nature communications

Article

https://doi.org/10.1038/s41467-024-53457-1

# Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery

Received: 11 March 2024

Accepted: 11 October 2024

Published online: 26 October 2024

Check for updates

Xiaoning Qi[1,2,8], Lianhe Zhao[1,2,8], Chenyu Tian[3,8], Yueyue Li[3], Zhen-Lin Chen[2,4], Peipei Huo[5], Runsheng Chen[6], Xiaodong Liu[7], Baoping Wan[1], Shengyong Yang[3] ✉ & Yi Zhao[1,2] ✉

Understanding transcriptional responses to chemical perturbations is central to drug discovery, but exhaustive experimental screening of disease-compound combinations is unfeasible. To overcome this limitation, here we introduce PRnet, a perturbation-conditioned deep generative model that predicts transcriptional responses to novel chemical perturbations that have never experimentally perturbed at bulk and single-cell levels. Evaluations indicate that PRnet outperforms alternative methods in predicting responses across novel compounds, pathways, and cell lines. PRnet enables gene-level response interpretation and in-silico drug screening for diseases based on gene signatures. PRnet further identifies and experimentally validates novel compound candidates against small cell lung cancer and colorectal cancer. Lastly, PRnet generates a large-scale integration atlas of perturbation profiles, covering 88 cell lines, 52 tissues, and various compound libraries. PRnet provides a robust and scalable candidate recommendation workflow and successfully recommends drug candidates for 233 diseases. Overall, PRnet is an effective and valuable tool for gene-based therapeutics screening.

Transcriptional responses to chemical perturbations reveal fundamental insights into biological functioning and play an integral role in both disease understanding and drug discovery. Bulk and single-cell RNA-sequencing (scRNA-seq) experiments facilitate the high-throughput screen (HTS) of chemical perturbations at the omics level. Recent HTS studies[1,2] have experimental profiled thousands of independent perturbations exposing cells or cell lines to compounds. These transcriptional responses to chemical perturbations revealed coherent interpretable gene-level programs representing individual and cellular processes and quantified them in response to chemical perturbation. Although encouraging progress has been made, experimentally screening chemical perturbations remains a time-consuming and expensive process with a low discovery rate of new therapies[3]. It is unfeasible to conduct an exhaustive exploration of the vast, novel chemical perturbation space by experimentally screening disease and compound combinations.

In the past decades, deep learning-based methods have emerged as important tools for modeling transcriptional responses to perturbations. Many approaches have been proposed recently for modeling HTS perturbation responses. CPA[4] utilized an auto-encoder-based model to map chemical induce transcriptomic effect into a latent space to reconstruct perturbation response. Biolord[5], scGen[6], and

[1]Research Center for Ubiquitous Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. [2]University of Chinese Academy of Sciences, Beijing, China. [3]Department of Biotherapy, Cancer Center and State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, Sichuan, China. [4]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. [5]Luoyang Institute of Information Technology Industries, Luoyang, Henan, China. [6]West China Hospital, Sichuan University, Chengdu, Sichuan, China. [7]University of Chinese Academy Sciences, Nanjing, Jiangsu, China. [8]These authors contributed equally: Xiaoning Qi, Lianhe Zhao, Chenyu Tian. ✉e-mail: yangsy@scu.edu.cn; biozy@ict.ac.cn

scVIDR[7] performed counterfactual prediction via deep encoder-decoder/generator-based generative frameworks, which take a control cell and unseen labels as input to predict the gene expression of the unseen cellular states. These recently published cell perturbation modeling tools precisely simulated chemical perturbations and predicted chemical perturbations in gene expression of unseen cell types, but few predict responses to novel chemicals. Following CPA[4], chemCPA[8] introduced a new encoder-decoder architecture that incorporated the compounds' structure to predict the perturbational effects of unseen drugs. Deep generative frameworks with encoder-decoder and its variants effectively predicted single-cell gene expression perturbations. CellOT and CINEMA-OT[9,10] leveraged optimal transport to match paired unperturbed-perturbed observations. Optimal-transport-based matched the observations experimentally perturbed but were incapable of modeling novel perturbations, such as novel compounds or novel cell types. Linear regression-based methods[11,12] estimate the impact of perturbations on gene expression by linearly combining the effects of genetic perturbation. However, this linear combination approach leads to limitations in accurately modeling the nonlinear chemical perturbations across diverse cell types and compound combinations. GEARS and CellOracle[13,14] leveraged a knowledge graph of gene-gene relationships to predict genetic perturbation outcomes. However, graph-based models relied on accurate prior knowledge leading to the lack of scalability. Given that most diseases are associated with characteristic gene expression profiles, Connectivity Map (CMap)[15] proposed a concept that connected genes, drugs, and diseases by virtue of common gene-expression signatures, leading to projects such as CMap[15], L1000[1], etc. Inspired by this concept, DLEPS[16], OCTAD[17] and other studies (such as[18–20], etc.) utilized gene signature matching methods to screen candidates by finding drugs that reverse the disease signature. DLEPS[16] predicted chemically induced changes in transcriptional profiles directly from molecular structure, and OCTAD[17] virtually screened compounds by matching the cancer-specific expression signature to compound-induced gene expression profiles. Although DLEPS and OCTAD screened candidates effectively based on bulk HTSs, they failed to predict cell-type-specific transcriptional response to novel perturbations and model cellular heterogeneity, which is highly relevant to treatments. Consequently, perturbations response models are required to address the limited exploration power to novel perturbations of existing experimental and computational methods, which are also needed for predicting the response to unseen perturbations and discovering promising therapeutic drug candidates. Deep generative models, including Generative Adversarial Network (GAN[21]), Variational Auto-Encoder (VAE[22]), Denoising Diffusion Probabilistic Model (Diffusion[23]), Normalization Flow (NF[24]), Generative Pre-Trained Transformer (GPT[25]), and so on, learn the probability density of observable samples and generate new samples. Deep generative models have greatly improved diverse areas (for example, natural language processing[25,26], computer vision[27,28], chemicals[29], and so on), suggesting the potential for applications in drug discovery.

Here we present PRnet, which is a flexible and scalable perturbation-conditioned deep generative model for predicting transcriptional responses to novel chemical perturbations that were never experimentally perturbed at bulk and single-cell levels. PRnet is a new encoder-decoder architecture based generative model which comprises three components, including the Perturb-adapter, the Perturb-encoder, and the Perturb-decoder. PRnet adapts novel compounds and diseases in various perturbation scenarios by taking compound structures and unperturbed transcriptional profiles as input to predict transcriptional responses. The Perturb-adapter uses simplified molecular-input line-entry system[30] (SMILES) chemical encoding as input, enabling generalization to unseen compounds without prior knowledge and annotation. The learnable latent space of PRnet facilitates gene-level response interpretation and capturing heterogeneity. PRnet was trained with close to one hundred million bulk HTS observations perturbed by 175,549 compounds and tens of millions single cell HTS observations perturbed by 188 compounds. Crucially, the model operates as a data-driven model, allowing for effective generalization to novel perturbations. The evaluation indicated that PRnet outperformed alternative approaches in predicting changes and expression in transcription response to novel compounds, pathways, and cell lines in bulk and single-cell HTS data. To further validate the effectiveness, PRnet has been utilized to identify novel bioactive compounds against small cell lung cancer (SCLC) and seek novel natural compounds against colorectal cancer (CRC). Experimental validation demonstrated the activity of novel candidate compounds against SCLC and CRC cell lines within the appropriate predicted ranges of concentration.

This model's flexibility and scalability make it a valuable tool to screen candidates for various diseases. We, therefore, leveraged PRnet to in silico screen various compound libraries and generated a virtual large integration atlas of perturbation profiles, covering 88 cell lines and 52 tissues, as well as compound libraries comprising 935 FDA-approved drugs, 4158 active compounds, 30,456 natural compounds, and 29,670 drug-like compounds. PRnet also provided a robust and scalable candidate recommendation workflow for diseases according to the reference changes in gene sets. Given disease-specific or sensitive compound gene signatures, gene set enrichment analysis (GSEA) is employed to assess the potential efficacy of compounds against these diseases. PRnet successfully recommended 577 drug candidate lists from 577 studies for 233 different diseases based on the profiles atlas. In three cases of metabolic disorders, including non-alcoholic steatohepatitis (NASH), polycystic ovary syndrome (PCOS) patients, and inflammatory bowel disease (IBD), drugs recommended by PRnet have been supported by previous literature with human or animal studies. PRnet enables effective prediction of transcriptional responses to novel complex chemical perturbations and screening large-scale compound libraries for specific diseases, thus being a valuable tool for gene-based therapeutics screening.

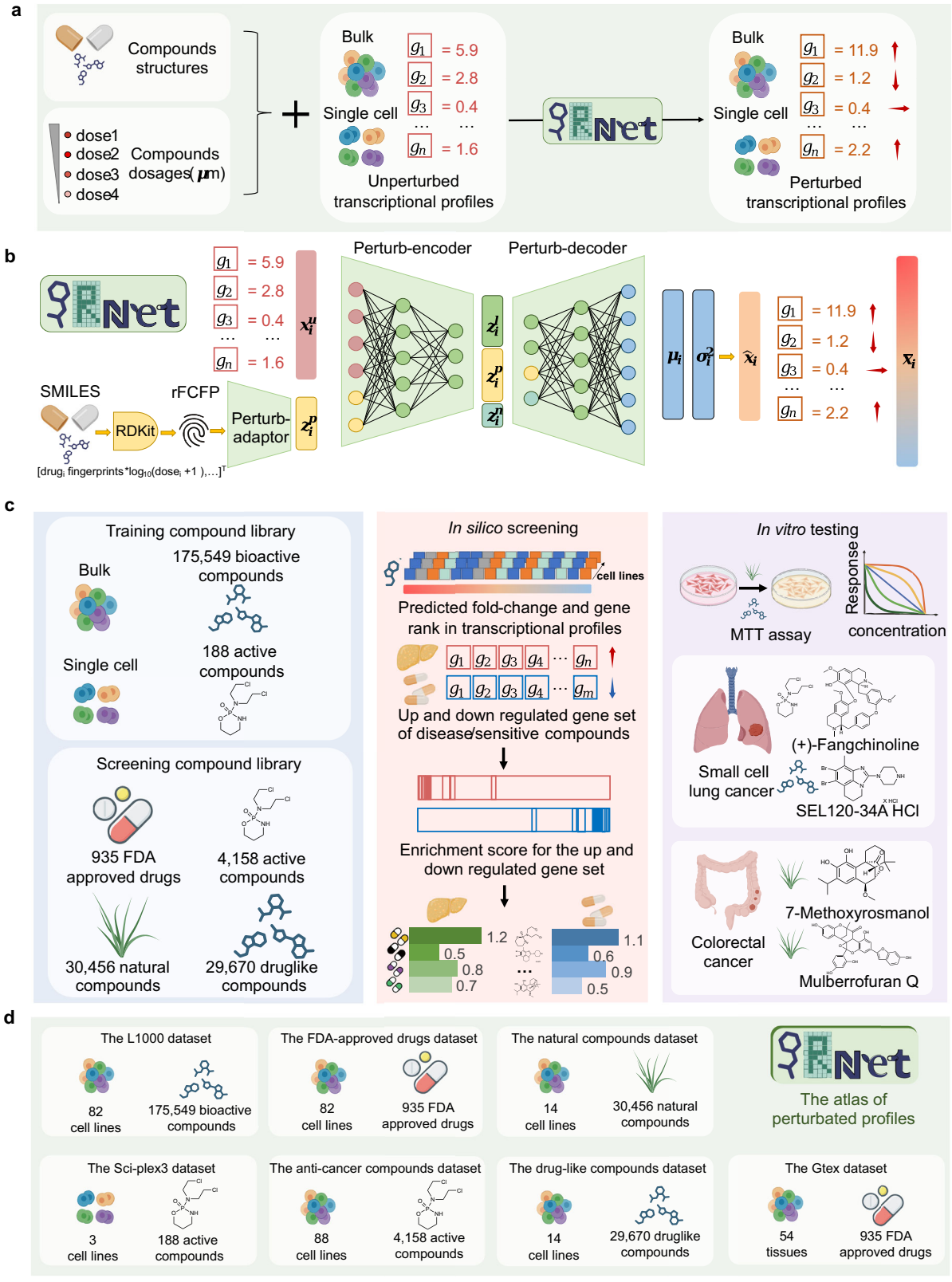
## Results

### PRnet: perturbation-conditioned deep generative model for transcriptional response prediction

In this paper, we formulate the transcriptional response prediction as a distribution generation problem conditioned on perturbations. Cells inherently recognize and respond to chemical stimuli, with their responses influenced by both external stimuli and their intracellular state. In the single cell or bulk HTS, transcriptional responses to chemical perturbations are affected by multiple conditions, such as compound structures, compound dosages, and covariates, such as cell types, and cell lines. Given an n-dimensional unperturbed transcriptional profile (a single cell or bulk RNA-seq observation) and a chemical perturbation imposed by a compound with a specific structure and dosage, PRnet aims to predict the distribution of the perturbed transcriptional profile. Modeling perturbation patterns enables capturing the novel chemical perturbation effect on gene-level programs and quantifying them in various perturbation scenarios (Fig. 1a).

