nature computational science

Supplementary information https://doi.org/10.1038/s43588-025-00887-6

Predicting drug responses of unseen cell types through transfer learning with foundation models

In the format provided by the authors and unedited

Contents

List of Figures

A Supplementary Notes

A.1 Implementation of baseline methods and scFMs

In order to conduct a reliable comparison, we use their default hyperparameter setting when conducting experiments on baselines. For each experiment for comparison, three replicates were run for each method with the same seeds. For CellOT and scGen, each drug is trained separately due to the characteristics of their frameworks. For chemCPA-FM model, we replace the input of the gene expression profile by scGPT-encoded embeddings. chemCPA-Pre model is first pretrained on L1000, a bulk drug perturbation transcriptome dataset, and fine-tuned on target scRNA-seq data.

chemCPA&CPA Hyperparameters used for chemCPA are listed below: - batch size: 128, - hidden dim: 32, drug encoder MLP: 128×4, - doser MLP: 64×3, - autoencoder MLP: 256×4, - adversary MLP: 256×3, learning rate: 5.6×10^{-4}, - adversary learning rate: 3.6×10^{-4}, - adversary reg: 3.98, - adversary penalty: 0.20. CPA is run based on chemCPA framework but without initialization of drug embedding using RDKit.

CellOT Hyperparameters used for CellOT is shown below: - hidden units: 64×4, - latent dimension: 50. The Adam optimization with learning rate as 1×10^{-4}, beta1 as 0.5 and beta2 as 0.9 is used for training.

scGen scGen is trained with batch size as 32, latent dimension as 50, drop out ratio as 0.2, hidden MLP as 512×2, and learning rate as 1×10^{-3}.

Geneformer Geneformer is a single-cell large language model trained from 29.9 million of scRNA-seq data from human tissues and a total vocabulary of 25,424 protein-coding or miRNA genes. Genes and cells are encoded into a 256-dimension hidden space. The pretrained Geneformer model contains 6 transformer layers with attention heads. The max length of input sentence is 2048.

scGPT The input of scGPT is preprocessed single cell gene expression data without highly variable genes subset. The size of scGPT's hidden embedding is 512. Batch size is 64, and max length of each cell's gene features is 1200. The scGPT-human model is trained from 33 million scRNA-seq samples from 51 human organs.

scFoundation The pretraining dataset of scFoundation contains 50 million human single-cell transcriptome profile and 19,264 protein-coding and common mitochon-drial genes in its gene vocabulary. The dimension of hidden embeddings of cells is 3072.

A.2 Ensuring no information Leakage

To ensure the integrity and reliability of our experimental results, we create a rigorous train-valid-test split for each experiment. We strictly adhered to a clear separation between training and held-out datasets. The training set was exclusively used to train the model, while the validation and test sets were withheld until the model evaluation phase. This separation was maintained regardless of the experimental scenario, ensuring that no overlap occurred between the training data and the data used for performance assessment. In particular, we describe the manner of train-valid-test split for different scenarios below:

Predict for unseen cell types : If we denote unseen cell type is C. All perturbed samples and 1/4 unperturbed samples of cell type C are extracted as test data. In the zero-shot prediction, 100% perturbed and unperturbed samples of cell type C are all extracted as test data. Samples of the remaining other cell types is split into a train and a validation subset according to the ratio of 4:1. For each dataset, we create three different splits to ensure the set-out unseen cell types are covered in all possible cell types.

Predict for unseen drugs and cell types : On the basis of the previous split manner, the perturbed samples of selected held-out drugs are also removed from the train and validation subset and extracted as test data. It ensures the model cannot see any observed perturbed features of these unseen drugs and cell types during training.

Lastly, we ensure that all evaluation metrics are computed solely based on the outputs from the held-out data in the test subset. We aimed to maintain the independence of our evaluation process and to provide a robust assessment of the model's generalization capabilities.

Algorithm 1: Cell Type-Specific Optimization Framework of `CRISP`

Input: Perturbation data $\mathbf{Y} \in \mathbb{R}^{(n+m)\times d}$ with unpaired perturbed profiles $\mathbf{Y_P} \in \mathbb{R}^{n\times d}$ and control profiles $\mathbf{X} \in \mathbb{R}^{m\times d}$; cell type labels $\mathcal{C} \in \mathbb{R}^{n+m}$, drug treatment labels $\mathcal{D} \in \mathbb{R}^{n+m}$ and dosage labels $\mathcal{S} \in \mathbb{R}^{n+m}$; maximum iteration ξ.

Output: Predicted post-perturbation profiles $\hat{\mathbf{Y}}$, and trained model f_θ.

Initialization:

Initialize model parameters θ;

Compute paired control pre-embeddings: $\hat{\mathbf{e}}_\mathbf{X} \leftarrow scFM(\mathbf{X})$;

Compute drug-conditioned pre-embedding: $\hat{\mathbf{e}}_\mathcal{D} \leftarrow DrugModel(\mathcal{D}, \mathcal{S})$;

Get set of index of contrastive samples: $\mathcal{I} \leftarrow \text{ContrastiveSample}(\mathbf{Y}, \mathcal{C}, \mathcal{D})$;

for $k \leftarrow 1$ ***to*** ξ **do**

  Forward Pass:

  Compute predicted perturbation profile and cell type label:

$$\hat{\mathbf{Y}}, \hat{\mathcal{C}}, \mathbf{Z}, \mu, \sigma \leftarrow f_\theta(\hat{\mathbf{e}}_\mathbf{X}, \hat{\mathbf{e}}_\mathcal{D})$$

  Loss Computation:

  Compute total loss:

$$\begin{aligned} \mathcal{L} = \;& \lambda_1 \text{MSE}(\mathbf{Y}, \hat{\mathbf{Y}}) + \lambda_2 \text{MMD}(\mathbf{Y}, \hat{\mathbf{Y}}) \\ &+ \lambda_3 \big(\text{ContrastiveLoss}(\mathbf{Z}, \mathbf{Z}_\mathcal{I}) + \text{CrossEntropy}(\mathcal{C}, \hat{\mathcal{C}})\big) \\ &+ \lambda_4 \text{KLD}(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(0,1)) \end{aligned}$$

  Update Model Parameters:

