anticipate the complex, system-level responses that underlie most clinical trial failures. Virtual cells have been posed as a solution by predicting gene expression responses to drugs, but they remain weakly validated as screening tools, as expression only provides an intermediate to reconstruct the pathways and mechanisms underpinning drug success. To address this, we propose **CellVS-Net**, a model that maps multivariate context (cell type, drug, dose) to predict post-perturbation gene network restructuring. CellVS-Net accurately predicts gene network restructuring for unseen perturbations covering small molecules, large molecules, genetic knockdowns, and over-expressions, providing a new way to compare therapeutic effects from different perturbation modalities. To evaluate CellVS-Net's utility and move virtual screening toward understanding drug efficacy and mechanism of action we introduce two benchmarks for preclinical screening. First, we introduce **DR-Bench** (Disease Retrieval), which evaluates a method's ability to predict effective diseases for drugs with unseen target profiles. Second, we introduce **DTR-Bench** (Drug-Target Retrieval), which tests bidirectional drug-gene lookup across small-molecule and genetic perturbations, to enable deeper investigation into mechanism of action and off-target effects. Compared to both molecular and expression representations of drugs, CellVS-Net improves efficacy prediction in DR-Bench and mechanism prediction in DTR-Bench. Overall, CellVS-Net is posed to address a critical blind spot in drug development, and associated benchmarks enable hill-climbing on clinically relevant tasks. We provide source code for models and data curation, as well as Kaggle-style submission functions and running public leaderboards at https://github.com/SohanAddagudi/contextpert/tree/main.

Introduction

90% of clinical drug development fails [1]. 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 virtual screening, cell screening, and animal studies. Virtual docking and crystallography quickly rule out drugs that do not bind their target, while cell screens rule out crude toxicity. At later stages, efficacy and toxicity failures are limited to unforeseen system-level aspects of human biology, i.e. target mispecification and off-target effects. Despite improvements in virtual screening for molecular interactions, recently achieving proteome-scale screens [2], virtual screening approaches still remain blind to emergent failures at the level of cellular systems.

[*]These authors contributed equally

Preprint.

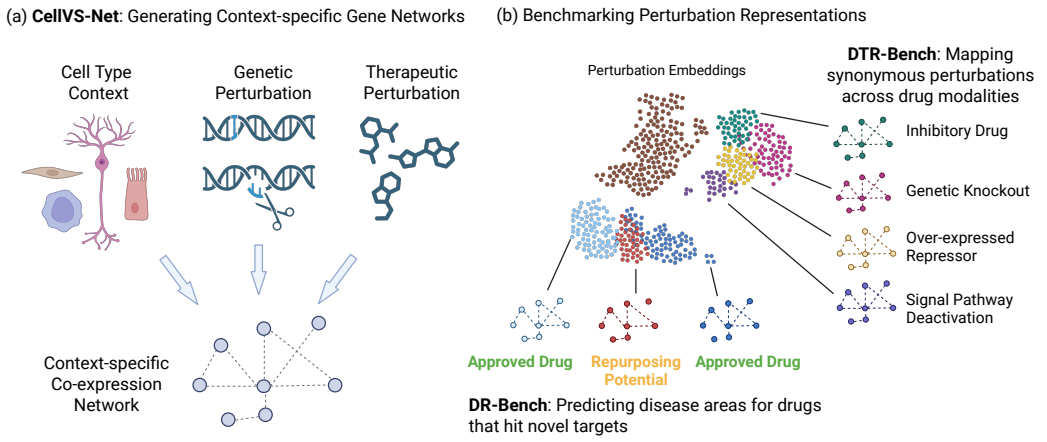

Figure 1: (a) CellVS-Net maps multivariate context (cell type, drug, dose) to context-specific gene co-expression networks. (b) We introduce two new benchmarks for evaluating drug representation approaches on clinically-relevant tasks. DR-Bench predicts effective diseases for drugs with previously unseen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

Moving beyond molecular interactions, large-scale expression and morphology profiling efforts such as LINCS L1000 [3], Tahoe-100M [4], scPerturb [5], JUMP-CP [6], and Recursion's phenomics platform [7] 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 [8, 9]. 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) [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. 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. [21] 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 relevant to drug discovery and (ii) generalize arbitrarily to unseen therapeutics to promote virtual exploration.

