Virtual Screening on Cellular Systems

Anonymous Author(s)
Affiliation
Address
email

Abstract

Virtual screening has traditionally focused on molecular targets, often failing to anticipate the complex, system-level failures that arise during clinical trials. To address this, we propose (i) two new benchmarks for evaluating the clinical relevance of virtual screening methods, such as virtual cells, and (ii) a framework for virtual screening against entire cellular systems. Our framework uses contextualized modeling, a multi-task learning approach for inferring context-specific network models, to infer perturbation-specific coexpression networks from large-scale screening datasets, enabling accurate prediction of network restructuring under diverse cellular and therapeutic contexts. We demonstrate that context-adaptive models outperform even observed expression profiles for predicting disease and drug mechanisms, suggesting low-cost improvements to common virtual cell objectives. At test-time, contextualized networks generate accurate models of gene network reorganization on-demand for completely unseen cell types and therapies. Across multiple independent runs, networks provide a standard, cohesive, and constrained latent space to compare therapeutic effects from different perturbation modalities (knockout, overexpression, small molecule). Comparing perturbations in terms of cell-level effects leads to a principled approach to drug repurposing, safety profiling, and interpreting mechanism of action. Rethinking virtual cell benchmarks to target clinical relevance and drug repurposing opens a path to hill-climbing on prelinical screening.

Introduction

Over 80% of drug candidates that reach clinical trials fail due to incorrect targets or unforeseen system-level effects [1]. Despite improvements in virtual screening for molecular interactions, recently achieving proteome-scale screens [2], such approaches remain blind to emergent failures at the level of cellular systems. Transcriptional response predictors ("virtual cells") [3, 4, 5] provide a snapshot view of transcriptional measurements, but do not directly reveal safety, efficacy, or off-target effects.

To address this, we curate two benchmarks from the OpenTargets and LINCS databases to test virtual screening methods on representing disease indication and drug mechanism of action, and enable hill climbing on clinically-relevant applications. We focus on enabling therapeutic comparisons, matching new therapies to well-characterized ones based on similar predicted effects, which enables us to compare both molecular and cell-level virtual screening methods.

We also turn to gene regulatory networks (GRNs) as a natural framework for modeling and comparing cellular circuitry. Condition-specific GRNs capture cellular responses to various interventions or intrinsic conditions, and display redundancies or fragilities for potential therapeutic development. However, existing GRN inference methods rely on partitioned cohorts [6, 7], which fail to capture the continuous and context-dependent rewiring observed in many diseases [8, 9], and plug-in estimators cannot generalize to new conditions.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

To address this, we propose a virtual screening framework based on contextualized modeling, which infers perturbation-specific GRNs conditioned on multivariate cellular and therapeutic contexts, and generates network models on-demand for unseen conditions, enabling a novel virtual screening approach. Furthermore, 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 and foundation models into cohesive models of disease pathology. By leveraging large-scale perturbation data, we aim to enable principled screening for efficacy, safety, and mechanistic similarity, while also supporting applications in drug repurposing, cohort refinement, and biomarker discovery, with implications in drug development for rare and heterogeneous diseases.

Methods

Data We curated a benchmark combining the OpenTargets and LINCS databases to test and compare perturbation representations. We select diseases with at least 2 targets, each with FDA-approved drugs, and merge this with LINCS small molecules, providing a dataset to test similar effects from different targets. The resulting dataset contains 50 approved drugs from 55 OpenTargets diseases with an average of 2.03 drugs per disease. The LINCS database contains 68 cell lines exposed to these drugs, with an average of 8.60 drugs per cell line.

We also apply the LINCS L1000 dataset for benchmarking model accuracy on held-out perturbations and cell lines. This dataset includes quantile-normalized expression values as well as metadata for cell line, perturbation type, dose, and post-dose measurement time. The final dataset contains 76 cell lines total, 71 with perturbations applied, and 12,053 total small-molecule perturbations. We produce 2 train-test splits: a context-held-out split, where certain contexts are entirely unseen during training, and a sample-held-out split where all contexts are seen during training and samples within each context randomly held out. Expression is PCA-compressed to 50 metagenes and all features are normalized.