  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$;

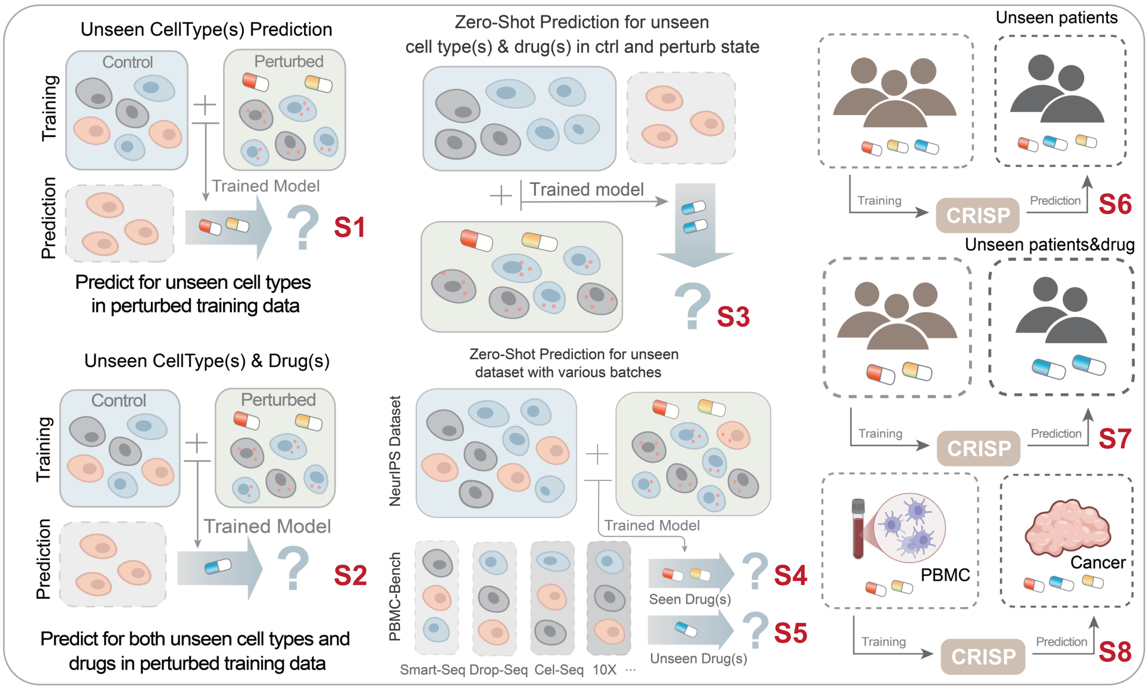
end

Output: Trained model f_θ with optimized parameters; Predicted post-perturbation profiles $\hat{\mathbf{Y}}$.

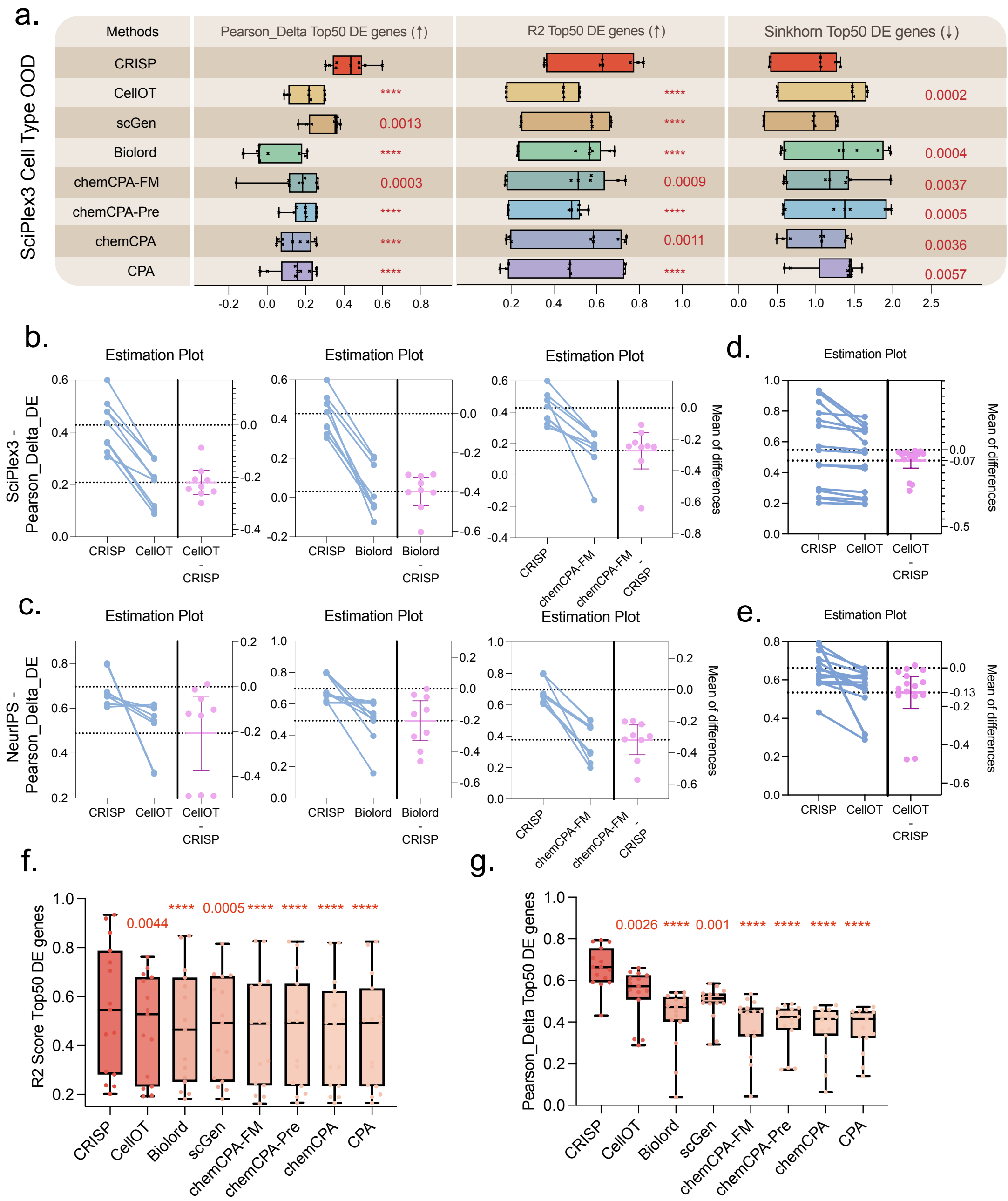
Supplementary Table 1: Selection of hyper-parameters using the NeurIPS dataset. The highest metric values are highlighted in bold, and the hyperparameters we selected are underlined.