To address this, we curate two benchmarks, DR-Bench and DTR-Bench, by combining the Open-Targets and LINCS databases. We use these benchmarks to evaluate virtual screening methods for predicting disease indication and drug mechanism of action, enabling hill climbing on clinically-relevant applications. To move beyond methods for predicting expression snapshots, we also propose gene regulatory networks (GRNs) as a natural framework for modeling how perturbations rewire cellular circuitry. However, most GRN inference methods rely on partitioned cohorts [22, 23], which fail to capture the continuous and context-dependent rewiring observed in many diseases [24, 25], and plug-in estimators cannot generalize to new conditions. To address this, we propose CellVS-Net, a model which maps multivariate cellular and therapeutic contexts (cell type, drug, dose) to predict a post-perturbation gene co-expression network. 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. The correlation loss which enables gene network prediction is a drop-in replacement for models which currently optimize for expression using mean squared error losses [26]. We test CellVS-Net by generating co-expression networks on-demand for

unseen cell lines and perturbations including small molecules, large molecules (biologics), genetic knockdowns, and over-expressions. These networks provide a structured latent space for comparing therapeutic effects across different conditions and even different model runs. Finally, context-adaptive network inference provides a general method to integrate a growing amount of multimodal and multi-omic biomedical data into cohesive models of therapeutic effects.

Results

Estimating Perturbation-specific Gene Networks

We first consider the problem of estimating perturbation-specific gene co-expression 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 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.

	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: 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.

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 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-

Model Variant	Mean Squared Error
Population	0.9721
Group-specific	2.12e6
CellVS-Net Onehot	0.572
CellVS-Net + dose, time, cell line expression	**0.541**

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.

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 [32, 26]. 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 correlation network 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 [31]. For a full overview of context representations, see Methods.

Model Type	Context Encoder	Chemical	shRNA	Over Expression	Ligand
Population	None	1.0594	0.9740	0.7769	0.9315
CellVS-Net Molecule	Morgan Fingerprint [27]	0.5787	—	—	—
	ChemBERTa [28]	0.5728	—	—	—
CellVS-Net Target	PCA	0.5824	0.6833	0.7389	0.6077
	AIDO.Cell [12]	0.5687	**0.6815**	0.7681	0.6016
	AIDO.DNA [29]	0.5757	0.6831	0.7880	0.6012
	AIDO.Protein [30]	0.5432	0.6819	0.6767	**0.5772**
	AIDO.Structure [31]	**0.5389**	0.6817	**0.6678**	0.5774

Table 3: Correlation 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.

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 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).

	Disease Hits			
	@1	@5	@10	@25
CellVS-Net	**0.1071**	0.3036	**0.5536**	**0.8571**
AIDO.Cell [12]	0.0357	**0.3214**	0.5179	0.8393
Expression	**0.1071**	0.2500	0.4286	0.8214
PCA expression	0.0714	0.2143	0.3929	0.8036
Fingerprint [33]	0.0357	0.1786	0.3571	0.6071
Random	0.0357	0.1071	0.2500	0.6429

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.

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 2.

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

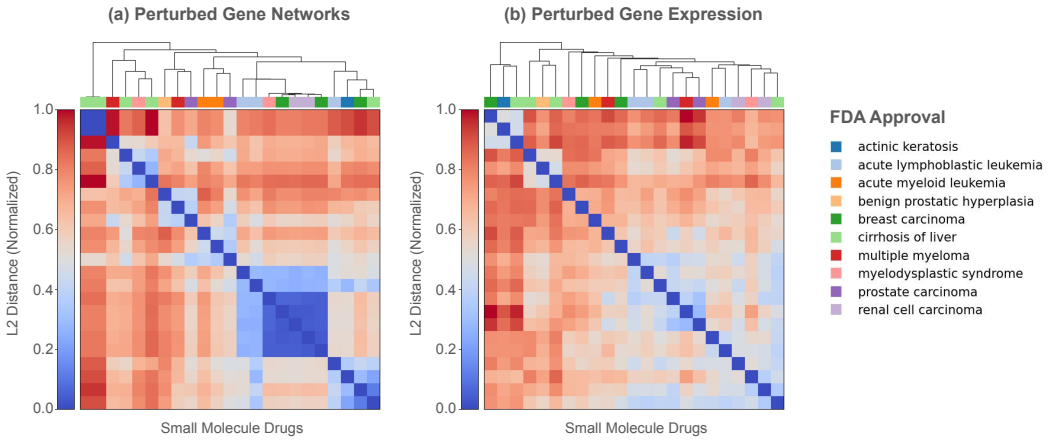

Figure 2: Organization of drugs based on (a) perturbed gene networks and (b) perturbed gene expression representations. Drugs are annotated with their FDA-approved disease indications. All samples are taken from the PC3 cell type.

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 approaches should induce a reliable similarity among perturbations with similar cell-level effects, even if they come from entirely 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. We validate this cross-modal repurposing 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). CellVS-Net consistently outperforms baselines on graph reconstruction and drug→target lookup. For target→drug lookup, PCA expression performs best.

			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 [12]	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

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. 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.

Lookup between inhibitory small molecules and genetic knockdowns provides a way to understand molecular mechanisms based on similar cell-level effects. To provide easy access to CellVS-Net and explore results, we provide a web tool for exploring mappings between shRNA and chemical perturbations (Figure 3).