Contextualized Coexpression Networks Contextualization is based on two modular components: a context encoder which translates sample context into model parameters, and a sample-specific model which represents the latent context-specific mechanisms of data generation. This view conveniently unifies both varying-coefficient models [10], and subpopulation and partition-based approaches, such as cluster analysis and cohort analysis [11]. By learning how models change in response to context, contextualization enables powerful control over high-dimensional and continuously varying contexts, discovering dynamic latent structures underlying data generation in heterogeneous populations and permitting GRN model inference at even sample-specific resolution.

We seek a context-specific density of network parameters $\mathbb{P}(\theta \mid C)$ such that

$$\mathbb{P}(X \mid C) = \int_{\theta} d\theta \mathbb{P}_M(X \mid \theta)\mathbb{P}(\theta \mid C)$$

is maximized, where $\mathbb{P}(X \mid \theta)$ is the probability of gene expression $X \in \mathbb{R}^p$ under network model class M with parameters $\theta \in \mathbb{R}^{p \times p}$, and C is sample context which can contain both multivariate and real features defining the cell line and therapeutic perturbation. To create a virtual screening approach that infers gene networks under unseen therapies and for new cell lines, plug-in estimators are insufficient. We instead apply a context encoder $f(C)$, implemented as a deep neural network, using a point-mass Dirac delta distribution, which gives the familiar contextualized model form [12]

$$P(X \mid \theta = f(C))$$

In this study, we use a multivariate gaussian as model class M to parameterize $\mathbb{P}_M(X \mid \theta)$, and transform this into a differentiable and convex negative-log-likelihood function which allows us to infer contextualized correlation networks. Correlation networks are simple to estimate and often state-of-the-art for gene regulatory network inference [13]; contextualized correlation expands this utility to the granularity of sample-specific network inferences. To estimate sample-specific correlation networks, we assume the data was drawn from $X \sim N(0, \Sigma)$ and use the well-known univariable regression view of Pearson's marginal correlation coefficient:

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

where the covariance matrix Σ has elements σ_{ij}, and $\widehat{\theta}_{ij} = \text{argmin}_\theta (X_j - X_i\theta)^2$. This form converts correlation into two separable univariate least-squares regressions that maximize the marginal conditional probabilities $P(X_i|X_j)$ and $P(X_j|X_i)$, leading to the differentiable correlation loss objective from [14]. We learn this model by applying the Contextualized [11] package.

Results

Generating Coexpression Networks On-demand for Unseen Therapies Contextualized networks learn to map contexts to context-specific model parameters through a context encoder. Context representations which contain some signal about the similarity or difference in downstream distributions greatly improve accuracy and generalization, even in the presence of noise features and non-linear effects [14]. For cell line, we use control cell profile gene expression. We apply morgan fingerprints [15] to represent small molecule contexts. For all other perturbations, we represent the target gene using AIDO.Cell gene embeddings [16]. We evaluate models on their ability to generalize to unseen perturbations (Table 1). Group-specific modeling and one-hot contexts fail in this regime, as unseen contexts cannot be mapped onto the original groups or feature set. Population modeling still applies as a context-ignorant method. To generalize effectively, the context encoder must learn a model of how small molecules affect cellular systems.

Model Variant	Small Molecule	shRNA	Over Expression	Ligand
Population	0.9914	0.9776	0.7944	0.9178
Cell-type-specific	—	—	—	—
One-hot	—	—	—	—
Fingerprint	**0.5943**	**0.6804**	**0.7504**	**0.6861**

Table 1: MSE of inferred networks on a context-held-out split for perturbed expression measurements from 71 cell lines perturbed with one of 12,053 small molecules. Small molecule contexts are encoded as morgan fingerprints [15].

Model Variant	MSE ↓
Population	0.978
Cell-type-specific	1.84e5
Contextualized	0.631
+ dose, time	0.685
+ dose, time, cell line embedding	0.549
+ 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.