Batch Size	Selections	64	<u>128</u>	256	512
	Pr_Δ	0.687	**0.693**	0.630	0.654
	R^2	0.536	0.538	0.530	**0.544**
Cell Type Loss Coefficient	Selections	0.01	0.1	<u>1</u>	10
	Pr_Δ	0.688	0.685	**0.693**	0.679
	R^2	0.536	0.531	**0.538**	**0.538**
Latent Dimensions	Selections	32	64	<u>128</u>	256
	Pr_Δ	**0.694**	0.684	**0.693**	0.687
	R^2	0.536	**0.539**	**0.538**	0.537
Learning Rate	Selections	0.00001	0.0001	<u>0.001</u>	0.01
	Pr_Δ	0.502	0.671	**0.693**	0.645
	R^2	0.486	**0.540**	0.538	0.512
MMD Loss Coefficient	Selections	0.001	0.01	<u>0.1</u>	1
	Pr_Δ	0.683	0.687	**0.693**	0.589
	R^2	**0.538**	**0.538**	**0.538**	0.508

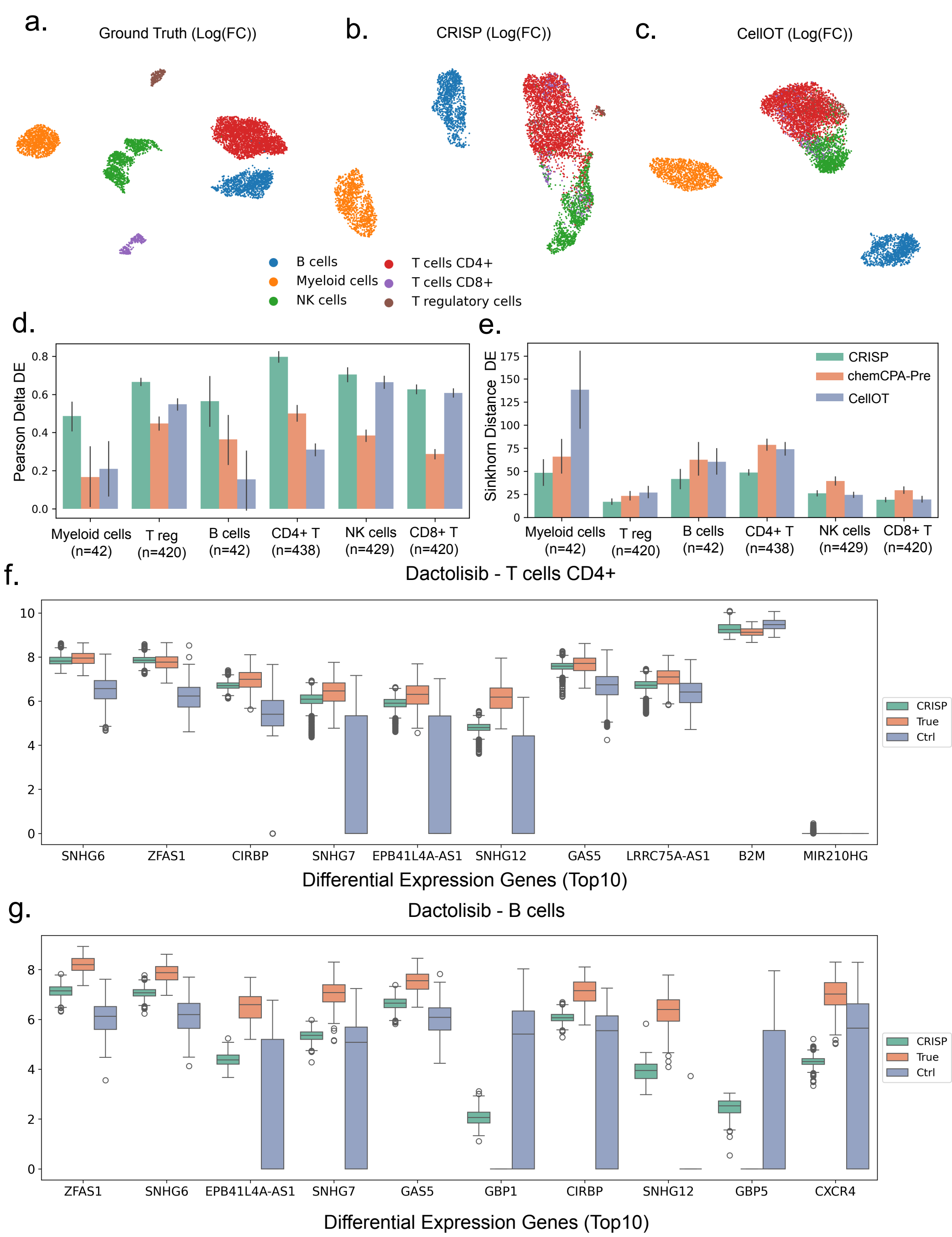
B Supplementary Figures



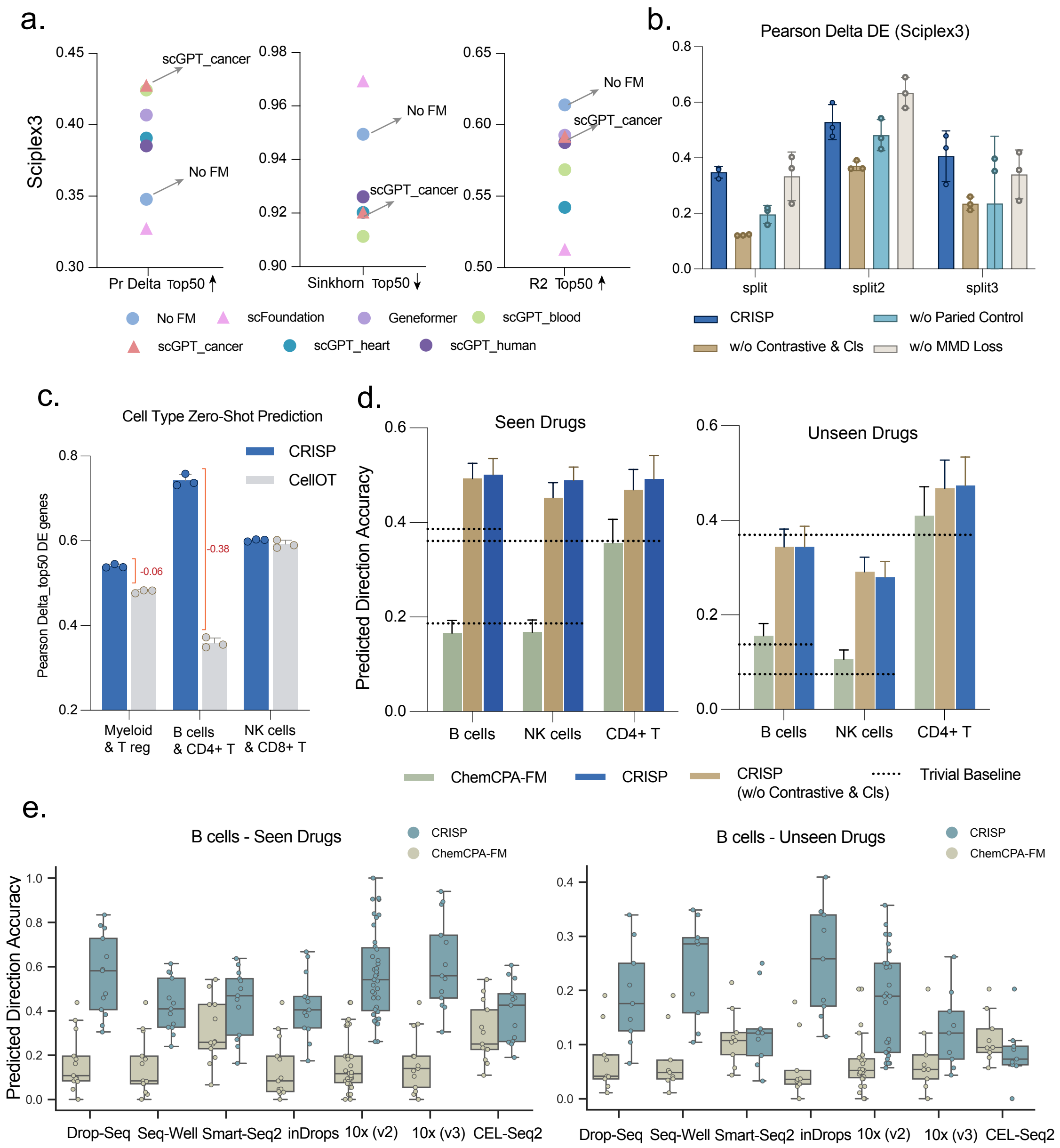
Supplementary Figure 1: **Overview of the different prediction scenarios for unseen data. S1.** Prediction for unseen cell types, where specific cell types are removed from the perturbed training data but kept in the control group. **S2.** Prediction for unseen cell types and drugs, where both held-out cell types and drugs are excluded from the perturbed training dataset. **S3.** Zero-shot prediction for unseen cell types and drugs, where held-out cell types are removed from the unperturbed group during training as well on the basis of **S2**. **S4,S5.** Zero-shot prediction for unseen batches, where the test and training data come from different datasets or protocols, but share the same cell types. This is separated into two settings: predicting for seen drugs (**S4**) and predicting for unseen drugs (**S5**). **S6,S7.** Prediction for unseen patients, where the trained model is applied to unseen patients with the same cell types. Predictions are made for seen drugs (**S6**) or unseen drugs (**S7**). **S8.** Prediction across tissues, where a model trained on the PBMC dataset is applied to cancer data. Some icons are created with BioRender.com.



