EVALUATING CELL AND DRUG REPRESENTATIONS FOR GENERALIZABLE DRUG RESPONSE PREDICTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models provide increasingly sophisticated representations of cells and drugs, with promising applications in precision oncology and drug discovery. Yet it remains unclear when these representations improve drug-response prediction or generalize across drugs, cellular contexts, and datasets. We systematically assess the effects of cell and drug representations, predictive architectures, and dataset characteristics by benchmarking five cell representations, six drug representations, and three predictive models across two multimodal single-cell settings: a large, controlled shared-label dataset with uniformly generated transcriptomes and standardized cell line–drug response labels, and a smaller, heterogeneous cell-specific-label setting with finer-grained response annotations. We evaluate increasingly stringent regimes, from random splits to zero-shot prediction of unseen cell line–drug pairs, drugs, and cellular contexts. Across both settings, performance depends more strongly on the cell representation and predictive model than on the drug representation, while pretrained encoders do not consistently outperform simpler alternatives. Under cell line–drug pair zero-shot evaluation, the best MCC reaches 0.36 in the shared-label setting and 0.17 in the cell-specific-label setting. Generalization degrades sharply with stringent splits compared with random splits: in the shared-label setting, the average within-drug MCC decreases from about 0.88 with a random split to 0.18 with cell line–drug pair zero-shot. In the cell-specific-label setting, it decreases from 0.63 to 0.03. Conclusions also depend strongly on the evaluation axis; for example, shared-label cell-line zero-shot yields MCC 0 when evaluated across cellular contexts but 0.52 when aggregated across drugs. Multimodal models likewise do not consistently outperform single-modality baselines, indicating that much of the predictive signal can often be captured by one modality alone. These results highlight the importance of realistic zero-shot evaluation, complementary aggregation views, and strong simple and single-modality baselines.

1 INTRODUCTION

Predicting how cancer cells respond to treatment is a central challenge in precision oncology and drug discovery. Because drug response is jointly determined by the cellular state and the chemical properties of the compound, predictive models must integrate informative representations of both. Recent advances in pretraining have transformed how these representations are constructed. On the cellular side, single-cell RNA sequencing (scRNA-seq) provides transcriptome-wide measurements at single-cell resolution, and the growing availability of large-scale cell atlases has enabled the development of single-cell foundation models. Models such as Geneformer, scGPT, and GeneJEPA are pretrained on millions of cellular profiles using self-supervised objectives with the promise of capturing transferable biological information that can be reused to generate representations of cells from new datasets for downstream tasks, potentially reducing the need of task-specific retraining (Theodoris et al., 2023; Cui et al., 2024; Litman et al., 2025). On the chemical side, molecular language models such as ChemBERTa and MoLFormer are pretrained on hundreds of millions of molecules written as SMILES strings, providing continuous embeddings of compounds intended to replace classical structural fingerprints (Singh et al., 2026; Ross et al., 2022). Whether such pretrained representations actually provide an advantage over representations learned directly from

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the downstream data, however, remains unclear. Here, we examine this question in the context of predicting the response of cancer cells to multiple drugs.

Bulk pharmacogenomic resources such as Genomics of Drug Sensitivity in Cancer (GDSC) or The Cancer Cell Line Encyclopedia (CCLE) have enabled predictive modeling of this relationship by pairing molecular profiles of cancer cell lines with response measurements across drug panels (Barretina et al., 2012; Yang et al., 2013; Iorio et al., 2016). Single-cell transcriptomics provides a finer view of the cellular state, capturing the heterogeneity and plasticity that are central to cancer biology (Patel et al., 2014; Tirosh et al., 2016; McGranahan & Swanton, 2017; Kinker et al., 2020). Tahoe-100M extends this setting to large scale, profiling roughly 100 million cells from 50 cancer cell lines under a uniform protocol and more than 1,100 perturbations (Zhang et al., 2025). Together with standardized drug-response measurements, this makes single-cell drug response a natural setting for testing whether pretrained cell and drug representations improve prediction.