Rich Context Representations Improve Performance To evaluate the impact of richer context, we incrementally augment input features for the context encoder, moving toward continuous contexts that put an implicit prior on the similarity between condition-specific modeling tasks (Table 2). In this experiment, we begin to evaluate prediction of post-perturbation networks, using one-hot encoded small molecule perturbations along with various representations of the cell type context. Post-perturbation prediction is more challenging than the control-only setup in Table 5. Contextualized networks again avoid over and underfitting, producing much-improved test performance even with one-hot contexts. Substituting the one-hot encoding of cell type for an embedding of the unperturbed transcriptomic profile from AIDO.Cell foundation model [16] improves generalization considerably for predicting post-perturbation networks. Providing the entire control expression profile as context improves further.

Networks Similarity Links Drugs with Disparate Targets to Shared Therapeutic Mechanisms A core requirement for cell-level virtual screening is that the representation space induces reliable similarity among perturbations with similar cell-level effects. To evaluate this, we gather a dataset of small molecule drugs from the OpenTargets platform that have different molecular targets, 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 compare therapeutic similarity in two spaces: (i) observed gene expression (an "oracle" virtual cell) and (ii) predicted cellular network representation. We label drugs by their disease indications and evaluate the representations in terms of silhouette score and pairwise AUROC to determine cohesion among indications, as well as a k-nearest neighbors model at $k \in [1, 5, 10, 50]$ to determine the representation's ability to predict new indications. We find that contextualized networks are consistently more representative of cell-level therapeutic effects than observed expression (Table 3).

Metric	Contextualized networks		Gene expression		PCA metagenes	
	micro	macro	micro	macro	micro	macro
Silhouette ↑	-0.1135	-0.1100	-0.0605	-0.0592	-0.0784	-0.0731
Pairwise AUROC ↑	0.5547		0.5334		0.5182	
kNN@1 ↑	0.4553	0.4697	0.4399	0.4209	0.4439	0.4201
kNN@5 ↑	0.4495	0.4480	0.3907	0.3310	0.3915	0.3294
kNN@10 ↑	0.4340	0.4044	0.3457	0.2759	0.3448	0.2712
kNN@50 ↑	0.3485	0.2592	0.2277	0.1603	0.2285	0.1604

Table 3: Systems-level similarity: disease-structure metrics in the context-inferred network representation vs. raw gene expression vs. PCA metagenes. "micro" weights classes by frequency; "macro" averages per-class scores.

Network Similarity Enables Drug Repurposing Across Modalities via Cell-level Effects Contextualized networks provide a constrained latent space across model runs, enabling lookup across modalities. Lookup between small molecules, knockdowns and knockouts, and over expressions provides a way to understand molecular mechanisms based on similar cell-level effects. 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 4).

	Contextualized networks	Gene expression	PCA metagenes
KNN @ 1 ↑	0.4322	0.4218	0.4162
KNN @ 5 ↑	0.4139	0.3835	0.3837
KNN @ 10 ↑	0.3856	0.3568	0.3533
KNN @ 50 ↑	0.2944	0.2862	0.2842

Table 4: Mapping small molecule drugs to genetic perturbations with identical targets using various representations of cell-level effects.

Discussion

Contextualized gene networks are a useful latent representation of cellular systems, allowing us to compare therapeutic, cell line, and pathological conditions under a cohesive and interpretable representation. Previous work has applied the similarity of these networks to identify prognostic disease types and relate disjoint biomarkers and patient cohorts through shared mechanisms [14]. Future work will focus on using network similarity to identify therapeutics which lead to similar disease states. By relating drugs to one another through predicted systems-level responses, we aim to enable a virtual drug discovery and repurposing platform which lends itself to predicting phenotypic response, while simultaneously predicting targets and mechanisms via relationships to well-characterized genetic and chemical perturbations.

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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Trevor Hastie and Robert Tibshirani. Varying-Coefficient Models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1993.tb01939.x.

[11] 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.

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

[13] Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, February 2020.

[14] 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.

[15] 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.

[16] 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.

A Contextualized Modeling

Context Encoding Addresses Over and Under Fitting

We first assess failure modes by comparing a contextualized network estimator to population and group-specific networks in a minimal regime. Contexts are one-hot encoded, containing no prior knowledge of cell line similarity, intending to disadvantage contextualized networks which leverage the similarity of contexts to share information and extrapolate between modeling tasks. In this experiment, perturbations are ignored and we only test across control measurements for each cell line. Notably, contextualized networks achieve the best performance on the full dataset by mitigating failure models of the population and condition-specific baselines (Table 5). Population models come with 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 ($n_c \leq 3$). Contextualized networks automatically switch between a population default when there are insufficient samples and group-specific modeling when there are sufficient samples, achieving stable performance across data regimes.