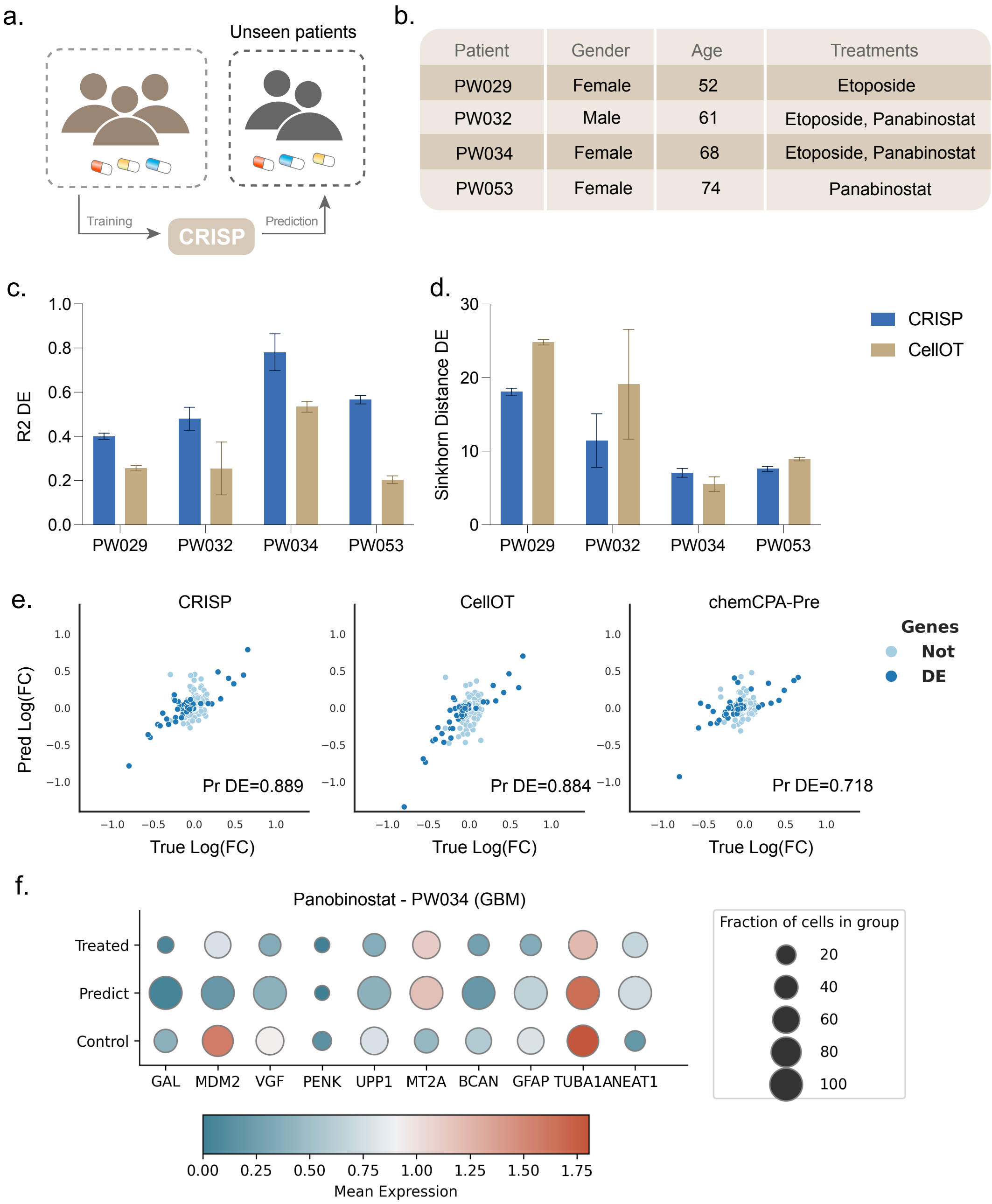
Supplementary Figure 2: **`CRISP` surpasses other methods across unseen cell types. a.** Comparison between different methods and `CRISP` using Pr_{Δ}, R^2, and Sinkhorn distance on Top 50 DE genes, performed on SciPlex3 dataset. Box plots display the distribution of 9 points with 3 split settings. Box: 25th–75th percentiles; center line: median; whiskers: minimum to maximum. Metric directionality is indicated by arrows. We use one-tailed paired t-test for statistical analysis of panels **a**, **f**, and **g**. The significance is annotated with red text (values ≤ 0.0001 are indicated by four stars). And these panels share the same boxplot settings. The corresponding significance is annotated with red text (values ≤ 0.0001 are indicated by four stars). **b.** The estimation plots of three pairs of comparisons, which are selected as representations. Each plot shows the distribution of differences between `CRISP` and the baseline of all experiments. These plots are generated from the evaluation results of the SciPlex3 dataset with Pr_{Δ}. **c.** Similar estimation plots with **b**, but performed on the NeurIPS dataset. **d, e, f, g.** Comparison of R2 score and Pearson Correlation Coefficient of Log(FC) on Top 50 DE genes across all methods. Each column contains 15 split settings (points).