Such an evaluation, however, depends strongly on what is meant by generalization. A model may need to predict a new combination of known cell lines and drugs, or transfer to an entirely unseen cell line or compound. Likewise, performance can be measured across cell lines for a fixed drug or across drugs for a fixed cell line. These settings probe different signals, and pooled metrics may reflect broadly sensitive cell lines or potent drugs rather than learned cell–drug interactions. Existing benchmarks typically vary only one representation modality and therefore do not disentangle these effects (Wang et al., 2025; Shen et al., 2026; Baptista et al., 2022; Xiao et al., 2025).

We address this by building a systematic benchmark of five cell representations, six drug representations, and three predictive models under matched preprocessing and evaluation. We compare dataset-fitted and pretrained representations across two settings: (1) Shared-label: Tahoe-100M single-cell profiles from 43 cell lines labeled with GDSC responses, and (2) Cell-specific-label: 19 published single-cell datasets with cell-level response labels. We evaluate random, unseen-pair, unseen-cell-line, and unseen-drug splits along both response-matrix axes using the Matthews correlation coefficient (MCC).

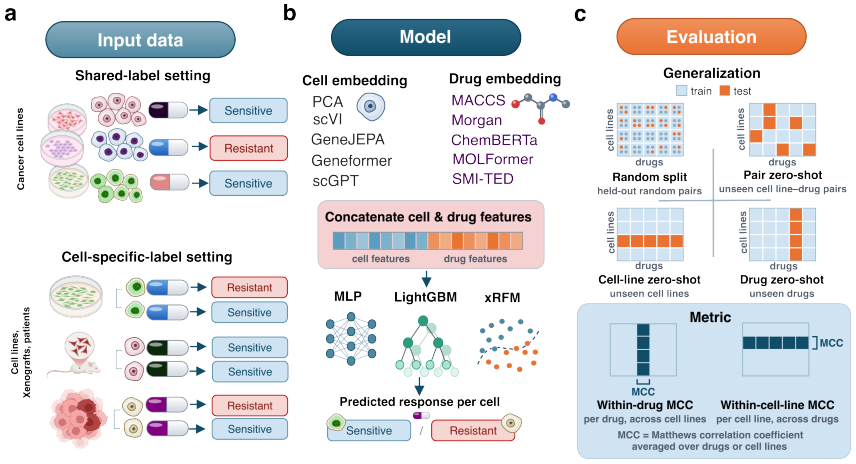

Figure 1: **Overview of the benchmark framework. (a)** Input data comprise shared-label and cell-specific-label single-cell drug-response datasets. **(b)** Early integration of Cell and drug representations is used by downstream classifiers to predict sensitivity or resistance. **(c)** Models are evaluated under random pair holdout, unseen cell line–drug pairs, unseen cell lines, and unseen drugs, using within-drug and within-cell-line MCC.

2 Related Work

Representation learning for cells and drugs. Single-cell representations range from dataset-fitted methods such as PCA and scVI to foundation models pretrained on millions of transcriptomes, including Geneformer, scGPT, and GeneJEPA (Gayoso et al., 2022; Theodoris et al., 2023; Cui et al., 2024; Litman et al., 2025). Although these models are intended to provide transferable representations, recent benchmarks have shown that pretrained embeddings do not consistently out-

perform simpler baselines across downstream tasks, and that conclusions can depend strongly on evaluation design (Ahlmann-Eltze et al., 2025). Molecular representations have similarly evolved from engineered structural descriptors like Morgan or MACCS fingerprints to pretrained foundation models such as ChemBERTa or MoLFormer (Rogers & Hahn, 2010; Durant et al., 2002; Ross et al., 2022; Singh et al., 2026). Benchmarks of chemical representations have compared fingerprints and learned embeddings for drug sensitivity, but generally hold the cellular representation fixed or omit it entirely (Baptista et al., 2022; Xiao et al., 2025). Our work examines this question in a multimodal setting, where cellular representations must contribute to prediction jointly with a chemical representation.