Additionally, continuous features, such as dose and time, are difficult to use with discrete group-based models, but contextualization naturally integrates these and further improves generalization.

	Full Data	$n_c > 3$	$n_c \leq 3$
Zeros	0.998	0.998	0.694
Population	0.978	0.980	**0.681**
Cell line-specific	51.576	**0.662**	1.38e6
Contextualized	**0.669**	0.665	0.730
+ dose, time	0.6433	0.638	0.767

Table 5: Mean-squared error (MSE) of inferred transcriptional networks on a sample-held-out split for control measurements from all cell lines. Contextualized and cell line-specific models use one-hot encoded celltype contexts. $n_c > 3$ subsets the conditions with more than 3 observations, while $n_c \leq 3$ subsets the conditions with less than 3 observations. Full data is the union of both.

B Hierarchical clustermap of Gene Expression & PCA Metagenes

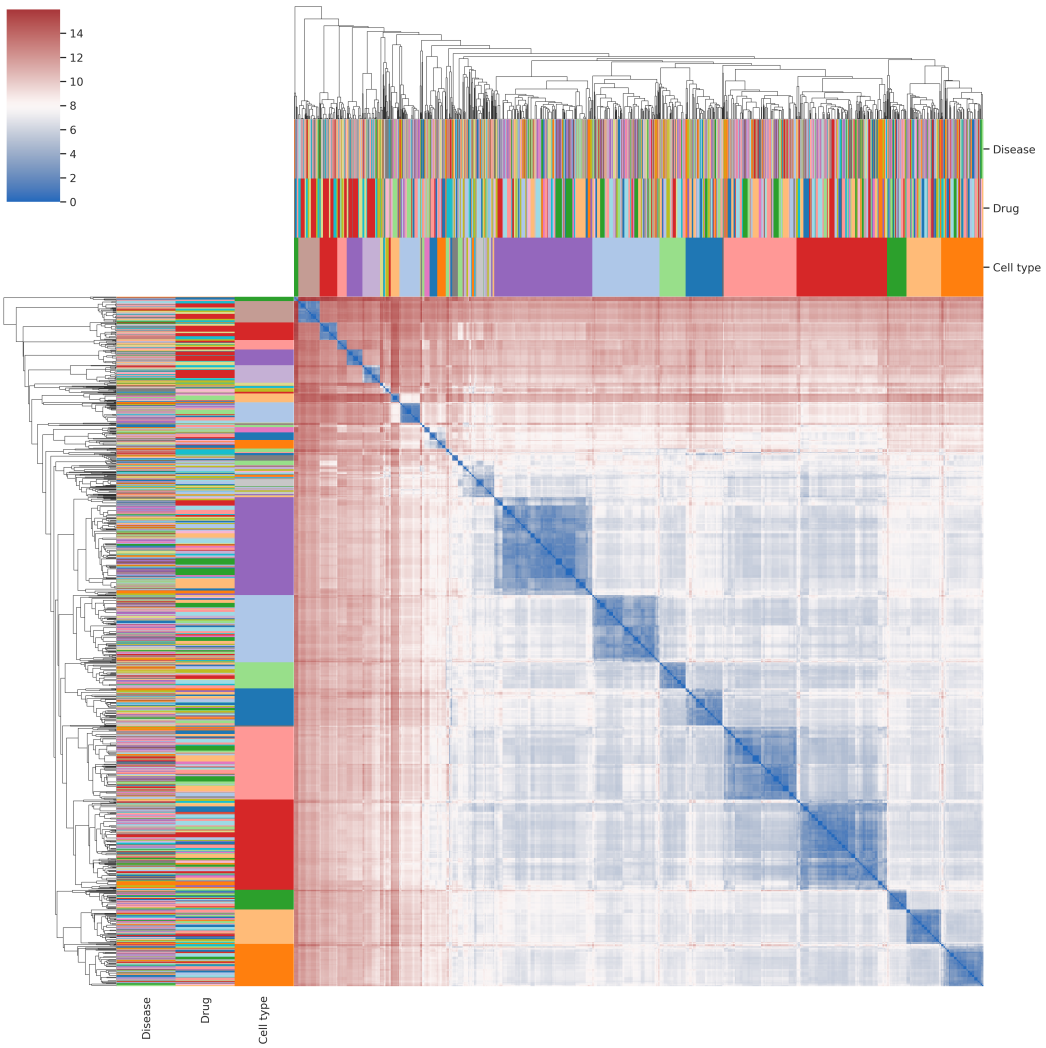

Figure 1: Hierarchical clustering of context-specific coexpression networks, annotated by cell line, disease, and small molecule drug.

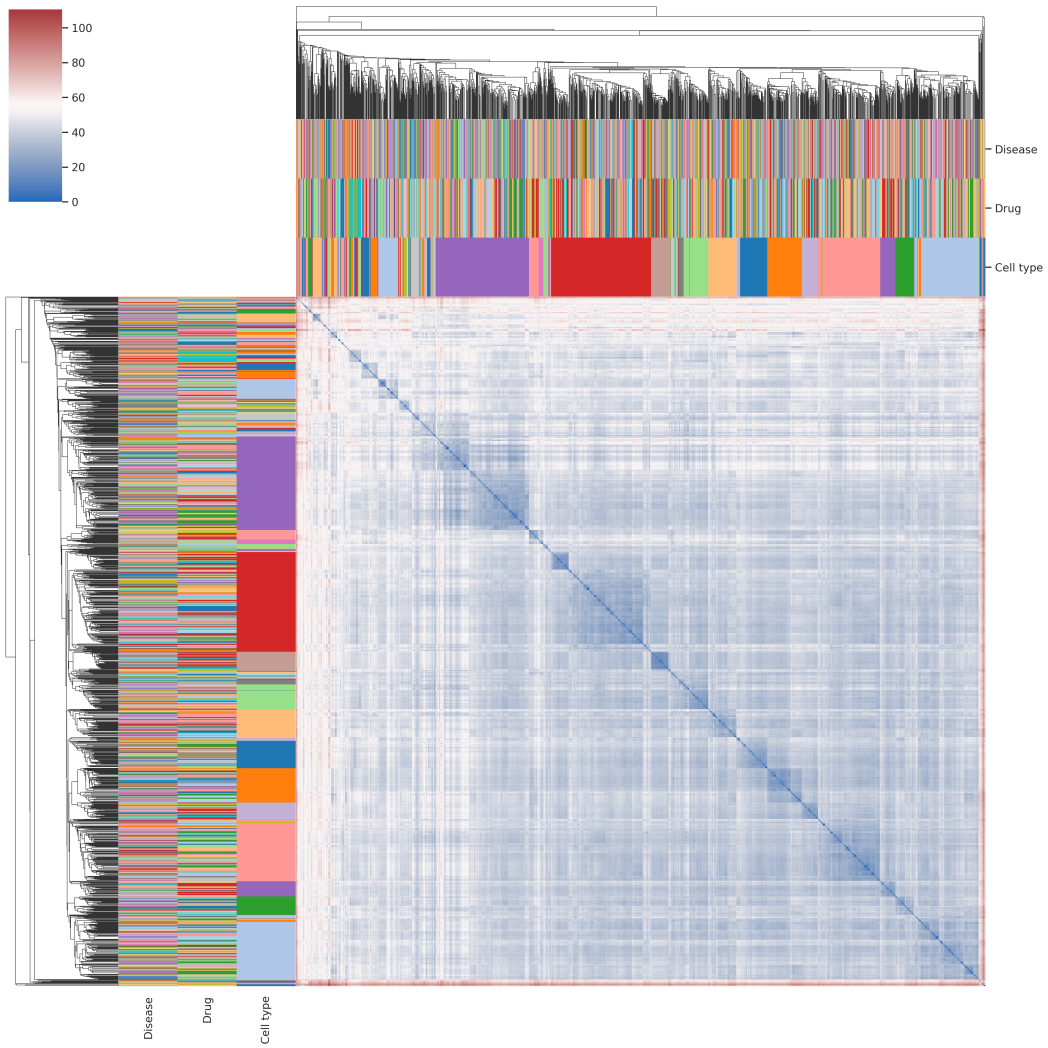

Figure 2: Hierarchical clustering of gene expression, annotated by cell line, disease, and small molecule drug.