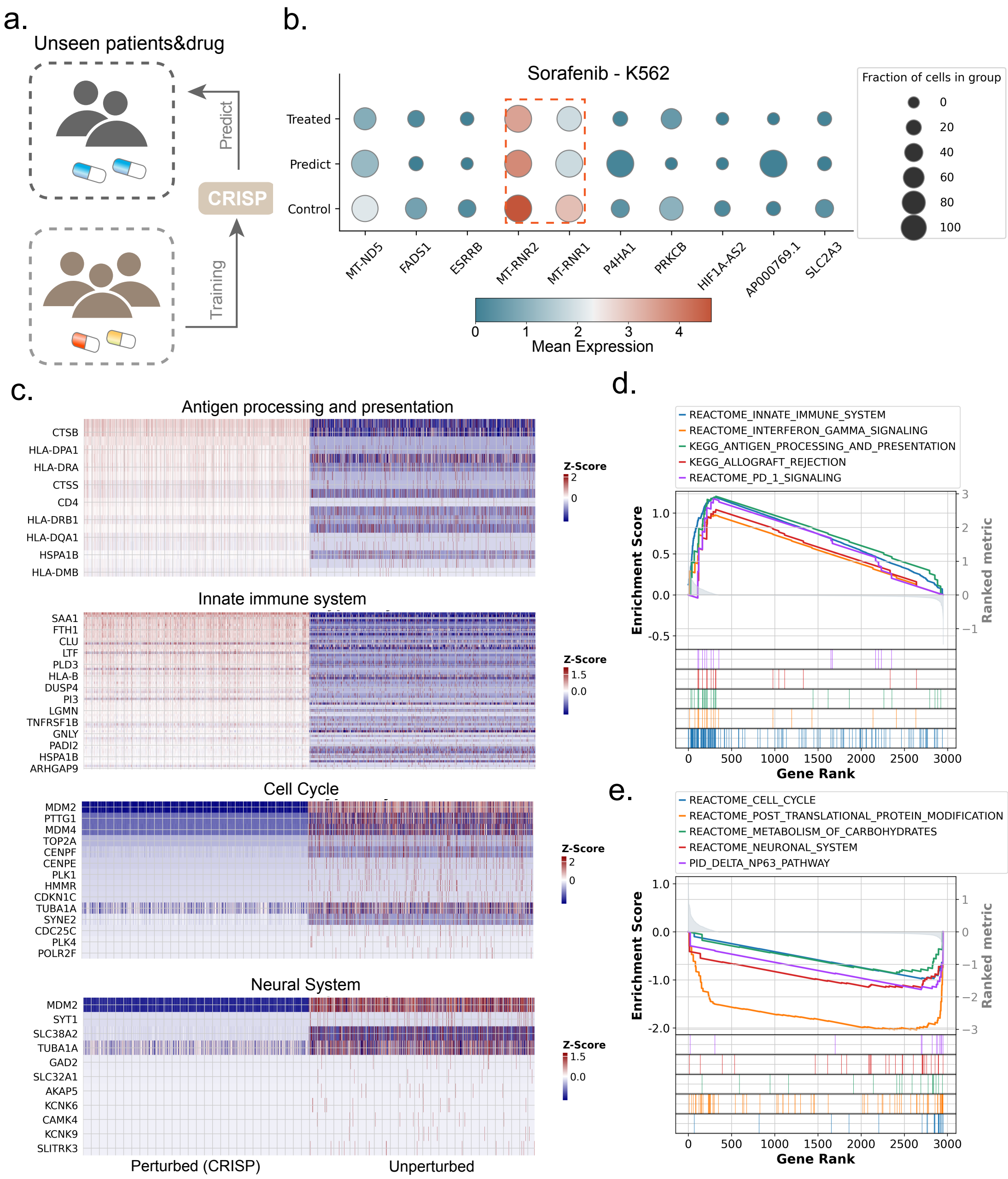
Supplementary Figure 3: **Visualization of predicted fold change across different cell types.** We show the UMAP visualization of log(FC) in ground truth (**a**), the prediction generated by CRISP (**b**), and CellOT (**c**). Six cell types are labeled by colors. Due to the unpairing of perturbed and control samples, the log(FC) of the ground truth is calculated from the difference between the perturbed gene expression profile and the mean of the control state profile grouped by cell type labels. **d, e.** Evaluation results of perturbation prediction across unseen cell types, measured with Pr_{Δ} and Sinkhorn distance on top50 DE genes. We show the number of cell type-drug combinations (n) in each cell type group. Error bar: 95% CI **f.** The distributions of gene expression value in untreated and treated states at the single-cell level for CD4+ T cells, focused on the Top 10 differential expression genes. For each DE gene, we show expression distribution in the control state (n=2598), perturbed state in the ground truth (n=760), and CRISP's predicted one (n=2695). Box: 25th–75th percentiles; center line: median; whiskers: data within 1.5 times the IQR from the quartiles. **g.** Similar illustration with **f** for B cells.