Discussion

Current gaps in drug screening approaches have led to the dominant failures of clinical trials today: lack of efficacy due to target misspecification, and late toxicity due to off-target effects. Virtual screening on cellular systems offers a path to address these failures by predicting higher-order therapeutic effects beyond isolated molecular interactions. To define virtual screening on cellular systems we formulate DR-Bench and DTR-Bench, which evaluate perturbation 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). To address the gap in current methods for virtual screening on cellular systems, we introduce CellVS-Net, a contextualized network generator that maps multivariate therapeutic contexts to perturbation-specific coexpression networks (Figure 1). This approach has desirable statistical properties for long-tail drug screening applications (Table 1), and generalizes smoothly to unseen drugs, doses, and cell types, while improving with richer context representations, reducing MSE by 40-50% (Tables 2, 3). On DR-Bench and DTR-Bench, CellVS-Net

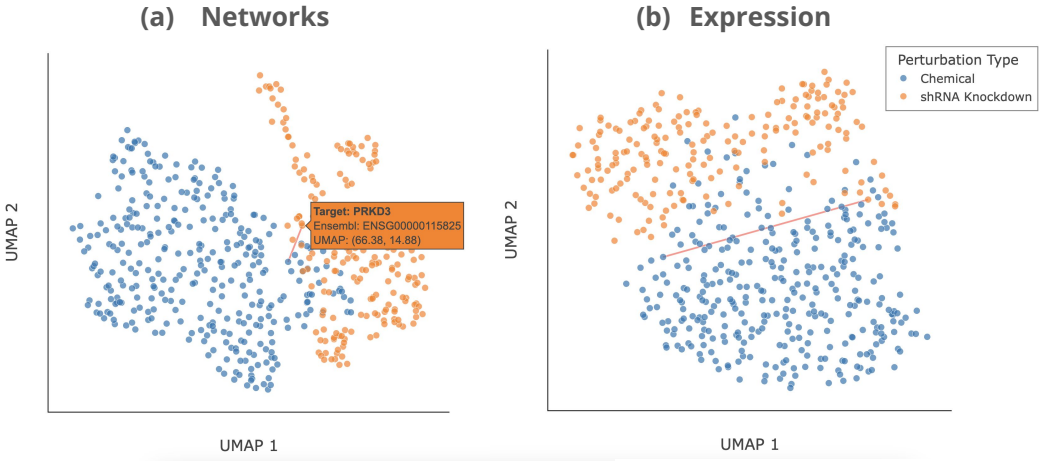

Figure 3: UMAP projection of (a) CellVS-Net networks and (b) expression snapshots for chemical and shRNA perturbations. We provide an interactive web tool on GitHub to explore this embedding space and highlight known interactions for a given drug or target. UMAP visualization provides a heuristic for 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).

improves over molecular and expression baselines in most cases, but leaves room for improvement on target→drug lookup.

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. By relating these representations to disease indication in DR-Bench, we provide a screen for efficacy which relates new drugs to known ones in a "samples-like-me" approach. DTR-Bench supports mechanism investigation by mapping drugs with unknown mechanisms to well-characterized ones, even across across drug classes. These benchmarks accommodate any vector representation of a perturbation, allowing unified comparison of molecular and cellular representations, and continuous evaluation as new methods and data types become available. To support this trajectory and enable hill-climbing over time, we maintain a public leaderboard on GitHub.

Methods

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.

DR-Bench First, we apply the combined OpenTargets-LINCS dataset to create the Disease-Retrieval Benchmark (DR-Bench). DR-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.

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 6: DR-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.

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 DR-Bench dataset (Table 6). 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 [2]. 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.

DTR-Bench 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).

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.

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 DR-Bench while specifically evaluating cross-modal mechanism-of-action prediction. We demonstrate retrieval as a search utility for drug-target mapping in Figure 3.

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 DR-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 [33]. 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.

CellVS-Net While these baseline representations capture aspects of chemical structure or transcriptional snapshots, they lack an explicit model of how a perturbation restructures the cellular system itself. 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.

CellVS-Net is designed to meet this need. It is a context-aware gene network model that maps a multivariate cellular and perturbation context to a predicted post-perturbation gene coexpression network. 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. This approach extends prior work on contextualized correlation networks [26], treating each perturbation as a request for an on-demand, condition-specific network governing cellular behavior.