PRnet (as detailed in "Methods") is a flexible and scalable perturbation-conditioned deep generative model designed to predict transcriptional responses to novel complex chemical perturbations at bulk and single-cell levels. The design of PRnet comprises three components, including the Perturb-adapter, the Perturb-encoder, and the Perturb-decoder (Fig. 1b). In preprocessing, each perturbed profile is assigned an unperturbed transcriptional profile of the same cell line. Initially, given a chemical perturbation imposed by the structure of compounds represented by Simplified Molecular Input Line Entry System[30] (SMILES) strings and the dosages of the corresponding compounds, PRnet leverages RDKit[31] to capture the functional topology information of the structures, generate the Functional-Class

https://doi.org/10.1038/s41467-024-53457-1




Fingerprints (*FCFP*) of compounds. These *FCFP*s were scaled by their dosages and then summed to generate rescaled Functional-Class Fingerprints (*rFCFP*) embedding. Without loss of generality, for the $i$-th compound perturbation, the Perturb-adapter encodes the fingerprint $rFCFP_i$ to an additive latent embedding $z_i^p$ which allows generalization to novel compounds and compound combinations. Then, the Perturb-encoder maps the chemical perturbation effect on heterogeneous unperturbed states $x_i^u$ into the interpretable latent space $z_i^l$. At last, the Perturb-decoder estimates the distribution of the transcriptional response $\mathcal{N}(x_i|\mu_i, \sigma_i^2)$ within the context of the chemical perturbation effect on the unperturbed state $z_i^l$, the applied perturbation $z_i^p$ and a noise $z_i^n$. PRnet encodes the chemical effect on the unperturbed state to a learnable latent space, estimates the distribution, and performs conditioned sampling to generate a transcriptional response with

**Fig. 1 | Overview of the methodological approach. a** Problem formulation: Given unperturbed transcriptional profiles (single cell or bulk) and applied perturbations (structures and the dosages of the compounds), predict transcriptional responses. Red arrows indicate changes in transcriptional profiles. **b** Model architecture: PRnet is a perturbation-conditioned deep generative model for transcriptional response prediction with three components, including the Perturb-adapter, the Perturb-encoder, and the Perturb-decoder. Crucially, the model operates as a data-driven model, allowing for effective generalization to novel perturbations. **c** Screening candidates with PRnet: The training compound library comprises a bulk high-throughput screening library (175,549 bioactive compounds), and a single-cell high-throughput screening library (188 active compounds). The screening library comprises four compound libraries, including 935 FDA-approved drugs, 4158 active compounds, 30,455 natural compounds, and 29,670 in-house druglike compounds, respectively. For in-silico screening, PRnet initially predicts the average transcriptional profile, fold-change in the gene, and gene rank after perturbing the specific cell line with screening compounds. Given the gene signature for a particular disease or sensitive drug, the enrichment scores of screening compounds were computed for compound ranking. In vitro validation experiments were conducted using MTT assays on colorectal cancer and small cell lung cancer cell lines to validate the activity of compound candidates. After in vitro validation, PRnet is utilized to recommend drug candidates for 233 different diseases. **d** The large-scale integration atlas of perturbation profiles, including the L1000 dataset, the Sci-plex3 dataset, the FDA-approved drugs dataset, the anti-cancer compounds dataset, the natural compounds dataset, the drug-like compounds dataset, and the Gtex dataset. Some icons were created in BioRender. Qi, X. (2024) biorender.com/l45e810.

biological and chemical contexts. Sampling generates a specific transcriptional profile $\hat{x}_i$ that provides gene-level up- and down-regulation information. For bulk HTS data, the predicted transcriptional responses of 978 landmark genes $\hat{x}_i$ are transformed into 12,328 genes by linear transformation. For single-cell HTS data, 5000 highly variable genes (HVGs) of transcriptional profile are selected. SMILES[30] is widely used for representing chemical structures due to its simplicity and efficiency in encoding complex molecules as strings. By taking SMILES of the compound as the input, the Perturb-adapter has sufficient flexibility to screen large-scale compound libraries without any prior knowledge. Driven by data, PRnet automatically identifies the heterogeneity in latent space corresponding to compound, dosage contexts, and cell-type specific contexts. This allows the model to directly generalize to novel perturbation scenarios involving novel compounds, pathways, cell types, and cell lines that have not been previously perturbed.

With the ability to predict transcriptional responses to novel perturbations, PRnet enables efficient screening of candidates for complex diseases (Fig. 1c). Inspired by the assumption embodied in the CMap[15] concept that gene signatures are used as indicators reflecting the underlying mechanisms of diseases, PRnet predicted new therapeutic candidates by finding drugs that reverse the disease signature. There are two steps to applying PRnet to downstream tasks. In step 1, for in silico screening, PRnet predicts the transcriptional profile of specific cell lines perturbed by a user-defined compound library with multiple gradient concentrations. In step 2, we calculate the average transcriptional profile and fold-change in gene expression of each compound, and rank genes according to their fold-change values. Then, given the query gene signature for a particular disease or known sensitive compounds, gene set enrichment analysis (GSEA[32]) was employed to evaluate compound efficacy with their enrichment scores. Finally, compounds are ranked based on these enrichment scores. Large-scale high-throughput screening data are initially fitted to the model, facilitating its adaptability to diverse compound libraries and diseases.

PRnet was trained on two compound libraries and screened four compound libraries (Fig. 1c). The training compound libraries contain a bulk high-throughput screening library consisting of over 883,269 transcriptional profiles of 175,549 bioactive compounds[1], and a single-cell high-throughput screening library consisting of 290,888 transcriptional profiles of 188 active compounds[2]. Being well trained in HTS observations, PRnet enables in silico high-throughput screening of novel compound libraries for various cell lines. PRnet has further been applied to screen active and drug-like compounds for SCLC and natural compounds for CRC. In vitro validation experiments with MTT assays confirmed the efficacy of the candidate compounds against SCLC and CRC cell lines. Lastly, PRnet screened four compound libraries and generated a large-scale integration atlas of perturbation profiles (Fig. 1d), including (1) 82 cell lines perturbed by 935 FDA-approved drugs, (2) 88 cell lines perturbed by 4158 active compounds, (3) 14 CRC cell lines perturbed by 30,456 natural compounds, (4) 6 SCLC cell lines perturbed by 29,670 drug-like compounds and (5) 54 tissues perturbed by 935 FDA-approved drugs. Based on the large-scale integration atlas of perturbation profiles, PRnet is capable of a variety of downstream applications. PRnet has been utilized to recommend drugs for 233 different diseases. PRnet successfully predicted drug candidates for these diseases and demonstrates the potential in drug discovery.

## PRnet robustly predicted the response of novel perturbation and learned interpretable latent embeddings

To evaluate the performance of PRnet for predicting responses to unseen perturbations, all datasets were strictly split by perturbation attributes (*compound_split*, *cell_line_split,* and *pathway_split*) into 3 subsets: train, validation, and test (Supplementary Fig. 1a). The held-out test sets were used to simulate datasets of novel perturbations. Three train-test data split strategies were employed to assess the performance of out-of-distribution perturbation scenarios, including (1) Random Split: randomly divides compounds and cell lines, (2) Unseen Compounds: testing compounds not seen perturbed during training, and (3) Unseen cell lines: testing cell lines not seen perturbed during training. Five-fold cross-validation was applied in each split strategy, and the average performance over five folds was computed as the overall metric for comparison. Two high-throughput screening data of different resolutions were used to test model performance, consisting of a bulk HTS dataset (from the L1000 project[1]) and a single cell HTS dataset (from the sci-Plex3 assay[2]). All models were trained and compared separately on two HTS datasets.

We employed bulk high-throughput screening data from the L1000 project[1] to fit the model, in which 978 genes (hereinafter called landmark genes) were selected to represent the diversity of biological pathways and processes in human cells. We first preprocessed these data (as detailed in "Methods") and obtained 836,352 paired bulk RNA-seq observations (represented by the expression levels of 978 landmark genes), covering 82 cell lines and their perturbation data perturbated by 175,549 compounds. To quantitatively evaluate the compound-induced gene expression changes, we compared the Pearson correlation between the true and predicted post-perturbation of the average logarithm of fold-change in gene expression (log(FC)) for the hold-out test set with alternative approaches. The "Pearson of log(FC) in compounds" metric evaluated the Pearson correlation between the true and predicted mean log(FC) perturbed by the same compound in the test set. We demonstrated the performance of the PRnet on the bulk HTS data in Fig. 2a, where a higher value indicates a better performance. PRnet consistently demonstrated the best performance across all three split strategies, particularly well-fitted in unseen compound prediction scenarios with an average Pearson Correlation (PCC) of 0.8. PRnet significantly outperformed in predicting unseen cell line log(FC) with an increase in PCC over 0.3 compared to other approaches. Some hold-out predicted cases in

https://doi.org/10.1038/s41467-024-53457-1

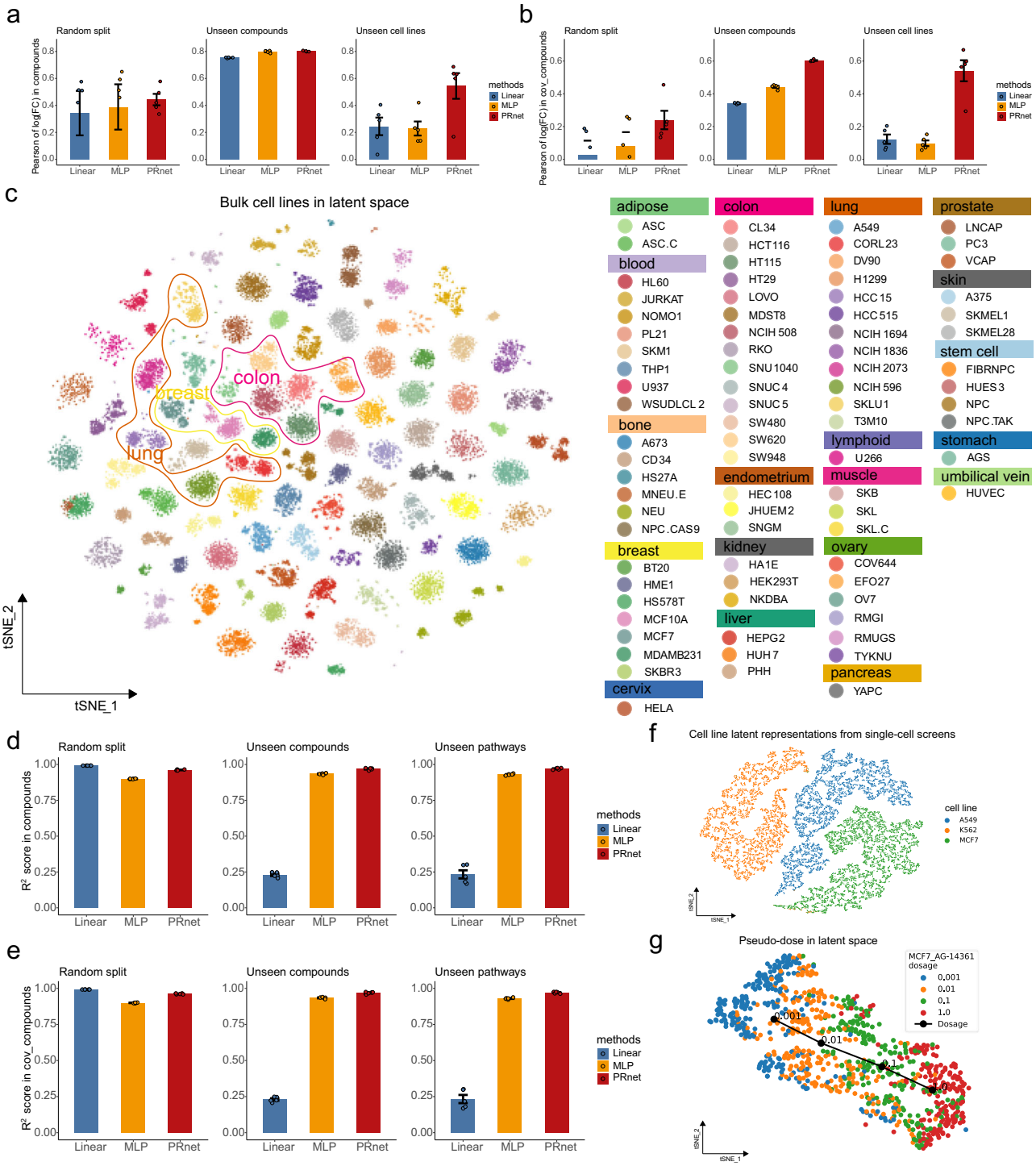



**Fig. 2 | PRnet outperforms alternative approaches in predicting post-perturbation change in gene expression. a** Performance of out-of-distribution perturbation scenarios in three train-test data split strategies (1) Random split, (2) Unseen compounds split, and (3) Unseen cell lines split. The "Pearson of log(FC) in compounds" metric evaluates the mean log(FC) perturbed by the same compounds in the test set, which are presented as mean values ± SD. Error bars indicate the standard error (SD) for each method, and the points ($n = 5$) refer to the 5-fold used for validation. **b** The "Pearson of log(FC) in cov_compounds" metric evaluated mean the (log(FC)) of compounds within the same cell line in the test set. Results are presented as mean values ± SD ($n = 5$). **c** T-SNE representations of latent embeddings learned by PRnet. Post-perturbation transcriptional profiles are from 82 cell lines from 20 different organs/tissues. **d** The "R2 in compounds" metric evaluates the R2 score of the average gene expression perturbed by the same compounds in the test set. Results are presented as mean values ± SD ($n = 5$). **e** The "R2 in cov_compounds" metric evaluates the R2 score of the average gene expression perturbed by the same compounds within the same cell line. Results are presented as mean values ± SD ($n = 5$). **f** T-SNE representation of latent embeddings from massive single-cell screens of 188 drugs across 3 cancer cell lines. **g** T-SNE representation of latent embedding of transcriptional profiles from MCF7 cell line perturbed with AG-14361. The pseudo-dose trajectory is displayed with a black line. Source data are provided as a Source Data file.

predicting transcriptional responses of unseen compounds and cell lines were illustrated in Supplementary Fig. 1b–d. In more challenging scenarios, we evaluated the performances of predicting cell-line-specific compound-induced changes in genes by the "Pearson of log(FC) in cov_compounds" metric. The "Pearson of log(FC) in cov_compounds" metric is the Pearson correlation between the true and predicted mean log(FC) perturbed by the same compound within the same cell line in the test set. PRnet achieved the best performance in the "Pearson of log(FC) in cov_compounds" metric across three scenarios. In particular, PRnet exhibited more than two times better performance in unseen cell line predictions compared to other methods and demonstrated improvements of 0.16 in unseen compound predictions, demonstrating the generalization of PRnet to novel perturbations (Fig. 2b).

To better characterize the heterogeneous gene-level change under certain perturbations, it is desirable to identify the set of cells or cell lines and isolate the precise variations enriched in the data from the corresponding cell lines or cells. After training, PRnet learned interpretable embeddings in the latent space in the context of the base unperturbed state and the applied perturbation. The low-dimensional (t-SNE[33]) representation (Fig. 2c) illustrates the latent embeddings of post-perturbation transcriptional profiles learned by PRnet. In the latent space, embeddings from the same cell line tend to form a cluster together. Each cancer cell lines form specific gene-level responses for corresponding perturbations. In a way, PRnet captures strong cell line-specific transcriptional profile variations under different conditions. Interestingly, we observed that the embedding learned by PRnet also represents cell line similarity in response to various perturbations. Figure 2c illustrates the t-SNE representation of the latent embeddings for all cell lines, where cell lines originating from the same organ exhibit a similarity preference in the latent space, resulting in close spatial locations, such as cell lines from the colon, breast, and lung. Supplementary Fig. 6a–c demonstrates in detail the low-dimensional (t-SNE) representation of the latent embeddings for cell lines from colorectal cancer, breast cancer, and lung cancer, respectively. To quantitatively evaluate the similarity of perturbations among cell types, we computed the normalized cosine similarity among the mean embeddings in the latent space for all cell lines. The resulting cosine similarity heatmap (Supplementary Fig. 6g) shows that most cell lines from the same tissue exhibit higher similarity in the latent space compared to those from different tissues.