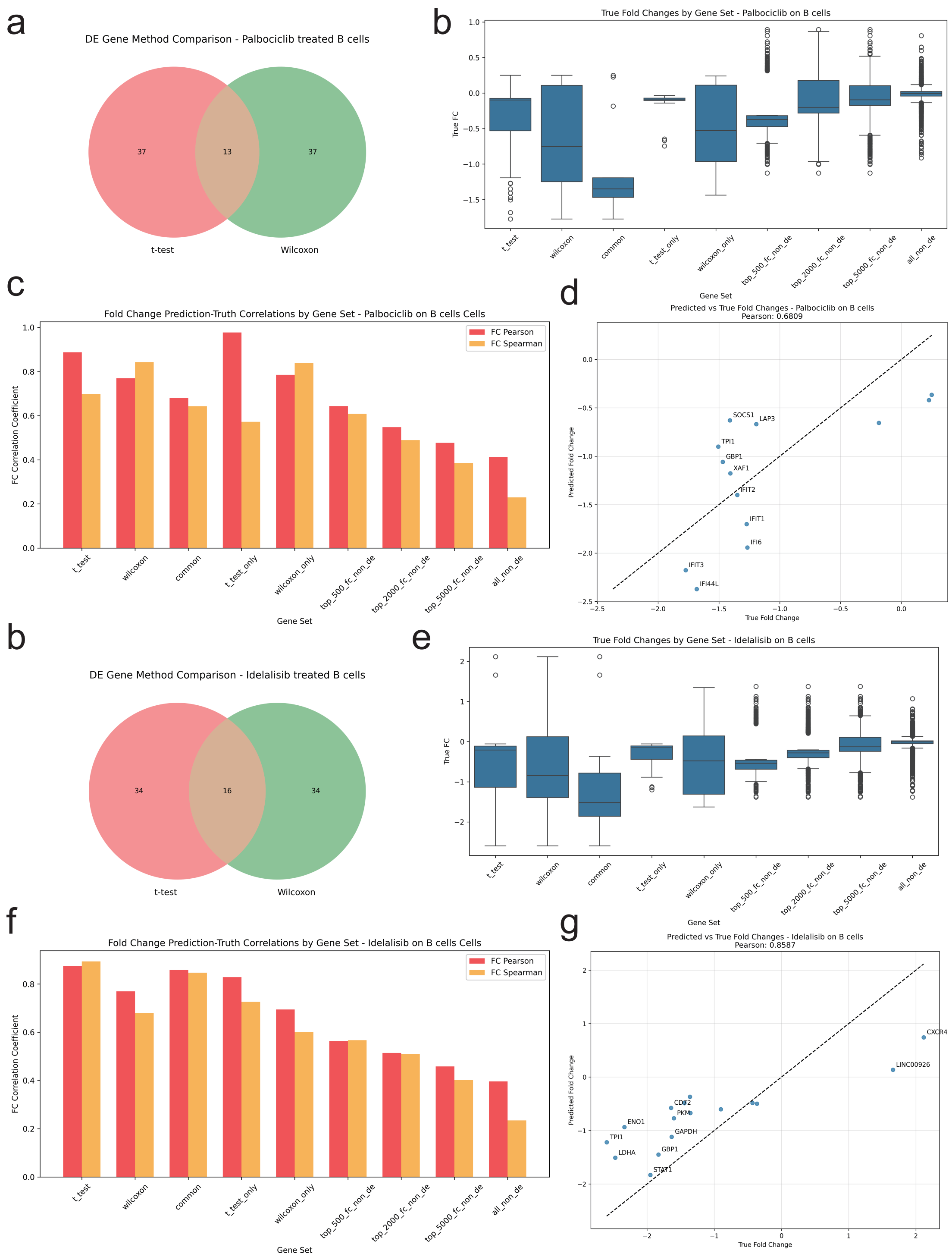
Supplementary Figure 4: **Ablation analysis and zero-shot prediction. a.** Benchmark of `CRISP` with various scFMs using SciPlex3 dataset. Overlap-FMs are labeled with a triangle symbol, and the overlap-scGPT model (scGPT-cancer) and the no-FM version are annotated by texts on plots. **b.** The ablation study results are evaluated on the SciPlex3 dataset with 3 replicated experiments. **c.** The evaluation of cell type zero-shot prediction on the NeurIPS dataset across three splits. **d, e.** The predicted fold change direction accuracy on the top 50 DE genes of zero-shot prediction of the PBMC-Bench dataset, compared with ChemCPA-FM model. Both scenarios of predicting seen and unseen drugs are presented. Each point corresponds to one of the 13 unseen drugs or 9 seen drugs. Error bar of **b, c, d** represents standard deviation.