CellVS-Net is an application of Contextualized modeling, a family of multi-task learning methods for learning models, distributions, and functions with context-specific parameters [32, 34]. Formally, CellVS-Net learns a context encoder $f(C)$ that maps a context vector C that may include cell type, perturbation, dose, and time to network parameters $\theta = f(C)$ that define a distribution on observed expression $X \sim P(X \mid \theta)$. Following [26], we use a Gaussian graphical model interpretation of coexpression, exploiting the univariate regression view of Pearson correlation:

$$\rho_{ij}^2 = \frac{\sigma_{ij}^2}{\sigma_{ii}^2 \sigma_{jj}^2} = \theta_{ij}\theta_{ji}.$$

This reduces correlation estimation to two symmetric one-dimensional regressions for each gene pair i and j, producing a differentiable correlation loss that is amenable to gradient-based optimization.

$$\widehat{f} = \operatorname*{argmin}_f \sum_{i,j} \|X_i - f(C)X_j\|_2$$

$$\mathbb{E}[\widehat{\rho}_{ij}^2 \mid C] = \widehat{f}(C)_{ij}\widehat{f}(C)_{ji}.$$

The encoder learns how network parameters vary smoothly with context, allowing CellVS-Net to interpolate to completely unseen perturbations or cell types, something that cluster-based or plug-in network estimators cannot achieve. It is noteworthy that this loss can serve as a drop-in replacement for regression MSE loss.

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 [34]. We test several methods for representing perturbations to improve generalization to unseen conditions, described in the Appendix.

References

[1] Duxin Sun, Wei Gao, Hongxiang Hu, and Simon Zhou. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica. B*, 12(7):3049–3062, July 2022.

[2] Andrew T. McNutt, Abhinav K. Adduri, Caleb N. Ellington, Monica T. Dayao, Eric P. Xing, Hosein Mohimani, and David R. Koes. Scaling Structure Aware Virtual Screening to Billions of Molecules with SPRINT, January 2025. arXiv:2411.15418 [q-bio].

[3] Amar Koleti, Raymond Terryn, Vasileios Stathias, Caty Chung, Daniel J Cooper, John P Turner, Dušica Vidović, Michele Forlin, Tanya T Kelley, Alessandro D'Urso, Bryce K Allen, Denis Torre, Kathleen M Jagodnik, Lily Wang, Sherry L Jenkins, Christopher Mader, Wen Niu, Mehdi Fazel, Naim Mahi, Marcin Pilarczyk, Nicholas Clark, Behrouz Shamsaei, Jarek Meller, Juozas Vasiliauskas, John Reichard, Mario Medvedovic, Avi Ma'ayan, Ajay Pillai, and Stephan C Schürer. Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic Acids Research*, 46(D1):D558–D566, January 2018.

[4] Jesse Zhang, Airol A. Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G. Jones, Vuong Tran, Joseph Pangallo, Efthymia Papalexi, Ajay Sapre, Hoai Nguyen, Oliver Sanderson, Maria Nigos, Olivia Kaplan, Sarah Schroeder, Bryan Hariadi, Simone Marrujo, Crina Curca, Alec Salvino, Guillermo Gallareta Olivares, Ryan Koehler, Gary Geiss, Alexander Rosenberg, Charles Roco, Daniele Merico, Nima Alidoust, Hani Goodarzi, and Johnny. Yu. Tahoe-100m: A giga-scale single-cell

perturbation atlas for context-dependent gene function and cellular modeling, February 2025. bioRxiv.

[5] Stefan Peidli, Tessa D. Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, Augustin Luna, Nils Blüthgen, and Chris Sander. scPerturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, March 2024. Publisher: Nature Publishing Group.

[6] Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D. Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D. Boyd, Laurent Brino, Patrick J. Byrne, Hugo Ceulemans, Carolyn Ch'ng, Beth A. Cimini, Djork-Arne Clevert, Nicole Deflaux, John G. Doench, Thierry Dorval, Regis Doyonnas, Vincenza Dragone, Ola Engkvist, Patrick W. Faloon, Briana Fritchman, Florian Fuchs, Sakshi Garg, Tamara J. Gilbert, David Glazer, David Gnutt, Amy Goodale, Jeremy Grignard, Judith Guenther, Yu Han, Zahra Hanifehlou, Santosh Hariharan, Desiree Hernandez, Shane R. Horman, Gisela Hormel, Michael Huntley, Ilknur Icke, Makiyo Iida, Christina B. Jacob, Steffen Jaensch, Jawahar Khetan, Maria Kost-Alimova, Tomasz Krawiec, Daniel Kuhn, Charles-Hugues Lardeau, Amanda Lembke, Francis Lin, Kevin D. Little, Kenneth R. Lofstrom, Sofia Lotfi, David J. Logan, Yi Luo, Franck Madoux, Paula A. Marin Zapata, Brittany A. Marion, Glynn Martin, Nicola Jane McCarthy, Lewis Mervin, Lisa Miller, Haseeb Mohamed, Tiziana Monteverde, Elizabeth Mouchet, Barbara Nicke, Arnaud Ogier, Anne-Laure Ong, Marc Osterland, Magdalena Otrocka, Pieter J. Peeters, James Pilling, Stefan Prechtl, Chen Qian, Krzysztof Rataj, David E. Root, Sylvie K. Sakata, Simon Scrace, Hajime Shimizu, David Simon, Peter Sommer, Craig Spruiell, Iffat Sumia, Susanne E. Swalley, Hiroki Terauchi, Amandine Thibaudeau, Amy Unruh, Jelle Van de Waeter, Michiel Van Dyck, Carlo van Staden, Michał Warchoł, Erin Weisbart, Amélie Weiss, Nicolas Wiest-Daessle, Guy Williams, Shan Yu, Bolek Zapiec, Marek Żyła, Shantanu Singh, and Anne E. Carpenter. JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations, March 2023. Pages: 2023.03.23.534023 Section: New Results.