Human cancer cell lines have facilitated drug discoveries in cancer biology, but they are neither clonal nor genetically stable. This instability can generate variability in drug sensitivity, as shown by Ben-David et al.[34]. The genomic evolution of cancer cell lines leads to a high degree of variation across cell line strains. To explore the impact of inter- and intra-heterogeneity of cell lines on drug responses, we conducted additional experiments focusing on A549 and MCF7 cell strains, as described by Ben-David et al.[34] in Supplementary Note 2. These experiments were designed to explore inter- and intra-cellular heterogeneity, similar patterns of the same MOA drug across cell strains, and screening candidate compounds for heterogeneous cells. These results indicated that drug response was highly similar to genetic or gene expression which aligned with findings in the original study. These findings underscore the utility of our model in capturing both inter- and intra-heterogeneity, enhancing its application in screening for specific cancer strains.

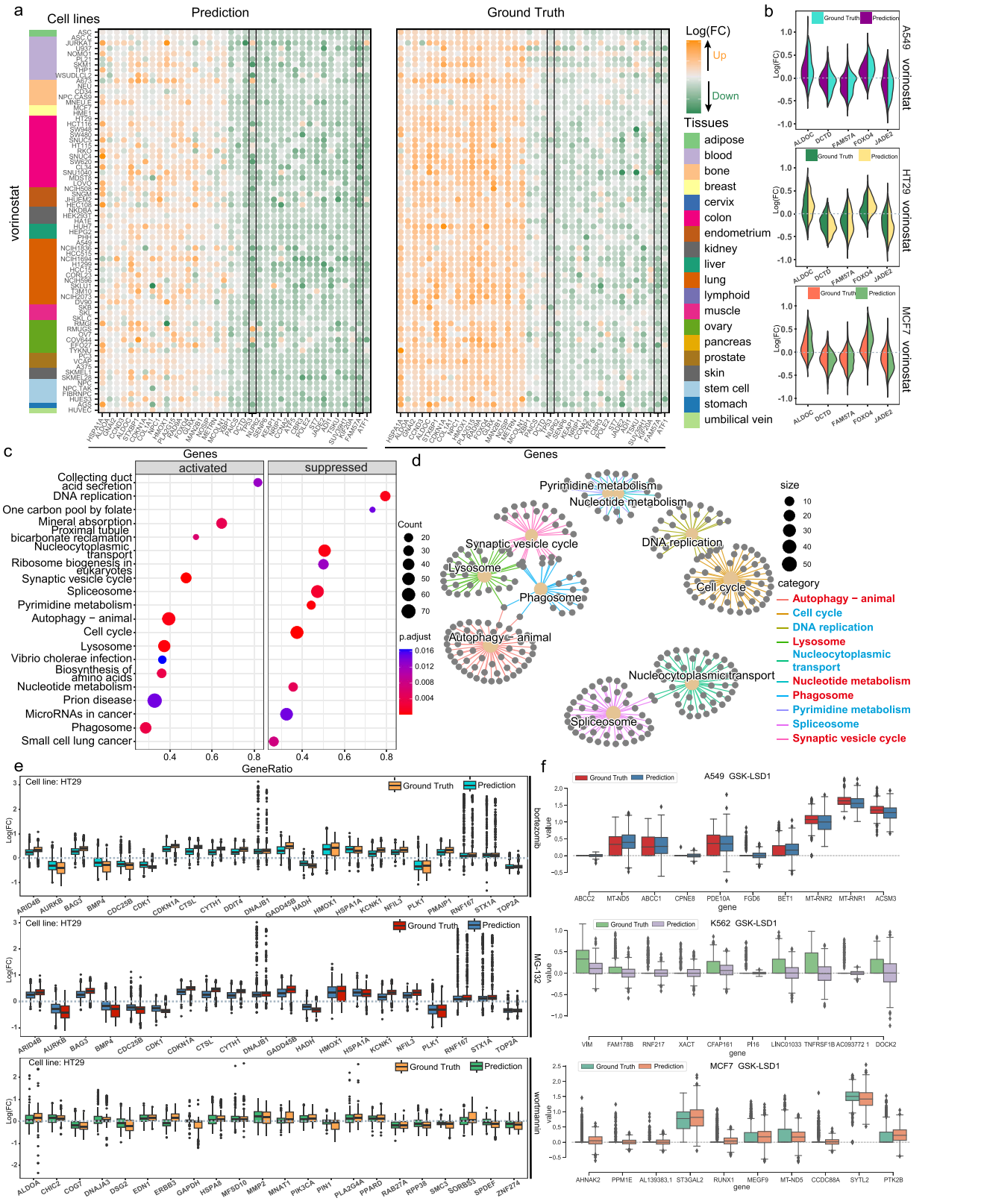
PRnet adapted to model profiles of HTS at different resolutions. The sci-Plex assay[2] screened 188 compounds in 3 cancer cell lines at single-cell resolution, measuring millions of cells. The screened cell lines A549 (lung adenocarcinoma), K562 (chronic myeloid leukemia), and MCF7 (breast adenocarcinoma) were treated to each of these 188 compounds at four dosages (10 nM, 100 nM, 1 μM, 10 μM). To quantitatively evaluate the performance of PRnet in single-cell HTS data, we followed commonly used metric[4–6] to compare the $R^2$ score between the true and predicted post-perturbation gene expression for the hold-out test set with alternative approaches (Fig. 2d, e). The "$R^2$ in compound" metric evaluated the $R^2$ score of the average post-perturbation gene expression of the same compound in the test set. We also compared the cell-type-specific performance "$R^2$ in cov_compound" metric, which evaluated the $R^2$ score of the average post-perturbation of the same compound within each specific cell line. PRnet outperformed alternative models in the "$R^2$ in the compound" and "$R^2$ in cov_compound" metrics in unseen compounds ($R^2$ in the compound: 0.969) and unseen pathways scenarios ($R^2$ in the compound: 0.97) than the other methods. The low-dimensional (t-SNE) representation illustrates latent embeddings learned by PRnet from massive single-cell screens of 188 compounds across 3 cancer cell lines (Fig. 2f). The latent embeddings automatically cluster the concordance of cells to their cell types. The results demonstrated that PRnet not only captures the heterogeneous responses of large-scale HTS but also resolves similar responses of homogeneous cells to various perturbations. The t-SNE embedding (Fig. 2g) in the latent space of MCF7 cells treated with AG-14361 produces a pseudo-dose trajectory, indicating that AG-14361 induced heterogeneous responses. Several other pseudo-dose trajectory examples were illustrated in Supplementary Fig. 6d–f.

To validate the clinical relevance of PRnet, we incorporated scRNA-seq data from a cohort of pediatric acute myeloid leukemia (AML) patients[35]. We trained PRnet to predict transcriptional responses to chemotherapy treatments and validated its potential for clinical application. PRnet showed robustness in predicting transcriptional responses for pediatric AML patients post-chemotherapy, demonstrating PRnet's potential to assist in clinical applications. For detailed experimental results, please see Supplementary Note 1 and Supplementary Fig. 8.

## PRnet captures gene programs in various perturbation scenarios

We reason that PRnet can be a valuable tool to analyze and capture functional transcriptional experiments that aim to uncover gene programs or effects on various conditions by introducing perturbation to the system. To investigate this in greater detail, we first collected and tested the hold-out perturbation data of compound vorinostat. Vorinostat (suberoylanilide hydroxamic acid) is the first FDA-approved HDAC inhibitor for the treatment of cutaneous manifestations of cutaneous T-cell lymphoma (CTCL) and is also currently being studied as monotherapy and in combination therapy for other types of cancers[36,37]. Results demonstrated a comprehensive ability of PRnet to capture gene-level fold changes in all of the 71 cell lines (Fig. 3a). By comparing the transcriptional responses of cell lines from different organs, we observed that PRnet captured cell-type-specific responses. For instance, cell lines from muscle exhibited relatively weaker responses and smaller fold changes compared to those from the lung. Figure 3a shows that PRnet correctly captures both the right trend and the magnitude of perturbation of the top up- and down-regulation genes across 71 cell lines from 16 tissues/organs. Taking gene FAM57A and TP53 as examples, PRnet made accurate predictions in cases of both up- and down-regulation in all cell lines after perturbation. In addition, PRnet even correctly predicted the fold change values in the expression of FAM57A across all cell lines. Figure 3b shows a detailed comparison between the predicted and actual distributions of fold changes in gene expression for three representative cell lines from different organs (HT29: colorectal adenocarcinoma, A549: lung adenocarcinoma and MCF7: breast adenocarcinoma,) treated with vorinostat. It can be observed that PRnet aligns consistently with the distribution of predicted and true observations and accurately predicts the up- and down-regulation trends of the top 5 genes with high log(FC) values. We employed the KEGG pathway Gene Set Enrichment Analysis (GSEA) on the average predicted post-perturbations gene rank of Vorinostat across all cell lines. The GSEA results (Fig. 3c, d and

https://doi.org/10.1038/s41467-024-53457-1




Supplementary Fig. 3a) reveal that Vorinostat is enriched in pathways related to fundamental cellular processes related to tumor suppressor mechanisms. The GSEA results indicated that Vorinostat suppresses pathways such as Cell cycle, DNA replication, and Spliceosome, and activates pathways including Autophagy - animal, Lysosome, Phagosome, and so on, which are all associated with autophagy and apoptosis in tumor cells[38].

To demonstrate the generalization of PRnet, we also analyzed perturbation observations of other hold-out compounds on a gene-by-gene basis. We collected and tested some case perturbation data of HT29 with the most observations. Figure 3e illustrates the log(FC) of the top 20 up- and down-regulated genes in multiple perturbations of HT29 cell lines treated with the hold-out compounds bortezomib, MG-132, and wortmannin, respectively. These results suggested that

**Fig. 3 | PRnet captures gene programs in various perturbation scenarios. a** The heatmap illustrates the average log(FC) of the gene expression profiles predicted by PRnet and the ground truth for 71 cell lines under the drug Vorinostat. The horizontal axis represents genes, while the vertical axis represents cell lines, with the color orange indicating upregulation and the color green indicating downregulation. **b** The violin plots illustrate the distribution of log(FC) for the top 5 up- and down-regulated genes in multiple perturbations of cell lines, including A549, HT29, and MCF7, from three cancer types treated with the drug Vorinostat. **c** The GSEA results of average post-perturbation changes in expression of 71 cell lines perturbed by Vorinostat. GSEA was performed for KEGG pathway enrichment. Enriched terms were identified by adjusted *p*-values (p.adjust) < 0.05 (Benjamini-Hochberg method). **d** The category net plot visualizes the functional enrichment result of Vorinostat with suppressed pathways colored in blue and activated pathways colored in red. **e** The box plots illustrate the log(FC) of the top 20 up-and down-regulated genes in cell lines HT29 with the drug bortezomib ($n = 865$ observations), MG-132 ($n = 892$ observations), and wortmannin ($n = 207$ observations), respectively. The middle line in the box plot, median; box boundary, IQR; whiskers, 1.5 × IQR; minimum and maximum, not indicated in the box plot; gray dots, points beyond the minimum or maximum whisker. **f** The box plots illustrate the expression of the 10 different expression genes of cell lines A549 ($n = 450$ cells), K562 ($n = 430$ cells), and MCF7 ($n = 988$ cells) perturbed by the drug GSK-LSD1, respectively. The middle line in the box plot, median; box boundary, IQR; whiskers, 1.5 × IQR; minimum and maximum, not indicated in the box plot; gray dots, points beyond the minimum or maximum whisker. Source data are provided as a Source Data file.

PRnet is able to capture regulated gene level information that is consistent with evidence from corresponding compounds in inferring different cancer transcriptional profile conditions that can be missed by perturbations analysis. The predicted distribution of fold changes in gene expression after perturbation closely aligned with the actual observed distribution, indicating the accuracy of PRnet in capturing the perturbation effects. More predicted gene-level perturbation responses of cell lines of breast cancer exhibited similar performance (Supplementary Fig. 2a). Besides, Fig. 3f demonstrates the ability of PRnet to predict cell-type-specific gene-level perturbation transitions in single-cell HTS observations. In the case of predicting the response of perturbing A549, K562, and MCF7 cell lines with GSK-LSD1, PRnet correctly captured both the right trend in gene expression and the magnitude of response across all 10 differentially expressed genes (Fig. 3f). Similar performance was observed for several other examples across perturbation conditions (Supplementary Fig. 4a–c). The ability to capture changes in gene-level programs under different compound conditions and resolutions indicates the robustness, generalization, and precise performance of PRnet in predicting perturbation responses.

## PRnet identified active compounds against small-cell lung cancer

Having been trained to simulate experimental measurements of high-throughput screening, PRnet was applied to identify potential novel compound candidates for the treatment of small cell lung cancer (SCLC). SCLC is an extremely aggressive lung cancer characterized by small cells with limited cytoplasm forming clusters or spheroids[39]. Despite an initial positive response to conventional chemotherapy and radiation, SCLC often recurs rapidly, with less than 5% of patients surviving five years. Currently, highly effective treatment options for this disease remain unresolved, making drug development efforts for this cancer a high priority.