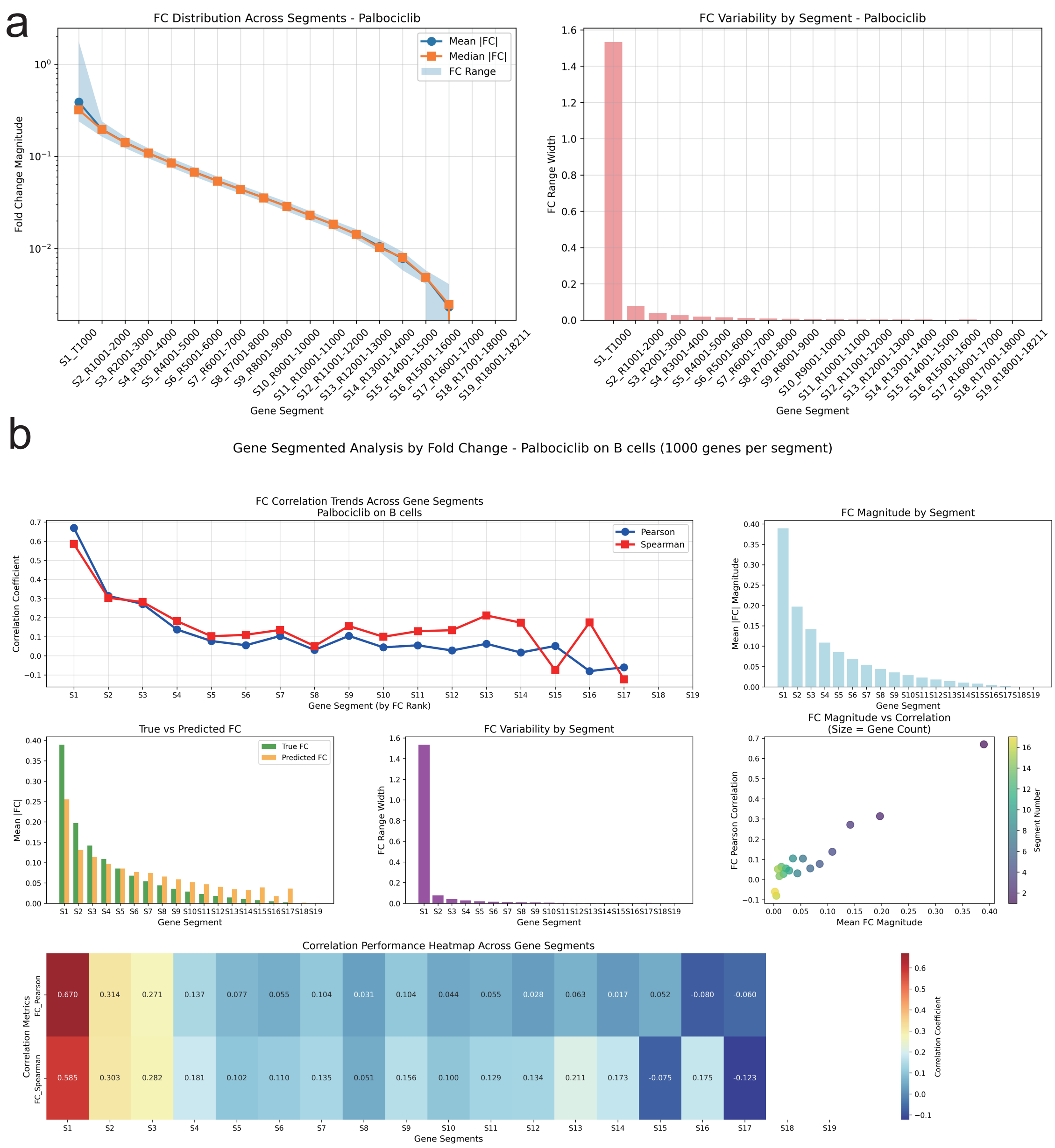
Supplementary Figure 5: **CRISP generalizes to unseen patient. a.** Evaluation of prediction on an unseen patient with the GBM dataset containing drug-treated glioblastoma samples. **b.** Information of four patients' treatment data used in experiments, including their gender, age, and treatment. These patients are treated with either Etoposide or Panabinostat which are both anticancer drugs. **c, d.** The mean and standard distance of the R^2 score and sinkhorn distance measured on DE genes are shown. Each point represents one of 3 experiments (PW029 and PW053) or 6 experiments (PW032 and PW034). Error bar: standard deviation. **e.** Distribution of observed log(FC) versus predicted log(FC) on Panobinostat treated PW034, compared CRISP with CellOT and chemCPA. DE genes are annotated as dark blue. **f.** Evaluation for unseen drug&patient prediction. It shows the gene expression visualization of Panobinostat-treated PW034, including the control state, true perturbed state, and predicted group. CRISP correctly predict the down-regulation of MDM2 and VGF genes related to multiple biological processes in cancer cells.



Supplementary Figure 6: **Gene expression pattern and enrichment analysis of unseen cell type and unseen patient prediction. a.** The illustration of prediction for unseen patients and drugs. **b.** Visualization of gene expression patterns of unseen cell type in Sorafenib-treated K562 cells, showing control, treated, and predicted outcomes by `CRISP`, with K562 cells excluded from the training dataset. The shift of MT-RNR1 and MT-RNR2 genes is correctly predicted by `CRISP`. **c, d, e.** Detailed GSEA analysis results of glioblastoma treated with the HDAC inhibitor panobinostat in an unseen patient PW034. Panobinostat has shown anti-tumor effects in glioma models[1]. Heatmaps display expression changes in the leading genes of each enriched process before and after treatment. Panobinostat's known mechanism of action aligns with the upregulation of innate immune system pathways[2]. Neuronal system-related genes, MDM2 and TUBA1A, are down-regulated and may induce the apoptosis of glioblastoma cells. It has been reported that MDM2 gene encodes a protein that inhibits the tumor suppressor p53 and cell cycle[3], and several HDAC inhibitors, like panobinostat, exert an anti-tumor effect via targeting MDM2-P53 signaling pathway[4, 5]. It highlights the capability of `CRISP` in predicting the post-treatment effect of panobinostat for unseen patients. Additionally, the positive enrichment of antigen processing and presentation indicates an activation of the immune process response to glioblastoma tumor cells.