[7] Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T. Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M. Gustafsdottir, Christopher C. Gibson, and Anne E. Carpenter. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774, September 2016.

[8] Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B. Burkhardt, Andrea Califano, Jonah Cool, Abby F. Dernburg, Kirsty Ewing, Emily B. Fox, Matthias Haury, Amy E. Herr, Eric Horvitz, Patrick D. Hsu, Viren Jain, Gregory R. Johnson, Thomas Kalil, David R. Kelley, Shana O. Kelley, Anna Kreshuk, Tim Mitchison, Stephani Otte, Jay Shendure, Nicholas J. Sofroniew, Fabian Theis, Christina V. Theodoris, Srigokul Upadhyayula, Marc Valer, Bo Wang, Eric Xing, Serena Yeung-Levy, Marinka Zitnik, Theofanis Karaletsos, Aviv Regev, Emma Lundberg, Jure Leskovec, and Stephen R. Quake. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, December 2024. Publisher: Elsevier.

[9] Le Song, Eran Segal, and Eric Xing. Toward AI-Driven Digital Organism: Multiscale Foundation Models for Predicting, Simulating and Programming Biology at All Levels, December 2024. arXiv:2412.06993 [cs].

[10] Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, June 2023.

[11] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo. Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21:1470–1480, February 2024.

[12] Nicholas Ho, Caleb N. Ellington, Jinyu Hou, Sohan Addagudi, Shentong Mo, Tianhua Tao, Dian Li, Yonghao Zhuang, Hongyi Wang, Xingyi Cheng, Le Song, and Eric P. Xing. Scaling dense representations for single cell with transcriptome-scale context, November 2024. bioRxiv.

[13] Yusuf Roohani, Kexin Huang, and Jure Leskovec. GEARS: Predicting transcriptional outcomes of novel multi-gene perturbations, July 2022. Pages: 2022.07.12.499735 Section: New Results.

[14] Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with state, June 2025. bioRxiv.

[15] Hengshi Yu, Weizhou Qian, Yuxuan Song, and Joshua D. Welch. Perturbnet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular Systems Biology*, 2025.

[16] Siyu He, Yuefei Zhu, Daniel Naveed Tavakol, Haotian Ye, Yeh-Hsing Lao, Zixian Zhu, Cong Xu, Shradha Chauhan, Guy Garty, Raju Tomer, Gordana Vunjak-Novakovic, James Zou, Elham Azizi, and Kam W. Leong. Squidiff: predicting cellular development and responses to perturbations using a diffusion model. *Nature Methods*, 2025.

[17] Yuge Ji, Alejandro Tejada-Lapuerta, Niklas A. Schmacke, Zihe Zheng, Xinyue Zhang, Simrah Khan, Ina Rothenaigner, Juliane Tschuck, Kamyar Hadian, Veit Hornung, and Fabian J. Theis. Scalable and universal prediction of cellular phenotypes enables in silico experiments, September 2025. ISSN: 2692-8205 Pages: 2024.08.12.607533 Section: New Results.

[18] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, Jay Shendure, Jose L McFaline-Figueroa, Pierre Boyeau, F Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, June 2023. Publisher: John Wiley & Sons, Ltd.

[19] Philip Fradkin, Puria Azadi, Karush Suri, Frederik Wenkel, Ali Bashashati, Maciej Sypetkowski, and Dominique Beaini. How Molecules Impact Cells: Unlocking Contrastive PhenoMolecular Retrieval, September 2024. arXiv:2409.08302 [q-bio].

[20] Ding Bai, Caleb N Ellington, Shentong Mo, Le Song, and Eric P Xing. AttentionPert: accurately modeling multiplexed genetic perturbations with multi-scale effects. *Bioinformatics*, 40(Supplement_1):i453–i461, July 2024.

[21] Djordje Miladinovic, Tobias Höppe, Mathieu Chevalley, Andreas Georgiou, Lachlan Stuart, Arash Mehrjou, Marcus Bantscheff, Bernhard Schölkopf, and Patrick. Schwab. In silico biological discovery with large perturbation models. *Nature Computational Science*, October 2025.