Given several novel cell lines of SCLC, namely NCI-H69, NCI-H526, NCI-H446, NCI-H209, and NCI-H196, and DMS114, we first employed PRnet to predict the transcriptional response of SCLC cell lines to sensitive compounds. Then, we in silico screened two user-defined compound libraries to identify potential compound candidates against SCLC (Fig. 4a), namely an active compound library (4158 compounds) from Selleckchem, and an in-house druglike compound library (29,670 compounds). Through in silico screening, PRnet predicted the transcriptional responses of each compound across eight concentration gradients perturbing 6 cell lines, with each scenario repeated three times for computational robustness and calculated gene rank of changes in the average post-perturbation expression of each compound. After that, the predicted up/downregulated genes of sensitive compounds on their cell lines were used as the GSEA gene signature input to calculate the enrichment scores. We then performed GSEA to calculate the enrichment scores of compounds in libraries and ranked them according to their scores (Supplementary Datas 1, 2). Ultimately, three compounds ((+)-Fangchinoline, (+)-JQ-1, and SEL120-34A HCl) at the top rank (rank ≤ 3 for enrichment score of sensitive compounds) were chosen as the candidate set (Fig. 4b). Among them, it has been proved that small cell lung cancer (SCLC) cells are exquisitely sensitive to growth inhibition by the BET inhibitor (+)-JQ-1 (CAS No. : 1268524-70-4)[40], and (+)-Fangchinoline and SEL120-34A HCl were assessed experimentally. We also explored the suitable activity concentrations of candidate compounds by calculating the enrichment scores of perturbed transcriptional profiles at concentration gradients. As shown in Fig. 4c, d, concentrations in the range of 1–10 μmol/L might be the proper inhibitory concentration for these candidate compounds.

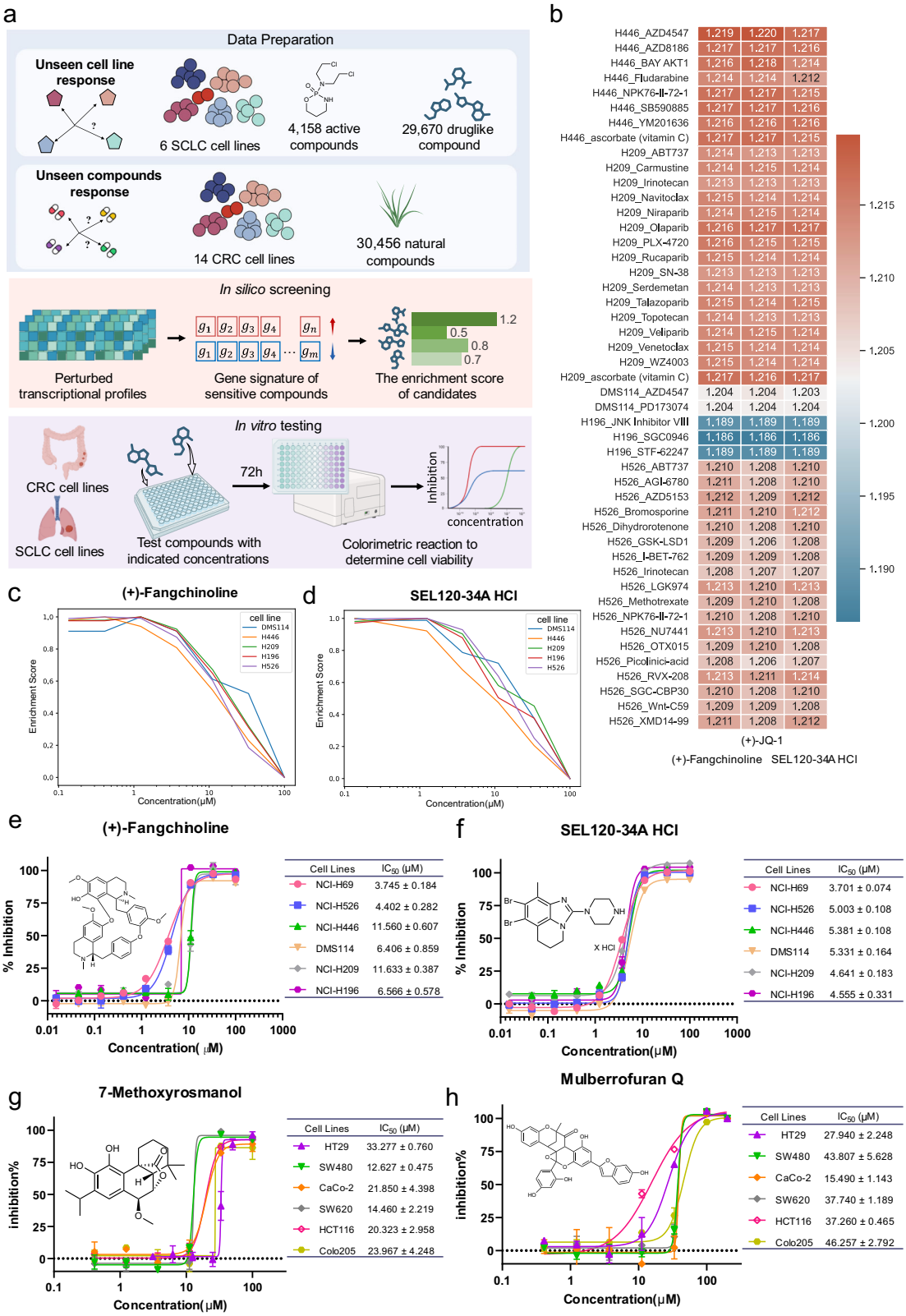
We utilized MTT assays to examine the activities of compound candidates against SCLC cells. Six human SCLC cell lines (NCI-H69, NCI-H526, NCI-H446, NCI-H209, and NCI-H196, and DMS114) were chosen for experiments. Results revealed the activity of SEL120-34A HCl (CAS No. : 1609452-30-3) and (+)-Fangchinoline (CAS No. : 436-77-1) against small cell lung cancer (SCLC) cell lines. SEL120-34A HCl and (+)-Fangchinoline showed significant inhibitory effects on the proliferation of SCLC cells, and exhibited an $IC_{50}$ (half-maximal inhibitory concentration) of less than 10 $\mu mol/L$, indicating their inhibitory effect on SCLC cell viability (Fig. 4e, f). (+)-Fangchinoline and SEL120-34A HCl moderately inhibited the viability of SCLC cell lines. Detailed anti-viability activities of two candidates are shown in Supplementary Table 3. These findings suggested the potential therapeutic efficacy of SEL120-34A HCl and (+)-Fangchinoline in the context of SCLC treatment, highlighting their promising role as active compounds against this aggressive lung cancer subtype.

## PRnet found natural compounds against colorectal cancer

Colorectal cancer (CRC) ranks as the third most common cancer and the second leading cause of cancer death worldwide[41]. Advances in molecularly targeted therapy and immunotherapy over the past decades have significantly improved patient survival[41]. However, some CRC patients, initially responsive to these treatments, quickly develop insensitivity. The emergence of drug resistance in cancer treatment significantly reduces patient outcomes. We aim to explore more potential novel treatment options for colorectal cancer to mitigate the impact of drug resistance through in silico screening of a large-scale library of natural compounds (30,456 compounds[42]).

We extended the application of PRnet to screening novel natural compounds for the treatment of CRC. We first leveraged PRnet to predict the post-perturbation transcriptional profiles of sensitive compounds in 14 CRC cell lines (HT29, LOVO, MDST8, RKO, HT115, SW948, SNU1040, SNUC4, SNUC5, SW480, SW620, HCT116, CL34, and NCIH508). Then, we used the predicted up/downregulated genes of sensitive compounds on their cell lines as the GSEA gene signature. After that, we also in silico screened a large-scale natural compounds library containing 30,456 natural ingredients. PRnet predicted the transcriptional responses of each compound across eight concentration gradients perturbing 14 cell lines with three repeats and calculated the post-perturbation average gene rank of compounds. Two natural

https://doi.org/10.1038/s41467-024-53457-1




compounds (7-Methoxyrosmanol and Mulberrofuran Q) at the top rank (rank ≤ 5 for enrichment score of sensitive compounds) were chosen as the candidate set (see Supplementary Tables. 1, 2).

We utilized MTT assays to examine the activities of candidate compounds against CRC cells. Six human CRC cell lines, namely HT29, SW480, Caco-2, SW620, HCT116, and Colo205, were chosen for the experiments. The results revealed that 7-Methoxyrosmanol (CAS No. : 113085-62-4) and Mulberrofuran Q (CAS No. : 101383-35-1) inhibited the viability of CRC cell lines (Fig. 4e, f). 7-Methoxyrosmanol and Mulberrofuran Q moderately inhibited the viability of CRC cell lines. Detailed antiviability activities of two candidates are shown in Supplementary Table 3. These findings showed the activity of 7-Methoxyrosmanol and Mulberrofuran Q against CRC cell lines, suggesting their potential efficacy in CRC treatment.

**Fig. 4 | The PRnet-identified candidates against small cell lung cancer and colorectal cancer. a** PRnet predicted the perturbed transcriptional profiles of 6 SCLC cell lines and 14 CRC cell lines. For in silico screening, PRnet first predicted the transcriptional profiles of SCLC cell lines perturbed by a multi-concentration gradient of 4158 active compounds, as well as 29,670 in-house druglike compounds. PRnet also predicted the transcriptional profiles of 14 CRC cell lines perturbed by 30,456 natural compounds. Then, post-perturbation fold-changes and average fold-changes of compounds across cell lines are computed. The gene ranking is performed based on the fold-change values. Given the predicted gene signature of sensitive compounds, the model computes the enrichment scores for up- and down-regulated gene sets of screening compounds. Finally, compounds are ranked based on the enrichment scores. For in vitro testing, MTT assays were performed for the evaluation of cell viability. **b** The heatmap illustrates the enrichment score of candidates for sensitive compounds. **c**, **d** The line charts plot the enrichment score of SCLC cells exposed to (+)-Fangchinoline and SEL120-34A HCl. **e**, **f** The cell survival curve of SCLC cells exposed to (+)-Fangchinoline and SEL120-34A HCl ($n = 3$ replicates). $IC_{50}$ (half-maximal inhibitory concentration) are presented as mean values ± SD. **g**, **h** The cell survival curve of CRC cells exposed to 7-Methoxyrosmanol and Mulberrofuran Q ($n = 3$ replicates). $IC_{50}$ are presented as mean values ± SD. Source data are provided as a Source Data file. Some icons were created in BioRender. Qi, X. (2024) biorender.com/c85y849.

## PRnet generated a large-scale integration atlas of perturbation profiles

With the ability to characterize specific gene-level perturbation responses and identify anti-cancer compounds, PRnet was applied to in silico screen novel compound libraries and cell lines and generate a large-scale integration atlas of perturbation profiles across various scenarios (Fig. 5a). PRnet was trained with two datasets: (1) the L1000 dataset, a bulk high-throughput screening library consisting of 883,269 transcriptional profiles from 82 cell lines perturbed by 175,549 biologically active compounds, and (2) the Sci-plex3 dataset, a single-cell high-throughput screening library consisting of 290,888 transcriptional profiles from 3 cell lines perturbed by 188 active compounds. The L1000 dataset[1] screened cell lines derived from over 20 diverse tissues and exposed to compounds targeting multiple genes and pathways (Supplementary Fig. 5a–c). The Sci-plex3 dataset screened three cancer cell lines treated by 188 compounds targeting a diverse range of targets and molecular pathways, covering various mechanisms of action (Supplementary Fig. 5d–f). After training, PRnet was applied to screen various perturbation scenarios to generate a large-scale integration atlas of perturbation profiles. Through virtual screening, PRnet has predicted over 25 million post-perturbation expression profile atlas of perturbation profiles which consists of five parts: (1) the FDA-approved drugs dataset: a bulk virtual high-throughput screening library containing 1,891,330 transcriptional profiles from 82 cell lines perturbed by 935 FDA-approved drugs, (2) the anti-cancer compounds dataset: a bulk virtual high-throughput screening library containing 8,781,784 transcriptional profiles from 88 cell lines perturbed by 4158 active compounds, (3) the natural compounds dataset: a bulk virtual high-throughput screening library containing 10,233,230 transcriptional profiles from 14 colorectal cancer cell lines perturbed by 30,456 natural compounds, (4) the bioactive compounds dataset: a bulk virtual high-throughput screening library containing 4,272,486 transcriptional profiles from 6 small cell lung cancer cell lines perturbed by 29,670 druglike compounds, and (5) the Gtex dataset: a bulk virtual high-throughput screening library containing 1,245,510 transcriptional profiles from 54 tissues perturbed by 935 FDA-approved drugs. Details of all cell lines are provided in Supplementary Data 3. PRnet offered a broad perspective by providing insights into how perturbations impact the transcriptional landscape on a large scale and extended its utility to diverse screening contexts. The large-scale integrated atlas of perturbation profiles can be applied to a variety of downstream application scenarios. For example, the FDA-approved drug dataset can be used for drug repositioning to recommend drugs for specific diseases based on gene signatures (see PRnet provided a robust and scalable drug 568 recommendation workflow based on the profiles atlas). The anti-cancer compounds dataset, the natural compounds dataset, and the bioactive compounds dataset are valuable for screening new anti-cancer compounds (see PRnet identified active compounds against small cell 430 lung cancer and PRnet found natural compounds against colorectal 482 cancer). In addition, the Gtex dataset can be useful to analyze the toxicity of compounds in different tissues. PRnet imports gene-level functionality in perturbations of different compounds and empowers flexibility by utilizing user-defined compounds' structures and transcription profiles to estimate the gene expression matrix. These profiles can be compared against various perturbation conditions(dosages, structures of compounds) using PRnet, to evaluate the impact of using different compounds on specific gene-expression profiles from single cell and bulk data. These diverse profiles provided potential solutions for drug discovery, disease treatment, and toxicity analysis.

## PRnet provided a robust and scalable drug recommendation workflow based on the profiles atlas

PRnet provides a comprehensive drug recommendation workflow based on the large-scale integration atlas of perturbation profiles (Fig. 5b). In step 1, given the compounds' structure of the screening library, PRnet predicts the post-perturbed transcriptional profiles of all compounds across multiple concentration gradients across cell lines. In step 2, the transcriptional profiles of 978 landmark genes are transformed into 12,328 genes through linear transformation. Subsequently, the perturbation fold-change and the average fold-change across cell lines for the transformed expression profile are computed. The gene ranking is then performed based on the fold-change values. In step 3, given the gene signature for a specific disease, PRnet computes the enrichment scores for up- and down-regulated gene sets of screening compounds. Finally, compounds in the screening library are ranked based on these enrichment scores. These downstream applications demonstrated the versatility and utility of PRnet in addressing diverse challenges in the field of drug discovery and perturbation analysis.