a
DE Gene Method Comparison - Palbociclib treated B cells
37
13
37
t-test
Wilcoxon
b
True Fold Changes by Gene Set - Palbociclib on B cells
True FC
t_test
wilcoxon
common
t_test_only
wilcoxon_only
top_500_fc_non_de
top_2000_fc_non_de
top_5000_fc_non_de
all_non_de
Gene Set
c
Fold Change Prediction-Truth Correlations by Gene Set - Palbociclib on B cells Cells
FC Pearson
FC Spearman
FC Correlation Coefficient
Gene Set
d
Predicted vs True Fold Changes - Palbociclib on B cells
Pearson: 0.6809
SOCS1
LAP3
TPI1
GBP1
XAF1
IFIT2
IFIT1
IFI6
IFIT3
IFI44L
Predicted Fold Change
True Fold Change
b
DE Gene Method Comparison - Idelalisib treated B cells
34
16
34
t-test
Wilcoxon
e
True Fold Changes by Gene Set - Idelalisib on B cells
True FC
Gene Set
f
Fold Change Prediction-Truth Correlations by Gene Set - Idelalisib on B cells Cells
FC Pearson
FC Spearman
FC Correlation Coefficient
Gene Set
g
Predicted vs True Fold Changes - Idelalisib on B cells
Pearson: 0.8587
CXCR4
LINC00926
CD52
PKM
ENO1
GAPDH
TPI1
GBP1
LDHA
STAT1
Predicted Fold Change
True Fold Change

Supplementary Figure 7: **Differential expression analysis and correlation patterns across gene sets. a.** Venn diagram showing overlap between t-test and Wilcoxon test DE gene selection methods for Palbociclib treatment on B cells. Only 13 genes are shared between the two methods, highlighting the method-dependent nature of DE gene selection. **b.** True fold change distributions for Palbociclib treatment across different gene sets, demonstrating that DE genes show larger magnitude changes compared to non-DE genes. Non-DE genes are ranked by the magnitude of their fold changes, with top_500_fc_non_de, top_2000_fc_non_de, and top_5000_fc_non_de representing increasingly larger sets of non-DE genes with the largest absolute fold changes. The progressively narrower distributions centered around zero illustrate the decreasing effect sizes as more non-DE genes are included. **c.** Fold change correlation bar chart for Palbociclib treatment, showing both Pearson and Spearman correlation values across gene sets. The clear downward trend from t-test (0.87) to all non-DE genes (0.40) supports our conclusion that correlation is driven by biological signal rather than systematic artifacts. **d.** Correlation scatter plot for common DE genes in Palbociclib-treated B cells' group, showing strong correlation between predicted and true fold changes for genes identified by both statistical methods. **e-g.** Similar results for Idelalisib treatment on B cells, demonstrating the consistency of our findings across different perturbations. The Venn diagram (e) shows 16 common genes between the t-test and Wilcoxon methods for Idelalisib. The boxplot of true fold changes (f) displays similar patterns of effect size distribution as observed with Palbociclib. The scatter plot for common DE genes (g) shows high correlation (Pearson = 0.86), validating our model's prediction accuracy for genes with robust differential expression.



Supplementary Figure 8: **Genome-wide gene segmentation analysis for Palbociclib-treated B cells. a.** Left: FC distribution across segments showing mean —FC— (blue line), median —FC— (orange line), and FC range (shaded area) on log scale, demonstrating a systematic decrease in fold change magnitude across gene segments. Right: FC variability (range width) per segment, with the highest variability in the top-ranked segment containing genes with the strongest responses. **b.** Comprehensive performance overview for Palbociclib. Six-panel analysis of genes segmented by fold change magnitude (1000 genes per segment). Top row: (left) Correlation trends showing Pearson (blue) and Spearman (red) correlations across segments, revealing significant performance drop around segments 4-5; (right) Mean fold change magnitude per segment confirming ranking validity. Middle row: (left) True vs predicted fold change comparison showing decreasing prediction accuracy in lower-ranked segments; (right) FC variability within segments. Bottom: Correlation performance heatmap summarizing Pearson and Spearman correlations across all segments, with warm colors indicating higher correlations.

References

[1] Eudocia Q Lee, David A Reardon, David Schiff, Jan Drappatz, Alona Muzikansky, Sean A Grimm, Andrew D Norden, Lakshmi Nayak, Rameen Beroukhim, Mikael L Rinne, et al. Phase ii study of panobinostat in combination with bevacizumab for recurrent glioblastoma and anaplastic glioma. *Neuro-oncology*, 17(6):862–867, 2015.

[2] Mikolaj Medon, Eva Vidacs, Stephin J Vervoort, Jason Li, Misty R Jenkins, Kelly M Ramsbottom, Joseph A Trapani, Mark J Smyth, Phillip K Darcy, Peter W Atadja, et al. Hdac inhibitor panobinostat engages host innate immune defenses to promote the tumoricidal effects of trastuzumab in her2+ tumors. *Cancer research*, 77(10):2594–2606, 2017.

[3] Shunbin Xiong, Carolyn S Van Pelt, Ana C Elizondo-Fraire, Geng Liu, and Guillermina Lozano. Synergistic roles of mdm2 and mdm4 for p53 inhibition in central nervous system development. *Proceedings of the National Academy of Sciences*, 103(9):3226–3231, 2006.

[4] J Sonnemann, C Marx, S Becker, S Wittig, CD Palani, OH Krämer, and JF Beck. p53-dependent and p53-independent anticancer effects of different histone deacetylase inhibitors. *British journal of cancer*, 110(3):656–667, 2014.

[5] Akihiro Ito, Yoshiharu Kawaguchi, Chun-Hsiang Lai, Jeffrey J Kovacs, Yuichiro Higashimoto, Ettore Appella, and Tso-Pang Yao. Mdm2–hdac1-mediated deacetylation of p53 is required for its degradation. *The EMBO journal*, 2002.