[22] Pau Badia-i Mompel, Lorna Wessels, Sophia Müller-Dott, Rémi Trimbour, Ricardo O. Ramirez Flores, Ricard Argelaguet, and Julio Saez-Rodriguez. Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews Genetics*, pages 1–16, June 2023. Publisher: Nature Publishing Group.

[23] Matthew Stone, Sunnie Grace McCalla, Alireza Fotuhi Siahpirani, Viswesh Periyasamy, Junha Shin, and Sushmita Roy. Identifying strengths and weaknesses of methods for computational network inference from single cell RNA-seq data. Publication Title: bioRxiv, June 2021.

[24] Oana Ursu, James T. Neal, Emily Shea, Pratiksha I. Thakore, Livnat Jerby-Arnon, Lan Nguyen, Danielle Dionne, Celeste Diaz, Julia Bauman, Mariam Mounir Mosaad, Christian Fagre, April Lo, Maria McSharry, Andrew O. Giacomelli, Seav Huong Ly, Orit Rozenblatt-Rosen, William C. Hahn, Andrew J. Aguirre, Alice H. Berger, Aviv Regev, and Jesse S. Boehm. Massively parallel phenotyping of coding variants in cancer with Perturb-seq. *Nature Biotechnology*, 40(6):896–905, June 2022. Number: 6 Publisher: Nature Publishing Group.

[25] Katherine A. Hoadley, Christina Yau, Toshinori Hinoue, Denise M. Wolf, Alexander J. Lazar, Esther Drill, Ronglai Shen, Alison M. Taylor, Andrew D. Cherniack, Vésteinn Thorsson, Rehan

Akbani, Reanne Bowlby, Christopher K. Wong, Maciej Wiznerowicz, Francisco Sanchez-Vega, A. Gordon Robertson, Barbara G. Schneider, Michael S. Lawrence, Houtan Noushmehr, Tathiane M. Malta, Joshua M. Stuart, Christopher C. Benz, and Peter W. Laird. 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.

[26] Caleb N. Ellington, Benjamin J. Lengerich, Thomas B. K. Watkins, Jiekun Yang, Abhinav K Adduri, Sazan Mahbub, Hanxi Xiao, Manolis Kellis, and Eric P. Xing. Learning to estimate sample-specific transcriptional networks for 7,000 tumors. *Proceedings of the National Academy of Sciences*, 122(21):e2411930122, May 2025. Publisher: Proceedings of the National Academy of Sciences.

[27] Suna Onengut-Gumuscu, Wei-Min Chen, Oliver Burren, Nick J. Cooper, Aaron R. Quinlan, Josyf C. Mychaleckyj, Emily Farber, Jessica K. Bonnie, Michal Szpak, Ellen Schofield, Premanand Achuthan, Hui Guo, Mary D. Fortune, Helen Stevens, Neil M. Walker, Lucas D. Ward, Anshul Kundaje, Manolis Kellis, Mark J. Daly, Jeffrey C. Barrett, Jason D. Cooper, Panos Deloukas, John A. Todd, Chris Wallace, Patrick Concannon, and Stephen S. Rich. 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. Number: 4 Publisher: Nature Publishing Group.

[28] Riya Singh, Aryan Amit Barsainyan, Rida Irfan, Connor Joseph Amorin, Stewart He, Tony Davis, Arun Thiagarajan, Shiva Sankaran, Seyone Chithrananda, Walid Ahmad, Derek Jones, Kevin McLoughlin, Hyojin Kim, Anoushka Bhutani, Shreyas Vinaya Sathyanarayana, Venkat Viswanathan, Jonathan E. Allen, and Bharath Ramsundar. ChemBERTa-3: An Open Source Training Framework for Chemical Foundation Models. *ChemRxiv*, 2025.

[29] Caleb N. Ellington, Ning Sun, Nicholas Ho, Tianhua Tao, Sazan Mahbub, Dian Li, Yonghao Zhuang, Hongyi Wang, Le Song, and Eric P. Xing. Accurate and general dna representations emerge from genome foundation models at scale. *bioRxiv*, 2024.

[30] Ning Sun, Shuxian Zou, Tianhua Tao, Sazan Mahbub, Dian Li, Yonghao Zhuang, Hongyi Wang, Xingyi Cheng, Le Song, and Eric P. Xing. Mixture of experts enable efficient and effective protein understanding and design. *bioRxiv*, 2024.

[31] Jiayou Zhang, Barthelemey Meynard-Piganeau, James Gong, Xingyi Cheng, Yingtao Luo, Hugo Ly, Le Song, and Eric Xing. Balancing locality and reconstruction in protein structure tokenizer. *bioRxiv*, 2024.

[32] Benjamin Lengerich, Caleb N. Ellington, Andrea Rubbi, Manolis Kellis, and Eric P. Xing. Contextualized Machine Learning, October 2023. arXiv:2310.11340 [cs, stat].

[33] Alice Capecchi, Daniel Probst, and Jean-Louis Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12(1):43, June 2020.