Single-cell drug-response benchmarks. Single-cell drug-response models either transfer predictors trained on bulk pharmacogenomic resources to single-cell profiles or learn directly from datasets with treatment-derived response annotations (Suphavilai et al., 2021; Chen et al., 2022; Zheng et al., 2023; Pellecchia et al., 2023; Sinha et al., 2024; Maeser et al., 2024). Recent benchmarks broaden this comparison: scDrugMap evaluates multiple single-cell foundation models across pooled and cross-dataset settings, while Shen et al. benchmark single-cell drug-response methods across heterogeneous cancer and treatment datasets (Wang et al., 2025; Shen et al., 2026). Both hold the drug fixed within each dataset and compare cell encoders or prediction methods, rather than systematically varying cellular and chemical representations under the same downstream predictors.

Generalization and evaluation in drug-response modeling. Evaluation design is particularly consequential in pharmacogenomic prediction. Random splitting can place related or repeated biological and chemical entities in both training and test sets, potentially producing optimistic performance estimates or direct information leakage (Sharifi-Noghabi et al., 2021; Kapoor & Narayanan, 2023; Bernett et al., 2024; Lenhof et al., 2024). Predicting an unmeasured response for a familiar cell line and drug is therefore distinct from transferring to a previously unseen cell line or compound. Multidrug prediction introduces an additional complication because every response is indexed by two entities. Metrics pooled across all cell line–drug pairs may therefore reflect between-drug or between-cell-line differences rather than the discrimination problem of interest. To our knowledge, existing benchmarks do not jointly vary modern cell and drug representations in a single-cell setting while controlling the predictor, comparing shared and cell-specific supervision, separating pair- from entity-level generalization, and evaluating both discrimination axes. Our benchmark is designed to isolate these factors within one framework.

3 METHOD

3.1 DATASET

Shared-label setting. We construct the shared-label dataset by integrating DMSO-treated control cells from Tahoe-100M with drug-sensitivity measurements from the Genomics of Drug Sensitivity in Cancer (GDSC) database (Yang et al., 2013; Zhang et al., 2025). The cellular input therefore reflects each cell line's baseline transcriptional state before drug exposure. Each cell line–drug pair receives a binary GDSC-derived sensitivity label shared by all cells in that pair, so the task is to predict cell line–level drug sensitivity from single-cell measurements of the untreated baseline state rather than to infer a direct cell-level treatment response. We retain drugs with Cmax viability labels and sensitive-class fraction between 0.15 and 0.85 to avoid highly imbalanced response distributions leading to unstable or undefined within-drug MCC estimates, yielding 43 cell lines and 10 drugs, 83,403 cells (806,030 cell–drug observations), and 416 distinct cell line–drug pairs. Further details on dataset construction and filtering are provided in the Supplementary Section B.1.

Cell-specific setting. We assembled 19 single-cell RNA-seq drug-response datasets spanning 10 compounds (Supplementary Table 4 and Fig. 9) and a range of biological systems, including cancer cell lines, patient-derived cell lines, xenografts, and primary chronic lymphocytic leukemia samples. Unlike the shared-label setting, response labels in these datasets are defined at the individual-cell level and arise from heterogeneous experimental designs, including treatment-survival labeling, transcriptional-state-based annotation, lineage tracing, and patient-level treatment response. Dataset composition and labeling procedures are summarized in Supplementary Section B.2.

The two settings capture different aspects of the drug-response prediction problem. The shared-label dataset emphasizes scale, consistency, and broad coverage across cell line–drug pairs, but response is defined at the pair level rather than for individual cells. The cell-specific datasets instead provide cell-level response information, offering more detailed supervision, while introducing substantial variation in how sensitivity and resistance are defined across datasets, being smaller, and less densely sampled across drugs and cellular contexts. Together, they allow us to examine model performance under a well-controlled benchmark and under a more granular but less complete data regime.