We leveraged the drug recommendation workflow to identify drug candidates for the treatment of 233 different diseases. For the user-defined compound library, we here chose the FDA-approved drug dataset, which comprises 935 FDA-approved drugs. All gene signatures of 233 diseases versus normal were collected from the CREEDS project (CRowd Extracted Expression of Differential Signatures[43]). The up/downregulated genes of a specific disease were used as the GSEA gene signature input to calculate enrichment scores (see "Methods") of drugs. Finally, we obtained 577 candidate lists from 577 studies for 233 unique diseases (Supplementary Data 4). Taking three metabolic disorder diseases, including nonalcoholic steatohepatitis (NASH), inflammatory bowel disease (IBD), and polycystic ovary syndrome patients (PCOS) as examples, we provided enrichment scores of each drug for the three diseases and the recommended drug candidates for them. The enrichment scores are plotted in Fig. 5c–e, drugs positioned at the upper right corner (rank≤ 10 for enrichment score) were chosen as the candidate set for literature verification. Supplementary Table 4 lists the Top 10 enrichment scores for compounds against NASH, Crohn's disease, and PCOS.

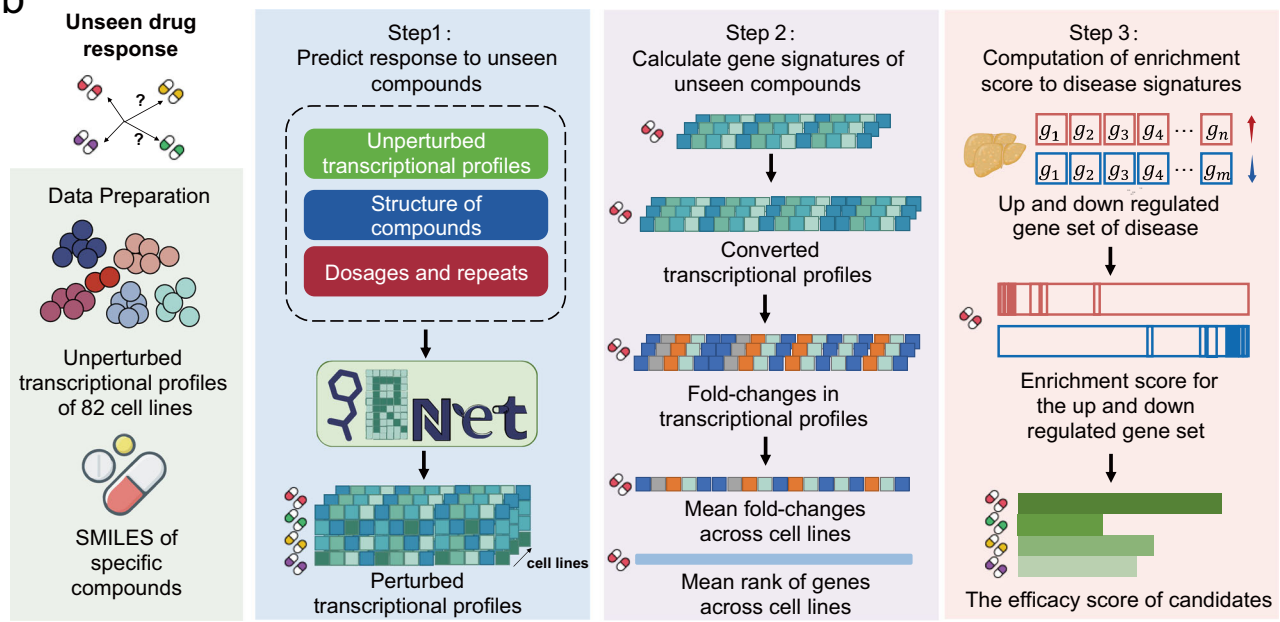
Nonalcoholic steatohepatitis (NASH) is liver inflammation and damage caused by a buildup of fat in the liver without standard treatment and any well-established drug targets. After calculation, we ultimately selected a set of candidate drugs from the upper right corner (rank≤ 6 for enrichment score) as candidate treatments for NASH (Fig. 5c). Literature verification revealed that Mirabegron (rank 1), Vidofludimus (rank 5), and Rifaximin (rank 6) have literature

https://doi.org/10.1038/s41467-024-53457-1

a

| Dataset | Compounds number | Cell line/tissue number | Profile number | Category |
|---|---|---|---|---|
| The L1000 dataset | 175,549 | 82 Cell lines | 883,269 | Bulk |
| The Sci-plex3 dataset | 188 | 3 Cell lines | 290,888 | Single cell |
| The FDA-approved drugs dataset | 935 | 82 Cell lines | 1,891,330 | Bulk |
| The anti-cancer compounds dataset | 4,158 | 88 Cell lines | 8,781,784 | Bulk |
| The natural compounds dataset | 30,456 | 14 Cell lines | 10,233,230 | Bulk |
| The drug-like compounds dataset | 29,670 | 6 Cell lines | 4,272,486 | Bulk |
| The Gtex dataset | 935 | 54 tissues | 1,245,510 | Bulk |

b



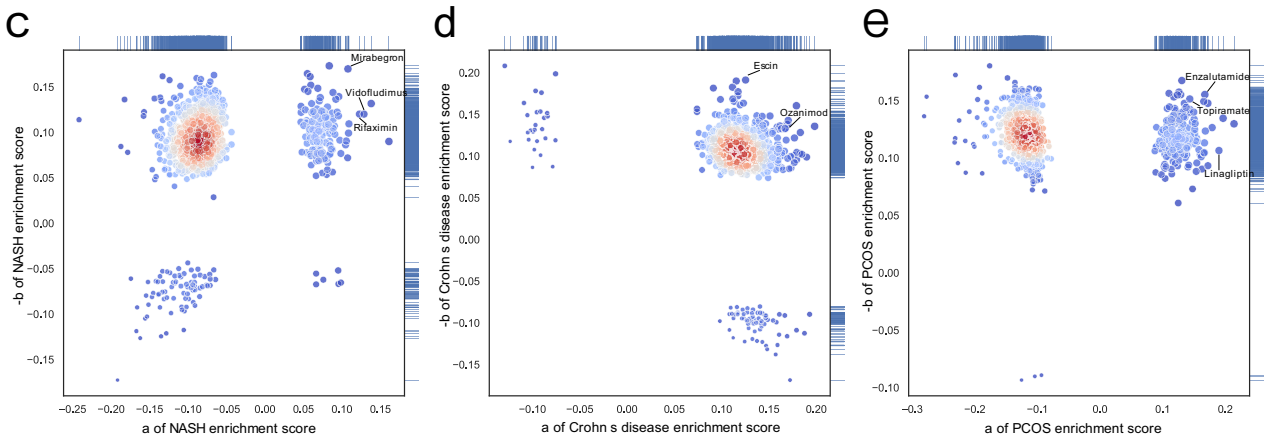




**Fig. 5 | PRnet provided a robust and scalable recommend workflow on a large integration atlas of perturbation profile. a** The large-scale integration atlas of perturbation profiles. **b** PRnet provided a recommended workflow for specific diseases. In step 1, given the structures of the screening library, PRnet predicts the post-perturbed transcriptional profiles of all compounds across multiple concentration gradients in 82 cell lines. In step 2, the transcriptional profiles of 978 landmark genes are transformed into 12,328 genes through linear transformation. Subsequently, the perturbation fold-change and the average fold-change across cell lines for the transformed expression profile are computed. Gene ranking is then performed based on the fold-change values. In step 3, given the gene signature for a particular disease, the model computes the enrichment scores for up- and down-regulated gene sets of screening compounds. Finally, compounds are ranked based on the enrichment scores. **c**–**e** Scatter plots of the enrichment score of compounds against NASH, Crohn's disease, and PCOS. The computationally predicted candidates are highlighted with their names in the upper right corner. The color gradient shows the density of the dots. Source data are provided as a Source Data file. Some icons were created in BioRender. Qi, X. (2024) biorender.com/z56i777.

support for use in the treatment of NASH. A study on high-fat diet rats[44] suggested that mirabegron may have a protective effect against NASH as it improves liver enzymes, lipids, serum HbA1c, fasting insulin and glucose, insulin resistance index, and serum adiponectin levels, as well as ameliorates hepatic histopathologic changes in NASH-induced rats. A study on obesity mice[45] suggested vidofludimus reduced hepatic steatosis and inflammation in obesity mice and has been repurposed to a therapeutic potential in the treatment of NASH by targeting FXR based on the newly established relationships among drugs, targets, and diseases. Rifaximin therapy appeared to be effective and safe in modifying NASH through reduction of serum endotoxin and improvement of insulin resistance, proinflammatory cytokines, CK-18, and NAFLD-liver fat score. Patients with biopsy-proven NASH with Rifaximin therapy showed that Rifaximin appeared to be effective and safe in modifying NASH[46,47]. We also performed KEGG pathway enrichment analysis of predicted up/down-regulated genes of Mirabegron, Vidofludimus, and Rifaximin (Supplementary Fig. 7). The KEGG pathway enrichment analysis of downregulated genes for both Mirabegron and Rifaximin suggested they may downregulate pathways associated with NASH.

Crohn's disease and ulcerative colitis are idiopathic inflammatory bowel disorders (IBD). Crohn's disease is a relapsing systemic inflammatory disease that primarily affects the gastrointestinal tract with extraintestinal manifestations and associated immune disorders[48]. After ranking the drugs, two drugs positioned at the upper right corner (rank≤ 9 for enrichment score) were chosen as the candidate set (Fig. 5d). The literature verification revealed that Escin (rank 3) and Ozanimod (rank 9) have literature support for use in the treatment of Crohn's disease. Escin is the main bioactive ingredient of Semen aesculi, which improves intestinal barrier dysfunction of IBD via Akt/NF-$\kappa$B signaling pathway[49]. Ozanimod is a once-daily sphingosine 1-phosphate receptor modulator for the treatment of inflammatory bowel disease and was approved by the FDA. Phase 2 clinical trial had proved that Ozanimod can be a novel oral small molecule therapy for the treatment of Crohn's disease[50,51]. The KEGG pathway enrichment analysis results of predicted up/down-regulated genes of Escin and Ozanimod are illustrated in Supplementary Fig. 7. The predicted upregulated pathways (Supplementary Fig. 7) suggested Escin may upregulate the PI3K/Akt signaling pathway which regulates a broad cascade of target proteins including nuclear factor kappa B (NF-$\kappa$B) and glycogen synthase kinase-3$\beta$ (GSK-3$\beta$)[52].

Polycystic ovary syndrome (PCOS) is a heterogeneous endocrine disorder in which the major endocrine disruption is excessive androgen secretion or activity, and a large proportion of women also have abnormal insulin activity. We also chose positively scored drugs as candidates for PCOS (rank≤ 9 for enrichment score, Fig. 5e). The literature verification revealed that Enzalutamide (rank 3), Linagliptin (rank 8), and Topiramate (rank 9) have literature support for use in the treatment of PCOS. Enzalutamide has been suggested for use as an antiandrogen to treat hirsutism and hyperandrogenism in women with polycystic ovary syndrome[53,54] investigated the effect of linagliptin and/or I3C on experimentally-induced PCOS in female rats, and the results suggested that linagliptin/I3C combination might represent a beneficial therapeutic modality for amelioration of PCOS. PCOS has completed phase 3 trials for Topiramate, and phentermine-topiramate extended-release (PHEN/TPM) resulted in the most loss of weight and total body fat[55]. The predicted regulated pathways (Supplementary Fig. 7) of Enzalutamide showed that Enzalutamide may upregulate the PI3K/Akt signaling pathway and MAPK signaling pathway, which are related to Androgen signaling[56]. The predicted downregulated pathways suggested Topiramate may downregulate the Lipid and atherosclerosis and the Fat digestion and absorption pathway, thereby contributing to weight loss. All three literature verification results illustrated that PRnet recommended reliable and effective drugs for different diseases, and hence is a valuable tool for drug discovery.

## Discussion

Recent advancements in high-throughput screening have profiled thousands of independent chemical perturbations, providing crucial insights into the fundamental responses of biological systems to perturbations. However, screening all possible disease and compound combinations is experimentally unfeasible. To address the limited exploration power of existing experimental methods, we developed PRnet, which supports the prediction of transcriptional responses to novel chemical perturbations that were never experimentally studied at bulk and single-cell levels. PRnet serves as a valuable tool that facilitates gene-level response interpretation in various novel perturbation scenarios and effectively recommends candidates for various diseases.

PRnet has the unique ability to screen various novel compound libraries for specific diseases and infer their post-perturbed transcriptional response. Given the structures of compounds in libraries and unperturbed transcriptional profile, PRnet encodes the perturbation and unperturbed state to an interpretable latent space as condition contexts to generate transcriptional responses. Trained on extensive data, PRnet is able to adapt to complex chemical perturbations of generalization to novel compounds and cell lines. To further validate the effectiveness of PRnet, we identified novel compounds candidates against SCLC and natural compounds against CRC. Experimental validations confirmed the activity of candidate compounds, showcased the capabilities of PRnet to guide the design of new screens, and reduced the time and costs of experiments. Lastly, PRnet generated a large-scale integration atlas of perturbation profiles and demonstrates its capability to recommend drug candidates for 233 complex diseases based on reference changes in gene sets. The flexibility and scalability of PRnet make it a valuable tool for guiding the design of screening strategies for gene-based therapeutics.

Although PRnet is effective in predicting drug candidates, there is still room for improvement. Due to the scalability of the SMILES format[30] and the RDKit[31], PRnet is able to encode unseen complex compounds as embeddings. SMILES is widely used for representing chemical structures due to its simplicity and efficiency in encoding complex molecules as strings. However, SMILES may not fully capture complex molecular features such as 3D geometry, conformational flexibility, or dynamic behavior, which can be important for understanding molecular interactions. Alternative encoding methods such as MOL/SDF files and graph-based representations are more flexible representations in capturing the 3D geometry structure of compounds. MOL/SDF formats represent chemical structures with explicit atom coordinates and bond connectivity. Graph-based representations use graph theory to represent molecules where atoms are nodes and bonds are edges. These methods can provide a more intuitive and flexible representation. However, they are more complex or require pre-training to obtain better representations. In the scenario of large-scale in silico screening, we chose SMILES to encode chemical structures for its simplicity and scalability. Alternative encoding methods, such as MOL/SDF files and graph-based representations, will be considered in the future work.

The reverse signature paradigm to connect disease and drug established by Lamb et al in CMap[15] has demonstrated its effectiveness in previous studies[16–20]. However, this may not hold true for all diseases, and there are instances where transcriptional changes do not correlate directly with drug sensitivity. Jie Cheng et al.[57] found the reverse signature paradigm may perform poor accuracy when some disease signatures are of low quality. Rasool Bhat[58] discusses how cancer cells exhibit phenotypic plasticity under drug treatment, leading to drug resistance which often involves complex regulatory networks that are not fully captured by changes in gene expression alone. The reverse signature paradigm is an effective matching algorithm in many diseases. But there are also cases that not fit this algorithm, more comprehensive omics data, such as genomics, epigenetics, and

proteomics should be considered in the future work to better characterize disease states and facilitate drug discovery.