[34] Caleb N. Ellington, Benjamin J. Lengerich, Wesley Lo, Aaron Alvarez, Andrea Rubbi, Manolis Kellis, and Eric P. Xing. Contextualized: Heterogeneous Modeling Toolbox. *Journal of Open Source Software*, 9(97):6469, May 2024.

[35] Euxhen Hasanaj, Elijah Cole, Shahin Mohammadi, Sohan Addagudi, Xingyi Zhang, Le Song, and Eric P. Xing. Multimodal benchmarking of foundation model representations for cellular perturbation response prediction. *bioRxiv*, 2025.

[36] Nicholas Ho, Caleb N. Ellington, Jinyu Hou, Sohan Addagudi, Shentong Mo, Tianhua Tao, Dian Li, Yonghao Zhuang, Hongyi Wang, Xingyi Cheng, Le Song, and Eric P. Xing. Scaling Dense Representations for Single Cell with Transcriptome-Scale Context, December 2024. Pages: 2024.11.28.625303 Section: New Results.

[37] Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science (New York, N.Y.)*, 365(6455):786–793, August 2019.

A 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 [35]. 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 [33], a substructure representation that encodes local atomic environments and has proven effective in traditional cheminformatics pipelines. Second, we apply ChemBERTa-100M-MLM [28], 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 [36], a full-transcriptome single-cell foundation model trained across diverse cellular contexts. Gene embeddings are computed using K562 control cells from Norman et al [37].

`AIDO.DNA (sequence-based).` We extract sequence-level gene representations using the AIDO.DNA model [29]. 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 [30], 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 [31], 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 [35], 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.

B 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 8: 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.

C Clustermap of Contexualized Network & Gene Expression & PCA Metagenes & Morgan fingerprints across cell types