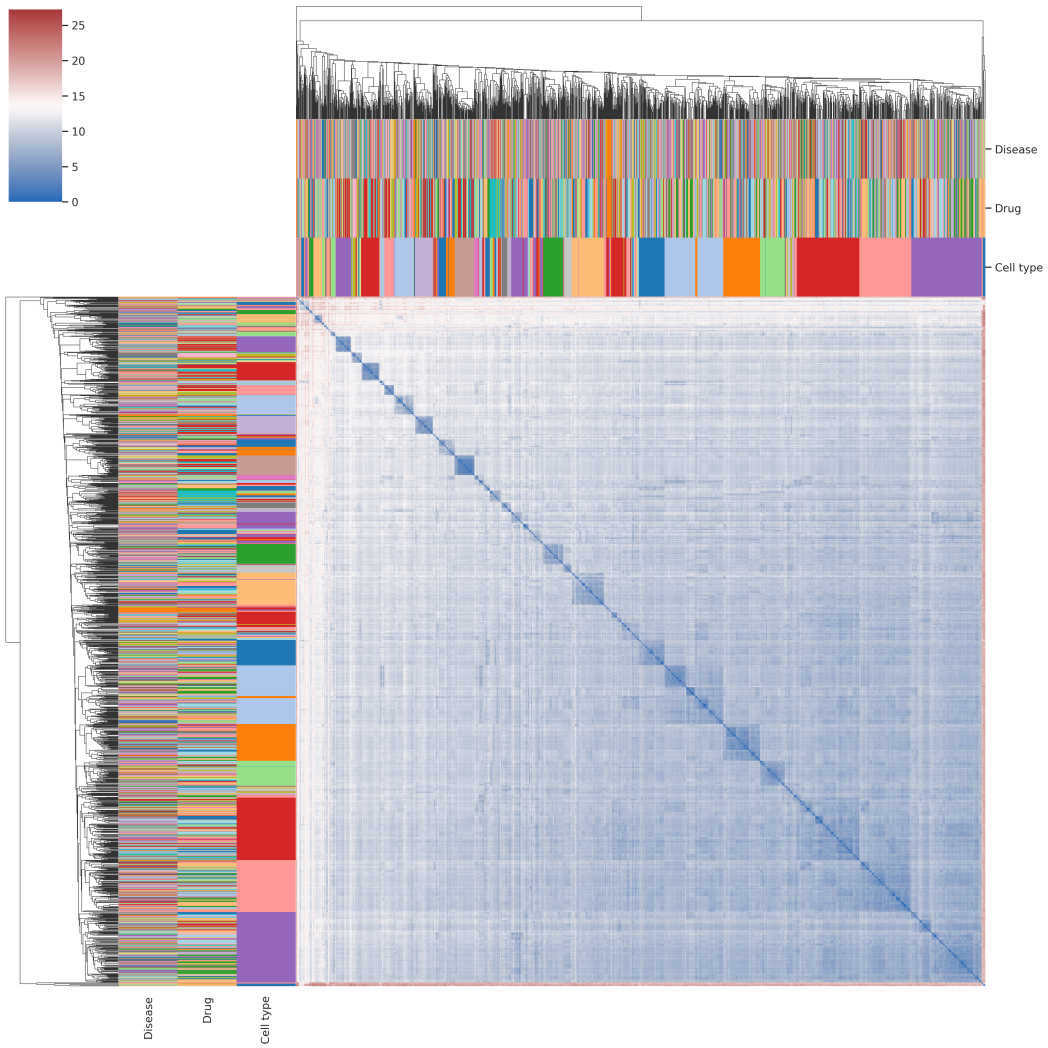

Figure 3: Hierarchical clustering of PCA metagenes, annotated by cell line, disease, and small molecule drug.