In addition, in the process of compound screening, we mainly focus on the impact at the gene level, lacking consideration for the effects on phenotypes, such as the Area Under the Curve (AUC) or half-maximal inhibitory concentration ($IC_{50}$) derived from the experimental dose-response curve. To comprehensively assess compound impacts, future research directions could involve incorporating more phenotype data for a holistic assessment of the relationship between genes and phenotypes. Moreover, expanding the scope of perturbation scenarios and incorporating extensive biological knowledge will be considered to enhance the model's predictive capabilities in the future. Furthermore, we expect PRnet to extend its utility beyond the prediction of chemical perturbations to encompass various perturbation experiments, including genetic perturbations and other forms, thereby contributing to the advancement of drug discovery.

## Methods

### The PRnet algorithm

In this work, we formulate transcriptional response prediction as a distribution generation problem conditioned on perturbations. Given a dataset $D=(x_i,P_i)_{i=1}^N$, where $x_i \in \mathbb{R}^n$ denotes as the $n$-dimensional gene expression and $P_i$ is the attribute set of perturbation, PRnet was trained to learn a function $f$ that predicts transcriptional responses $\widehat{x}_i=f(x_i^u,P_i)$ to novel perturbations. Here, $x_i^u$ represents the unperturbed gene expression, $\widehat{x}_i$ is the predicted perturbed gene expression, and $P_i=(s_i,d_i)$ is the attribute set of perturbation, which includes the chemical structures $s_i$ and dosages of the compounds $d_i$. As a generation model, the Gaussian negative log-likelihood loss is chosen as the loss function. For HTS RNA-seq datasets $D=(x_i,P_i)_{i=1}^N=(x_i,(s_i,d_i))_{i=1}^N$, the gene expression $x_i$, compounds $s_i$, and dosages $d_i$ attributes are usually considered, in which $s_i=(s_{i,1},s_{i,2},\ldots,s_{i,M})$ describes the Canonical SMILES format of $M$ compounds of perturbation $i$ and $d_i=(d_{i,1},d_{i,2},\ldots,d_{i,M})$ are the dosages of compounds. When $d_{i,j}=0$, the compound $j$ was not applied in perturbation $i$. If $M=0$, there is no perturbation performed, and in this case, perturbation $i$ is an unperturbed state $x_i^u$, otherwise is in a perturbed state. In addition, the covariates vector $c_i$ contains discrete covariates such as cell types, cell lines, or species, depending on the available data. While $c_i$ does not directly serve as input to the function $f$, it relates to the intermediate variable vector $x_i^u$ of the function $f$. Each perturbed profile $x_i$ is assigned an unperturbed transcriptional profile of the same cell line $x_i^u$ by random selection in the dataset. The goal of PRnet is to learn a function $f$ that predicts transcriptional responses $\widehat{x}_i=f(x_i^u,P_i)$ to novel perturbations. Given an unperturbed gene expression $x_i^u$, PRnet maps the unperturbed state to a distribution $\mathcal{N}(x_i\,|\,\mu_i,\sigma_i^2)$, and the sample to a perturbed state $\widehat{x}_i$.

PRnet is a perturbation-conditioned generative model aimed at predicting gene expression profiles under different perturbations. The design of PRnet consists of three components: (1) the Perturb-adapter, a scalable adapter $f_{pert}:\mathbb{Z}\to\mathbb{R}^k$ encodes complex perturbations $(s_i,d_i)$ into a $k$-dimensional perturbation embedding $z_i^p$, (2) the Perturb-encoder $f_{enc}:\mathbb{R}^{k+n}\to\mathbb{R}^e$ encodes the combined learnable perturbation embedding $z_i^p$ and unperturbed gene expression $x_i^u$ to a latent space to get the embedding $z_i^l$, and (3) the Perturb-decoder $f_{dec}:\mathbb{R}^{e+k+m}\to\mathbb{R}^{2n}$ decodes combines perturbation embedding $z_i^p$, latent state $z_i^l$ and noise $z_i^n$ into Gaussian likelihood distribution $\mathcal{N}(x_i\,|\,\mu_i,\sigma_i^2)$ of perturbed state and generates perturbed gene expression $\widehat{x}_i$ by sampling from $\mathcal{N}(x_i\,|\,\mu_i,\sigma_i^2)$. We elaborate on the three components in the following.

### The Perturb-adapter

The Perturb-adapter encodes the $i$-th perturbation $P_i$ as a fixed-size embedding $z_i^p\in R^k$ ($k=64$). Given the $i$-th perturbation $P_i$, the $j$-th compound of $P_i$ was first represented as the canonical SMILES format $s_{i,j}$, and was converted to a fixed size fingerprint embedding $FCFP_{i,j}$ (FCFP4 fingerprints: Functional-Class Fingerprints with $radius=2$ which focus on the functional topological pharmacophore) by RDKit[31] encoder $H:\mathbb{Z}\to\mathbb{R}^h$ ($h=1024$). Then, the rescaled Functional-Class Fingerprints $rFCFP_i$ embedding of $P_i$ was calculated by the weighted sum of all fingerprint embeddings of all compounds applied by $P_i$:

$$rFCFP_i=\sum_j \phi(d_{i,j})\times H(s_{i,j})=\sum_j log_{10}(d_{i,j}+1)\times H(s_{i,j}), \tag{1}$$

where $\phi$ is a log scale function using the dosage of $j$-th compound. At last, the final perturbation embedding $z_i^p$ was generated by the Perturb-adapter $E_{\theta_p}$:

$$z_i^p=f_{pert}(s_i,d_i)=E_{\theta_p}(G(H(s_i),\log_{10}(d_i+1))), \tag{2}$$

where $\theta_p$ are the parameters of the Perturb-adapter, which is a 2-layer feedforward neural network.

Thanks to the scalability of the SMILES[30] format and the RDKit[31], which can encode any compound to a latent embedding, the design of the Perturb-adapter is scalable to in silico screen novel compounds and cell line perturbations.

### The Perturb-encoder and the Perturb-decoder

Inspired by the Variational Auto-Encoder (VAE[22]) framework, PRnet uses the encoder and decoder framework to estimate the Gaussian distribution of each gene parameterized with the mean ($\mu_i$) and the variance ($\sigma_i^2$). Due to the technical limitation, only unpaired perturbed and unperturbed transcriptional profiles were observed. To match the expression profiles before and after perturbation, for scRNA-seq datasets, each perturbed cell was assigned with an unperturbed cell from the same cell type by random selection in the dataset, and for bulk RNA-seq datasets, the same cell line observation was assigned for the perturbed state. Given the perturbation embedding $z_i^p$ generated by the Perturb-apaptor, the Perturb-encoder $E_{\theta_e}$ mapped the $z_i^p$ and the unperturbed gene expression $x_i^u$ to a latent embedding $z_i^l\in R^e$ ($e=64$). Then, the perturb-decoder $E_{\theta_d}$ accepted $z_i^p$, $z_i^l$, and the noise $z_i^n$ to estimate the Gaussian likelihood distribution $\mathcal{N}(x_i\,|\,\mu_i,\sigma_i^2)$ of the perturbed profile:

$$z_i^l=E_{\theta_e}(x_i^u,z_i^p), \tag{3}$$

$$[\widehat{\mu}_i,\widehat{\sigma}_i^2]=E_{\theta_d}(z_i^p,z_i^l,z_i^n), \tag{4}$$

where both $E_{\theta_e}$ and $E_{\theta_d}$ are 2-layer feedforward neural networks, and $\theta_e$and $\theta_d$ are the parameters of $E_{\theta_e}$ and $E_{\theta_d}$, respectively. In the design of PRnet, the perturbation embedding $z_i^p$ was used as the input for both the perturb-encoder and the perturb-decoder. Multiplexing $z_i^p$ in the perturb-decoder combined the chemical and biology context helped assemble a more precise output. And the noise $z_i^n$ increases the robustness of the model. Once the estimated Gaussian likelihood distribution $\mathcal{N}(x_i\,|\,\mu_i,\sigma_i^2)$ was generated, PRnet sampled the estimated perturbed gene expression $\widehat{x}_i$ from it:

$$\mu_i=\widehat{\mu}_i, \tag{5}$$

$$\sigma_i^2=\varphi(\widehat{\sigma}_i^2), \tag{6}$$

$$\widehat{x}_i\sim\mathcal{N}(\mu_i,\sigma_i^2), \tag{7}$$

where $\varphi(.)$ denotes the non-linear softplus function. $\mu_i$ and $\sigma_i^2$ are the mean and the variance for each gene, which are $n$-dimensional vectors.

**Table 1 | Hyperparameters set of PRnet**

| Module | Hyperparameter | Default value | hyperparameters set |
|---|---|---|---|
| PRnet | batch size | 512 | 256, 512, 1024 |
| | learning rate | 1e-3 | 1e-3, 1e-5, 3e-3 |
| | weight decay | 1e-8 | 0,1e-3 |
| | scheduler factor | 0.5 | 0.5 |
| | scheduler patience | 5 | 5 |
| | early stopping patience | 20 | 20 |
| | num of epochs | 100 | 100 |
| | gradient descent | Adam | Adam,RMSprop |
| | loss function | GaussNLLLoss[1] | GaussNLLLoss, NBLoss[2], MSE[3], KLLoss[4] |
| Perturbe-adapter | drug dimension($h$) | 1024 | 1024 |
| | input size | $h$ | 1024, 2048 |
| | hidden size | 128 | 64, 128, 256 |
| | output size($k$) | 64 | 64, 128 |
| Perturbe-encoder | input size | 5000, 978 | 5000(scRNA-seq), 978(bulk RNA-seq) |
| | hidden size | 128 | 128, 256 |
| | output size($e$) | 64 | 64, 128 |
| Perturbe-decoder | input size($e+k+m$) | 138 | 138, 266 |
| | hidden size | 128 | 128, 256 |
| | output size | $n\times 2$ | 10000, 1956 |

[1] GaussNLLLoss is the Gaussian negative log-likelihood loss.
[2] NBLoss is the negative binomial negative log-likelihood loss.
[3] MSE is the mean squared error loss.
[4] KLLoss is the Kullback-Leibler divergence loss.

And $\hat{x}_i$ is the estimated $n$-dimensional transcriptional responses to perturbations.

## Training and testing

The training objective of PRnet is to minimize the Gaussian negative log-likelihood loss defined as:

$$\mathcal{L}(\theta_p,\theta_e,\theta_d)=\sum_i\frac{1}{2}\left(\log\left(\max\left(\sigma_i^2,eps\right)\right)+\frac{(\mu_i-x_i)^2}{\max\left(\sigma_i^2,eps\right)}\right), \quad (8)$$

Where $eps$ is used to clamp $\sigma_i^2$ for stability, and is set to 1e-6 by default. By minimizing the loss of all cells/samples, PRnet is able to learn the most proper parameters.

The training datasets (a bulk HTS dataset from the L1000 project[1] and a single cell HTS dataset from the sci-Plex3 assay[2]) were split by dividing the attribute set of perturb $P_i=(s_i, d_i, c_i)$.

To train and test PRnet, all datasets were split into three subsets: train, valid, and test in a ratio of 6:2:2, according to different split strategies. A total of four split strategies were designed:

- random_split: randomly divides compounds and cell lines
- compound_split: groups the datasets according to $s_i$ for each profile, and then splits groups of the dataset
- cell_line_split: groups the datasets according to $c_i$ for each profile, and then splits groups of the dataset
- pathway_split: groups the datasets according to $c_i$ from the same pathway for each profile, and then splits groups of the dataset

For bulk RNA-seq datasets, random_split, compounds_split, and cell_line_split were applied, and for scRNA-seq datasets, random_split, compound_split, and pathway_split were applied. Strict data split can prevent data leakage and bring better generalization performance. PRnet is trained using 5-fold cross-validation for each split category.

The table below (see Table 1) outlines the values for the hyperparameters involved in PRnet training. The same set of hyperparameters is used across all splits and datasets.

## Model evaluation

Four metrics were used to quantitatively evaluate the performance of PRnet:

1. *fold − change*: describes the ratio changes between the gene expression of unperturbed and perturbed state:

$$fc_i=\frac{x_i+1}{x_i^u}, \quad (9)$$

$$\hat{fc}_i=\frac{\hat{x}_i+1}{\hat{x}_i^u}. \quad (10)$$

Where $fc_i$ and $\hat{fc}_i$ are the ground truth and predicted perturbed fold change in gene expression $x_i$, respectively. A higher value of *fold − change* indicates a better performance.

2. $R^2$: is the mean coefficient of determination score between the predicted and true perturbed gene expression of all cells:

$$R^2=\frac{1}{O}\sum_{i=1}^{O}\left(1-\frac{\sum_{j=1}^{j=n}(\hat{x}_i^j-x_i^j)^2}{\sum_{j=1}^{j=n}(\hat{x}_i^j-\bar{x}_i^j)^2}\right), \quad (11)$$

where $O$ denotes the number of cells in test datasets, $\hat{x}_i$, $x_i$ and $\bar{x}_i$ are the predicted, true perturbed gene expression and mean of the ground truth perturbed gene expression, respectively. A higher value of $R^2$ indicates a better performance.

3. Pearson of log(FC) in compounds: evaluates the Pearson correlation between the true and predicted post-perturbation of the average logarithm of the fold-change in gene expression (log(FC)) perturbed by the same compound:

$$\overline{\log(fc_j)}=\frac{1}{n}\sum_{i=1}^{n}\log(fc_{i,j}), \quad (12)$$

$$\overline{\log(\hat{fc}_j)}=\frac{1}{n}\sum_{i=1}^{n}\log(\hat{fc}_{i,j}), \quad (13)$$

$$PCC_j=PCC(\overline{\log(fc_j)},\overline{\log(\hat{fc}_j)}), \quad (14)$$

$$\text{Pearson of log(FC) in compounds}=\frac{1}{m}\sum_{j=1}^{m}PCC_j, \quad (15)$$

where $fc_{i,j}$ and $\hat{fc}_{i,j}$ denote as the ground truth and predicted post-perturbation fold change in gene expression $x_i$ perturbed by compound $j$. The set $(fc_{i,j})_{i=1}^{n}$ represents all perturbed fold-changes for compound $j$. $\overline{\log(fc_j)}$ and $\overline{\log(\hat{fc}_j)}$ are the ground truth and predicted post-perturbation average logarithm of logarithm of the fold-change in gene expression for compound $j$, respectively. $PCC_j$ is the Pearson correlation coefficient of the mean log(FC) perturbed by compound $j$. The average of the Pearson correlations for all $m$ compounds in the test set is referred to as the "Pearson of log(FC) in compounds". A higher value of "Pearson of log(FC) in compounds" indicates a better performance.