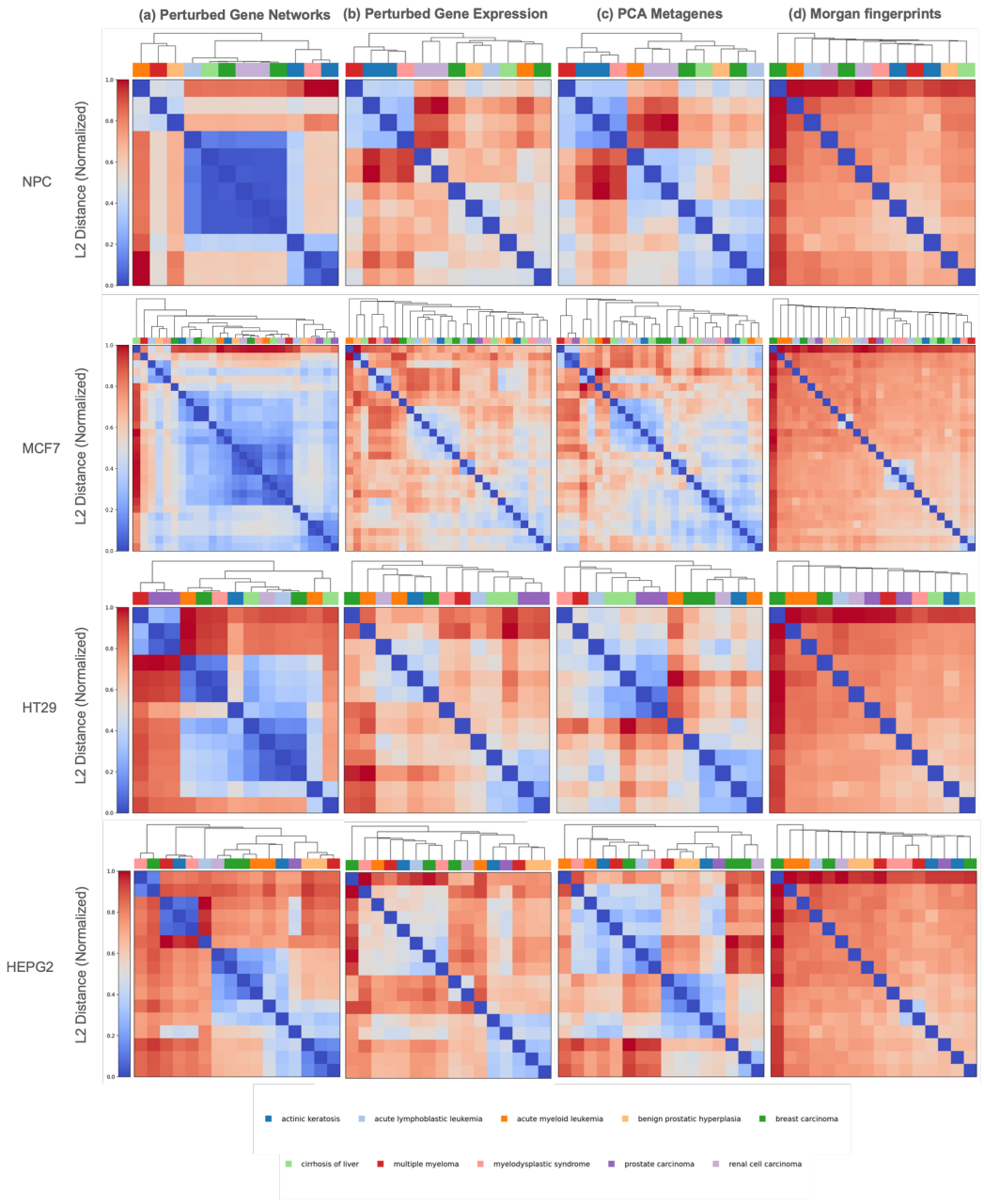

Figure 4: Organization of drugs based on four representations across cell types.

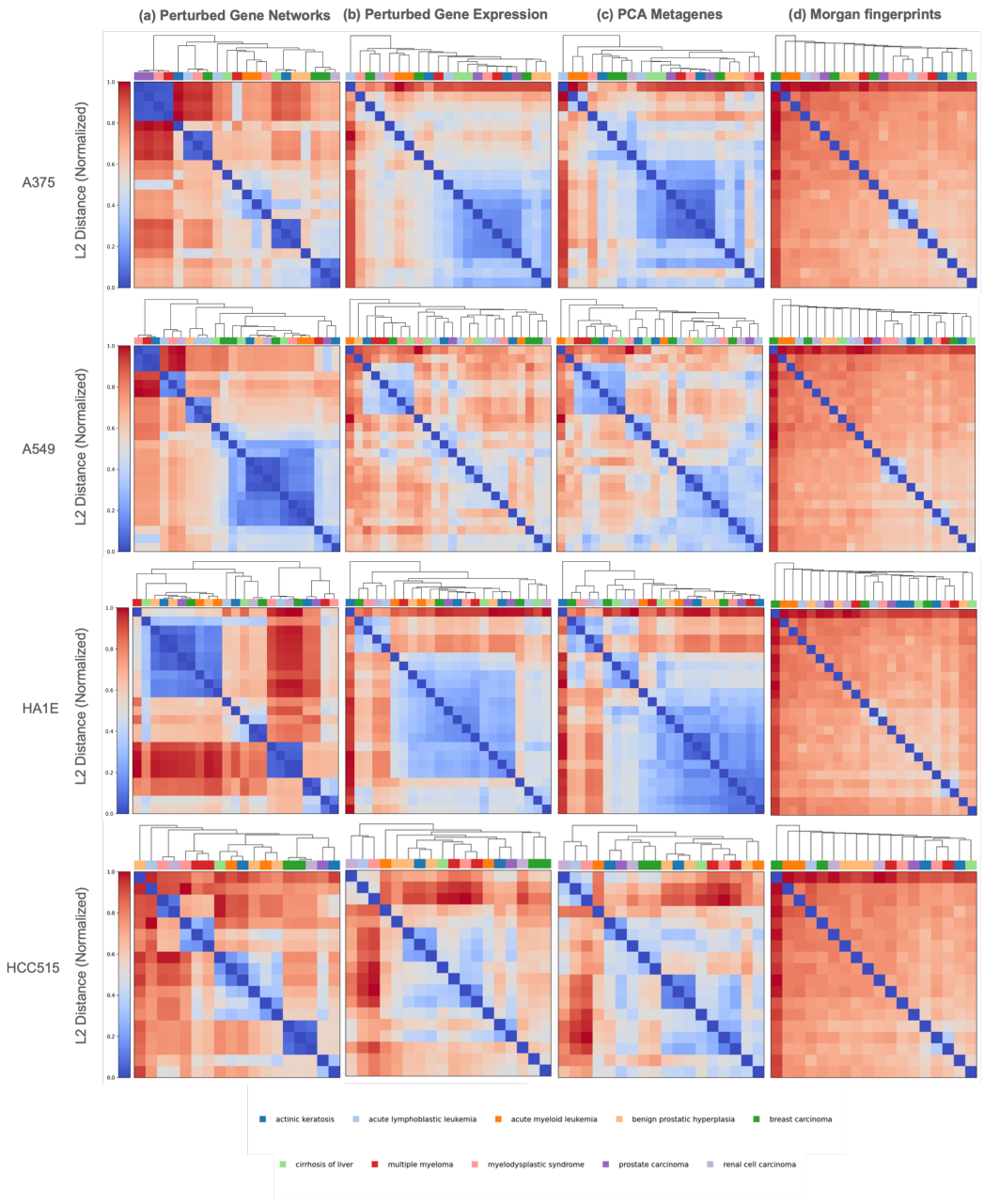

Figure 5: Organization of drugs based on four representations across cell types.