4. Pearson of log(FC) in cov_compounds: evaluates the Pearson correlation between the true and predicted post-perturbation of the average logarithm of the fold-change in gene expression (log(FC)) perturbed by the same compound within the same cell line:

$$\overline{\log(FC_j^c)} = \frac{1}{n}\sum_{i=1}^{n}\log(fc_{i,j}^c), \tag{16}$$

$$\overline{\log(\widehat{FC_j^c})} = \frac{1}{n}\sum_{i=1}^{n}\log(\widehat{fc_{i,j}^c}), \tag{17}$$

$$PCC_j^c = PCC(\overline{\log(FC_j^c)}, \overline{\log(\widehat{FC_j^c})}), \tag{18}$$

$$\text{Pearson of log(FC) in cov\_compounds} = \frac{1}{z}\sum_{c,j=1}^{z} PCC_j^c, \tag{19}$$

where $fc_{i,j}^c$ and $\widehat{fc_{i,j}^c}$ denote the ground truth and predicted post-perturbation fold change in gene expression $x_i$ perturbed by compound $j$ in covariate $c$, respectively. The covariate $c$ represents cell line or cell type of $x_i$. $(fc_{i,j}^c)_{i=1}^n$ represents all perturbed fold-changes for compound $j$ in covariate $c$. $\overline{\log(FC_j^c)}$ and $\overline{\log(\widehat{FC_j^c})}$ are the ground truth and predicted post-perturbation average logarithm of logarithm of the fold-change in gene expression for compound $j$ in covariate $c$, respectively. $PCC_j^c$ is the Pearson correlation coefficient of the mean log(FC) perturbed by compound $j$ in covariate $c$. The average of the Pearson correlations for all $z$ "cov_compounds" conditions in the test set is referred to as the "Pearson of log(FC) in cov_compounds". A higher value of "Pearson of log(FC) in cov_compounds" indicates a better performance.

5. $R^2$ in compounds: is the mean coefficient of determination score between the predicted and true post-perturbation gene expression of the same compound:

$$\overline{x_t} = \frac{1}{p}\sum_{i=1}^{p} x_{i,t}, \tag{20}$$

$$\overline{\widehat{x_t}} = \frac{1}{p}\sum_{i=1}^{p} \widehat{x_{i,t}}, \tag{21}$$

$$R_t^2 = R^2(\overline{x_t}, \overline{\widehat{x_t}}), \tag{22}$$

$$R^2\text{in compounds} = \frac{1}{T}\sum_{j=1}^{T} R_t^2, \tag{23}$$

where $x_{i,t}$ and $\widehat{x_{i,t}}$ denote as the ground truth and predicted post-perturbation gene expression $x_i$ perturbed by compound $t$. The set $(x_{i,t})_{i=1}^p$ represents all post-perturbation gene expression for compound $t$. $\overline{x_t}$ and $\overline{\widehat{x_t}}$ are the ground truth and predicted post-perturbation average gene expression for compound $t$, respectively. $R_t^2$ is the coefficient of determination score of the mean post-perturbation gene expression of compounds $t$. The average of the coefficient of determination score for all $T$ compounds in the test set is referred to as the "$R^2$ in compounds". A higher value of "$R^2$ in compounds" indicates a better performance.

6. $R^2$ in cov_compounds: is the mean coefficient of determination score between the predicted and true post-perturbation gene expression of the same compound within each specific cell line:

$$\overline{x_t^c} = \frac{1}{p}\sum_{i=1}^{p} x_{i,t}^c, \tag{24}$$

$$\overline{\widehat{x_t^c}} = \frac{1}{p}\sum_{i=1}^{p} \widehat{x_{i,t}^c}, \tag{25}$$

$$R_t^2 = R^2(\overline{x_t^c}, \overline{\widehat{x_t^c}}), \tag{26}$$

$$R^2 \text{ in cov\_compounds} = \frac{1}{Q}\sum_{j=1}^{Q} R^{2^c}_t, \tag{27}$$

where $x_{i,t}^c$ and $\widehat{x_{i,t}^c}$ denote as the ground truth and predicted post-perturbation gene expression $x_i$ perturbed by compound $t$ in covariate $c$, respectively. The covariate $c$ represents the cell line or cell type of $x_i$. The set $(x_{i,t}^c)_{i=1}^p$ represents all post-perturbation gene expression for compound $t$ in covariate $c$. $\overline{x_t^c}$ and $\overline{\widehat{x_t^c}}$ are the ground truth and predicted post-perturbation average gene expression for compound $t$ in covariate $c$, respectively. $R^{2^c}_t$ is the coefficient of determination score of the mean post-perturbation gene expression of compounds $t$ in covariate $c$. The average of the coefficient of determination score for all $Q$ "cov_compounds" conditions in the test set is referred to as the "$R^2$ in cov_compounds". A higher value of "$R^2$ in cov_compounds" indicates a better performance.

We use the following baseline models to compare model performance:

1. **Linear model**: This model uses a linear regression model to learn weights between all genes and perturbations. The linear model uses MSELoss to learn perturbation effects applied to the control sample/cell. Let $\theta_l$ represent the weight matrix of the linear model, the perturbed gene expression would be:

$$\widehat{x}_i = E_{\theta_l}(z_i^p, x_i^u), \tag{28}$$

$$\mathcal{L}_i(E_{\theta_l}) = MSELoss(\widehat{x}_i, x_i). \tag{29}$$

2. **MLP model**: MLP model utilizes Multilayer Perceptron (MLP) to fit the effect of perturbation to genes. The input is passed to an MLP model(Input Layer, Linear with output dimension 128, BatchNorm1d, LeakyReLU, Linear with output dimension as input size, and ReLU). Let $\theta_{MLP}$ represent the parameters of the linear model, the perturbed gene expression would be:

$$\widehat{x}_i = MLP_{E_{\theta_{MLP}}}(z_i^p, x_i^u), \tag{30}$$

$$\mathcal{L}_i(E_{\theta_{MLP}}) = MSELoss(\widehat{x}_i, x_i). \tag{31}$$

### The pseudo-dose trajectory

To calculate the pseudo-dose trajectory, we take the mean t-SNE embedding of cells with the same dosage as a point on the pseudo-dose trajectory:

$$\overline{z_d} = \frac{1}{n_d}\sum_{i=1}^{n_d} z_{i,d}, \tag{32}$$

where $z_{i,d}$ is the t-SNE latent embedding of cell $i$ at dose $d$, $n_d$ is the number of cells at dose $d$, and $\overline{z_d}$ is the mean t-SNE embedding for dose

$d$, representing a point on the pseudo-dose trajectory. The sequence of these mean latent embeddings $\overline{z_d}$ forms the pseudo-dose trajectory.

### Screening candidates based on gene signatures

Inspired by the CMap connectivity score[15] and L1000 project[1], PRnet employed a similar reverse signature paradigm to screen candidates based on gene signatures. PRnet provided a workflow for screening candidates for complex diseases according to the reference changes in gene sets.

Step 1: Given the SMILES of the screening compounds library and unperturbed transcriptional profiles of specific cell lines, PRnet predicts the perturbed transcriptional profiles of all compounds across multiple concentration gradients. This prediction is performed with three repeats to ensure computational robustness.

Step 2: The transcriptional profiles of 978 landmark genes are transformed into profiles of 12,328 genes using linear transformation vectors derived from the L1000 project[1]. The fold-change of transcriptional profiles is then calculated, which includes the average fold-change of each compound across cell lines. Genes are ranked based on their fold-change values. Gene expression signatures representing the disease states or sensitive compound responses are generated, known as query signatures. Query signatures include an up-regulated gene set and a down-regulated gene set which are the reversed signatures of the disease. Gene expression signatures of screening compounds are also generated, which are the ranked gene lists.

Step 3: The Kolmogorov-Smirnov test is employed to calculate enrichment scores, which measure the connectivity of screening compound profiles with the query signatures. The score to reverse disease up- and down-regulated features are calculated separately and then summed together:

$$a = \max_{m=1\sim p}\left[\frac{m}{p} - \frac{R(m)}{n}\right], \tag{33}$$

$$b = -\max_{m=1\sim q}\left[\frac{R(m-1)}{n} - \frac{m-1}{q}\right], \tag{34}$$

$$Score = \begin{cases} a-b, & when \quad a\times b<0, \\ 0, & when \quad a\times b>0, \end{cases} \tag{35}$$

where $p$ is the number of genes in the upregulated gene set of query signature, $q$ is the number of genes in the downregulated gene set of query signature, $n$ is the number of genes in the computed transcriptional profile, and $R(m)$ is the rank of a specific gene in the rank list. The enrichment score is commonly used in the field to evaluate the distribution of the predicted gene rank in the reference gene signature. The gene signature is an up- and down-regulated gene set. The up- and down-regulated gene set of the specific disease is identified from the reference study. The up- and down-regulated gene sets of the sensitive compounds to cell lines are computed by PRnet. Note that the up- and down-regulated gene sets represent the reverse gene signatures of a specific disease, indicating the up- and down-regulate gene sets that the compounds are expected to enrich. Finally, compounds are ranked based on these enrichment scores, and the top-ranked compounds are recommended for diseases.

### Other computational tools

Morgan fingerprints were calculated by the RDKit[31] in Python (2023.3.2). The atlas of perturbation profiles was organized into anndata format through the Scanpy package[59] in Python (1.9.1). T-SNE[33] clustering was performed by the Scanpy package in Python (1.9.1). Figures related to computation were plotted by matplotlib[60] (3.5.2), seaborn[61] in Python (0.12.0), and the ggplot[62] package in R (3.5.1). The GSEA and KEGG pathway enrichment analysis were performed by the Clusterprofiler package[63] in R (4.6.2), and plot by the gseaplot, enrichplot and cnetplot. A $P$-value $< 0.05$ was defined as the cutoff criterion. Experimental figures were plotted by GraphPad Prism. Some figures were plotted by Chiplot (https://www.chiplot.online/). And some icons in figures are created in BioRender (https://www.biorender.com/). The deep learning model was constructed by Pytorch[64] (1.12.1).

### Cell culture

All the cell lines used in this work were purchased from the American Type Culture Collection (ATCC). HT29, SW480, SW620, and HCT116 cells were cultured in DMEM (Gibco) medium supplemented with 10% fetal bovine serum, 100 *U/mL* penicillin, and 100 *U/mL* streptomycin. Colo205, NCI-H196, NCI-H209, NCI-H446, NCI-H526, and NCI-H69 cells were cultured in RPMI-1640 (Gibco) medium supplemented with 10% fetal bovine serum, 100 *U/mL* penicillin, and 100 μ/mL streptomycin. DMS114 cells were cultured in Waymouth's MB 752/1 (Gibco) medium supplemented with 10% fetal bovine serum, 100 μ/mLpenicillin, and 100 μ/mL streptomycin. All the cell lines were incubated at 37 °C in a humidified 5% $CO_2$ atmosphere. All cells were negative for mycoplasma, and these cell lines are not among those commonly misidentified by the International Cell Line Authentication Committee (ICLAC).

### Cell viability assays

MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) assay was performed for the evaluation of cell viability. The cells were seeded into 96-well plates at a density of 2000–20000 cells per well in full medium. Overnight, test compounds with indicated concentrations were added for 72 h. 20 μL of MTT (5 mg · $mL^{-1}$ in saline, Sigma) was then added per well for 2 h, and 50 μL SDS (20%, dissolved in $H_2O$ containing 1% HCl) was added per well, followed by incubating overnight. The absorbance at 570 nm was measured using a multiscan spectrum reader (BMG lab tech). The cell survival rate was calculated subsequently.

### Atlas of perturbation profile

**Data preprocessing.** The HTS RNA-seq datasets used for this study all underwent the same preprocessing. For single-cell RNA-seq data, the gene expression of each cell was normalized by total counts over all genes with log-transformation, and 5000 highly variable genes (HVGs) were selected using Scanpy. For bulk RNA-seq data, 978 different genes of level 3 were normalized with log transformation for training and evaluation. These data were split into training, validation, and test sets by dividing various perturbation conditions (such as compounds, cell type, and pathway) with a ratio of 6:2:2. And 5-fold cross-validation was applied for training.

**The L1000 dataset.** The L1000 project[2] contains more than 1 million bulk RNA-seq observations with 978 landmark genes of level 3. We performed data cleaning using the following criteria: deleted insufficient compound conditions(observations $< 5$); removed invalid compound SMILES which were not successfully parsed by RDKit; assigned each perturbed observation an unperturbed observation and removed the unpaired observations. Finally, we obtained 175,549 *cov_compounds_dose_name* condition (unperturbed-perturbed pair), which contains 82 cell lines and 17,202 compounds.

**The sci-Plex3 dataset.** The sci-Plex3 study[2] was subseted to 290,888 cells. We performed data cleaning using the following criteria: randomly subsampled the dataset to half size; selected 5000 highly variable genes (HVGs) using Scanpy; removed invalid molecule SMILES which could not successfully be parsed using RDKit[31]; assigned each perturbed observation an unperturbed observation and remove the unpaired observations. Finally, we obtained 2244 *cov_drug_dose_name*

condition (unperturbed-perturbed pair), which contains three cancer cell lines (A549, MCF7, K562), which were treated with 188 different compounds in 4 dosages (10, 100, 1000, and 10000 nM) and vehicle for unperturbed cells.

**The SCLC dataset.** The unperturbed profiles of small cell lung cancer collected from CCLE contain 5 cell lines, including NCI-H69, NCI-H526, NCI-H446, NCI-H209, and NCI-H196. We performed data preprocessing using the following criteria: selected 978 landmark genes of unperturbed profiles; and filled in the valid gene expression with the mean expression value of all genes in the current cell line. After the virtual screening, we obtained 4272486 transcriptional profiles.

**The compounds libraries.** Six compound libraries were used in this study. The training compound libraries are collected from the L1000[1] and the sci-Plex3 datasets[2]. in silico screening, an FDA-approved library (TargetMol, $n = 935$), an active compound library (Selleckchem, $n = 4158$), a natural compound library (Herb, $n = 30{,}456$), and an in-house drug-like compound library($n = 29{,}670$) were used to screen the positive chemicals. The sensitive compounds of cell lines are collected from Genomics of Drug Sensitivity in Cancer (GDSC) (https://www.cancerrxgene.org/) with z-score≥ 2.0.

**The gene signature of diseases.** The gene signature of diseases was collected from CRowd Extracted Expression of Differential Signatures (CREEDS)[43]. The CREEDS project annotated 839 diseases versus normal signatures from Gene Expression Omnibus (GEO). We collected the gene signature of the disease by removing the study with the intersected set of up- and down-regulated genes and 12,328 genes and set less than 1 gene and finally obtained 577 studies for 233 unique diseases. The gene signatures of diseases were downloaded from CREEDS.

**The scRNA-seq data of pediatric AML patients cohort.** We collected scRNA-seq data of paired pre- and post-chemotherapy whole bone marrow samples from 13 pediatric AML patients who achieved disease remission following chemotherapy, as described by Zhang et al.[35]. These patients were treated with either a low-dose chemotherapy (LDC) regimen or a standard-dose chemotherapy (SDC) regimen. The LDC regimen consisted of cytarabine (10 mg/$m^2$) and mitoxantrone or Idarubicin, administered concurrently with G-CSF (5 μg/kg). The SDC regimen consisted of cytarabine (100 mg/$m^2$), daunorubicin, and etoposide. We compiled the expression profiles of all 224,217 cells from the 13 patients and performed normalization and log transformation. Each cell was annotated with patient information, cell type, and chemotherapy regimen based on the information provided in the study. The LDC regimen was annotated with cytarabine and mitoxantrone, while the SDC regimen was annotated with cytarabine, daunorubicin, and etoposide. Through highly variable gene analysis, we retained 33,538 highly variable genes for training. The scRNA-seq data of patients were downloaded from the Genome Sequence Archive for Humans at the BIG data center, Beijing Institute of Genomics, Chinese Academy of Sciences, and China National Center for Bioinformation under accession number (OMIX005223).

**The data of A549 and MCF7 cell strains.** We collected expression profiles for 27 MCF7 strains and 23 A549 strains from Ben-David et al.[34]. Expression profiles for 978 L1000 landmark genes were used for each strain. We included 35 compounds from the chemical screen in the study[34], retaining 33 after removing those that RDKit could not generate to FCFP fingerprints. We also collected the drug response to 35 active compounds of MCF7 strains in the chemical screen, including decreasing (active), decreasing (weakly active), and inactive.

**Statistics and reproducibility.** All samples discussed in this manuscript (cell viability assays and perturbation profile atlas generation) were measured and analyzed as technical triplicates. No data were excluded from the analyses.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

For the bulk and single-cell datasets, we used expression profiles from the L1000[1] and the sci-Plex3[2] datasets. The L1000 and the sci-Plex3 datasets were downloaded from the Gene Expression Omnibus with the accession number (GSE92742) and (GSM4150378), respectively. Small cell lung cancer is available from The Cancer Cell Line Encyclopedia Project (CCLE) (https://sites.broadinstitute.org/ccle/). The gene signatures of diseases were downloaded from CREEDS. The scRNA-seq data of patients were downloaded from the Genome Sequence Archive for Humans at the BIG data center, Beijing Institute of Genomics, Chinese Academy of Sciences, and China National Center for Bioinformation under accession number (OMIX005223). We provided a website (prnet.drai.cn) to browse and download compound libraries and predicted results. The predicted signature results data in this paper have been deposited in the OMIX[65,66], China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences (https://ngdc.cncb.ac.cn/omixdatabase under accession code OMIX006910). Source data are provided in this paper.

## Code availability

The code for PRnet[67] is available at (https://github.com/Perturbation-Response-Prediction/PRnet).

## References


1. Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452 (2017).
2. Srivatsan, S. R. et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science* **367**, 45–51 (2020).
3. Drews, J. Drug discovery: a historical perspective. *Science* **287**, 1960–1964 (2000).
4. Lotfollahi, M. et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Mol. Syst. Biol.* **19**, e11517 (2023).
5. Piran, Z., Cohen, N., Hoshen, Y., Nitzan, M. Disentanglement of single-cell data with biolord. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-02079-x (2024).
6. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scgen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
7. Kana, O. et al. Generative modeling of single-cell gene expression for dose-dependent chemical perturbations. *Patterns* **4**, https://doi.org/10.1016/j.patter.2023.100817 (2023).
8. Hetzel, L. et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. *Adv. Neural Inf. Process Syst.* **35**, 26711–26722 (2022).
9. Bunne, C. et al. Learning single-cell perturbation responses using neural optimal transport. *Nat. Methods* **20**, 1759–1768 (2023).
10. Dong, M. et al. Causal identification of single-cell experimental perturbation effects with cinema-ot. *Nat. Methods* **20**, 1769–1779 (2023).
11. Dixit, A. et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).
12. Norman, T. M. et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* **365**, 786–793 (2019).

13. Roohani, Y., Huang, K., Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nat. Biotechnol.* **42**, 927–935 (2023).
14. Kamimoto, K. et al. Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**, 742–751 (2023).
15. Lamb, J. et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
16. Zhu, J. et al. Prediction of drug efficacy from transcriptional profiles with deep learning. *Nat. Biotechnol.* **39**, 1444–1452 (2021).
17. Zeng, B. et al. Octad: an open workspace for virtually screening therapeutics targeting precise cancer patient groups using gene expression features. *Nat. Protoc.* **16**, 728–753 (2021).
18. Sirota, M. et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* **3**, 96–779677 (2011).
19. Jahchan, N. S. et al. A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors. *Cancer Discov.* **3**, 1364–1377 (2013).
20. van Noort, V. et al. Novel drug candidates for the treatment of metastatic colorectal cancer through global inverse gene-expression profiling. *Cancer Res.* **74**, 5690–5699 (2014).
21. Goodfellow, I. et al. Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).
22. Kingma, D. P., Welling, M. Auto-encoding variational bayes. Preprint at https://doi.org/10.48550/arXiv.1312.6114 (2013).
23. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process Syst.* **33**, 6840–6851 (2020).
24. Rezende, D., Mohamed, S. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pp. 1530–1538 (2015).
25. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process Syst.* **33**, 1877–1901 (2020).
26. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
27. Betker, J. et al. Improving image generation with better captions. *Comput. Sci.* **2**, 3 (2023).
28. Razavi, A., Van den Oord, A. & Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. *Adv. Neural Inf. Process Syst.* **32**, 14837–14847 (2019).
29. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
30. Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inform. Comput. Sci.* **28**, 31–36 (1988).
31. Landrum, G. et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg. Landrum* **8**, 5281 (2013).
32. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).
33. Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
34. Ben-David, U. et al. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* **560**, 325–330 (2018).
35. Zhang, Y. et al. Single-cell transcriptomics reveals multiple chemoresistant properties in leukemic stem and progenitor cells in pediatric aml. *Genome Biol.* **24**, 199 (2023).
36. Grant, S., Easley, C. & Kirkpatrick, P. Vorinostat. *Nat. Rev. Drug Discov.* **6**, 21–22 (2007).
37. Kavanaugh, S. A., White, L. A. & Kolesar, J. M. Vorinostat: A novel therapy for the treatment of cutaneous t-cell lymphoma. *Am. J. Health Syst. Pharm.* **67**, 793–797 (2010).
38. Marks, P. A. & Breslow, R. Dimethyl sulfoxide to vorinostat: development of this histone deacetylase inhibitor as an anticancer drug. *Nat. Biotechnol.* **25**, 84–90 (2007).
39. Van Meerbeeck, J. P., Fennell, D. A. & De Ruysscher, D. K. Small-cell lung cancer. *Lancet* **378**, 1741–1755 (2011).
40. Lenhart, R. et al. Sensitivity of small cell lung cancer to bet inhibition is mediated by regulation of ascl1 gene expression. *Molecular cancer therapeutics* **14**, 2167–2174 (2015).
41. Siegel, R. L., Wagle, N. S., Cercek, A., Smith, R. A. & Jemal, A. Colorectal cancer statistics, 2023. *CA: a cancer journal for clinicians* **73**, 233–254 (2023).
42. Fang, S. et al. Herb: a high-throughput experiment-and reference-guided database of traditional chinese medicine. *Nucleic Acids Res.* **49**, 1197–1206 (2021).
43. Wang, Z. et al. Extraction and analysis of signatures from the gene expression omnibus by the crowd. *Nat. Commun.* **7**, 12846 (2016).
44. Makar, N. N. et al. Possible protective effects of mirabegron on experimentally induced non-alcoholic steatohepatitis in rats. *Benha Med. J.* **39**, 277–293 (2022).
45. Zhu, Y. et al. Repositioning an immunomodulatory drug vidofludimus as a farnesoid x receptor modulator with therapeutic effects on nafld. *Front. Pharmacol.* **11**, 590 (2020).
46. Abdel-Razik, A. et al. Rifaximin in nonalcoholic fatty liver disease: hit multiple targets with a single shot. *Eur. J. Gastroenterol. Hepatol.* **30**, 1237–1246 (2018).
47. Gangarapu, V. et al. Efficacy of rifaximin on circulating endotoxins and cytokines in patients with nonalcoholic fatty liver disease. *Eur. J. Gastroenterol. Hepatol.* **27**, 840–845 (2015).
48. Baumgart, D. C. & Sandborn, W. J. Crohn's disease. *Lancet* **380**, 1590–1605 (2012).
49. Li, M. et al. Integrated systematic pharmacology analysis and experimental validation to reveal the mechanism of action of semen aesculi on inflammatory bowel diseases. *J. Ethnopharmacol.* **298**, 115627 (2022).
50. Feagan, B. G. et al. Ozanimod induction therapy for patients with moderate to severe crohn's disease: a single-arm, phase 2, prospective observer-blinded endpoint study. *Lancet Gastroenterol. Hepatol.* **5**, 819–828 (2020).
51. Feagan, B. G. et al. Ozanimod as a novel oral small molecule therapy for the treatment of crohn's disease: The yellowstone clinical trial program. *Contemp. Clin. Trials* **122**, 106958 (2022).
52. Williams, D. L., Ozment-Skelton, T. & Li, C. Modulation of the phosphoinositide 3-kinase signaling pathway alters host response to sepsis, inflammation, and ischemia/reperfusion injury. *Shock* **25**, 432–439 (2006).
53. Moretti, C. et al. Combined oral contraception and bicalutamide in polycystic ovary syndrome and severe hirsutism: a double-blind randomized controlled trial. *J. Clin. Endocrinol. Metab.* **103**, 824–838 (2018).
54. Kabel, A. M., Al-Shehri, A. H., Al-Talhi, R. A. & Abd Elmaaboud, M. A. The promising effect of linagliptin and/or indole-3-carbinol on experimentally-induced polycystic ovarian syndrome. *Chem. Biol. Interact.* **273**, 190–199 (2017).
55. Elkind-Hirsch, K. E., Chappell, N., Seidemann, E., Storment, J. & Bellanger, D. Exenatide, dapagliflozin, or phentermine/topiramate differentially affect metabolic profiles in polycystic ovary syndrome. *J. Clin. Endocrinol. Metab.* **106**, 3019–3033 (2021).
56. Kaarbø, M., Klokk, T. I. & Saatcioglu, F. Androgen signaling and its interactions with other signaling pathways in prostate cancer. *Bioessays* **29**, 1227–1238 (2007).
57. Cheng, J., Yang, L., Kumar, V. & Agarwal, P. Systematic evaluation of connectivity map for disease indications. *Genome Med.* **6**, 1–8 (2014).

58. Bhat, G. R. et al. Cancer cell plasticity: From cellular, molecular, and genetic mechanisms to tumor heterogeneity and drug resistance. *Cancer Metastasis Rev.* **43**, 197–228 (2024).
59. Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 1–5 (2018).
60. Hunter, J. D. Matplotlib: A 2d graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
61. Waskom, M. L. Seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
62. Villanueva, R. A. M. & Chen, Z. J. ggplot2: elegant graphics for data analysis. Taylor Francis (2019).
63. Wu, T. et al. clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2**, https://doi.org/10.1016/j.xinn.2021.100141 (2021).
64. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (2019).
65. Chen, T. et al. The genome sequence archive family: toward explosive data growth and diverse data types. *Genom. Proteom. Bioinform.* **19**, 578–583 (2021).
66. CNCB-NGDC Members and Partners. Database resources of the national genomics data center, china national center for bioinformation in 2024. *Nucleic Acids Res.* **52**, 18–32 (2024).
67. Xiaoning, Q. et al. Predicting Transcriptional Responses to Novel Chemical Perturbations Using Deep Generative Model for Drug Discovery. Perturbation-Response-Prediction/PRnet: PRnet, https://doi.org/10.5281/zenodo.13751384 (2024).


## Acknowledgements

This work was supported by The National Key R&D Program of China (2021YFC2500203 to Y.Z.), The National Natural Science Foundation of China (32341019 to Y.Z., 32070670 to Y.Z.), Ningbo major project for high-level medical and healthcare teams (2023030615 to Y.Z.), Beijing Natural Science Foundation Haidian Origination and Innovation Joint Fund (L222007 to Y.Z.), Ningbo Science and Technology Innovation Yongjiang 2035 Project(2024Z229 to Y.Z.), Major Project of Guangzhou National Laboratory (GZNL2023A03001 to Y.Z.), Open Project of National Key Laboratory of Oncology Systems Medicine (KF2422-93 to Y.Z.), The National Key R&D Program of China (2022YFF1203303 to Y.Z.). The authors would like to acknowledge the Nanjing Institute of Infor-SuperBahn MLOps for providing the training and evaluation platform.

## Author contributions

X.Q. developed the kernel algorithms of PRnet, implemented the models, analyzed data, and wrote the manuscript. L.Z. contributed to data analysis, as well as revising and editing the manuscript. C.T. and Y.L. validated compound candidates through MTT assays and contributed to the writing. Z.-L.C. supported data analysis and participated in writing, reviewing, and editing. P.H. developed the PRnet website. R.C. facilitated the collaboration and supervised the research. X.L. and B.W. provided support for academic survey and platform development. S.Y. and Y.Z. directed the study and manuscript writing.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-53457-1.

**Correspondence** and requests for materials should be addressed to Shengyong Yang or Yi Zhao.

**Peer review information** *Nature Communications* thanks Sudin Bhattacharya, Xiling Shen, and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.