3.2 EMBEDDINGS

Cell embedding We evaluate five cell representations spanning increasing model complexity: PCA, scVI (Gayoso et al., 2022), GeneJEPA (Litman et al., 2025), Geneformer (Theodoris et al., 2023), and scGPT (Cui et al., 2024). These range from linear dimensionality reduction and dataset-fitted deep generative modeling to pretrained foundation models with different architectures and pretraining corpora. PCA and scVI are fitted within each cross-validation fold using training data only. PCA uses 3,000 highly variable genes and 50 components, while scVI uses 2,000 highly variable genes and a 50-dimensional latent space. GeneJEPA, Geneformer, and scGPT are used in inference mode with fixed pretrained weights and are not fine-tuned for the drug-response task, producing 768-, 768-, and 512-dimensional embeddings, respectively. Further model, checkpoint, preprocessing, and pretraining details are provided in Supplementary Methods Section C.1.

Drug embedding We compare six drug representations spanning classical structural fingerprints and pretrained molecular embeddings: Morgan fingerprints (Rogers & Hahn, 2010), MACCS fingerprints (Durant et al., 2002), ChemBERTa (Singh et al., 2026), MoLFormer (Ross et al., 2022), BioMed (Suryanarayanan et al., 2024), and SMI-TED (Soares et al., 2024). MACCS and Morgan yield 167- and 1,024-dimensional binary fingerprints, respectively. ChemBERTa, MoLFormer, BioMed, and SMI-TED provide dense pretrained representations of 384, 768, 512, and 768 dimensions, respectively, and differ in architecture, molecular input representation, and pretraining corpus. All pretrained molecular models are used only for feature extraction and are not fine-tuned for the drug-response task. Additional model and pretraining details are provided in Supplementary Methods Section C.2.

3.3 MODEL ARCHITECTURE AND EVALUATION

Multimodal integration. We formulate drug-response prediction as a binary classification task using three predictive architectures: a multilayer perceptron (MLP), LightGBM (LGBM), and xRFM. The models were chosen to represent complementary neural, tree-based, and alternative nonlinear prediction strategies, to test whether representation effects are consistent across classifiers. Cell and drug representations are concatenated and a single model is trained jointly across drugs and cell lines. Each configuration is evaluated with three random seeds. Architecture and hyperparameter details are provided in Supplementary Methods Section C.3.

Ablation analyses. To assess whether performance differences reflect information beyond marginal response structure, we performed two control analyses. First, prevalence controls ignored embeddings and predicted from marginal positive rates estimated on the training data. Second, single-modality ablations removed either the drug or cell representation while keeping the predictive model unchanged. These controls show how much performance can be explained by marginal class structure or either modality alone, providing a reference for the added value of the full multimodal model.

Evaluation units and MCC. The evaluation unit differs between settings. In the shared-label setting, all cells belonging to the same cell line–drug pair share a single GDSC-derived response label; held-out cell-level probabilities are therefore averaged within each pair before scoring, so that each cell line–drug pair contributes one prediction. In the cell-specific-label setting, response labels are defined at the individual-cell level, and predictions are evaluated directly at the single-cell level. Predictions are thresholded at 0.5 and evaluated using Matthews correlation coefficient (MCC). We report two complementary evaluation conventions: within-drug MCC, measuring discrimination across cellular contexts for each drug, and within-cell-line MCC, measuring discrimination across

drugs for each cell line. MCC is computed separately for each drug or cell line, macro-averaged across groups containing both classes ,and averaged across random seeds. Detailed definitions and equations are provided in the Supplementary Section D.1.

Representation-level ranking. To summarize cell-representation performance across drug representations and predictive models, the $5 \times 6 \times 3 = 90$ configurations are ranked within each split and evaluation convention using MCC averaged across seeds. Ranks r are converted to normalized scores defined as $1 - r/90$, with higher values indicating better relative performance and ties receiving their average rank. Each cell representation occurs in 18 configurations, and its plotted value is the mean normalized rank across these combinations. Error bars show the standard error across the 18 values and therefore summarize consistency across drug representation–model combinations rather than seed-to-seed uncertainty.

Generalizability. We evaluate model generalization under four splitting strategies that progressively reduce train–test overlap. The random split assigns individual observations to training and test sets and therefore evaluates prediction within previously observed cell line–drug combinations. The pair zero-shot split holds out complete cell line–drug pairs while allowing the constituent cell line and drug to remain represented in other combinations. The cell-line zero-shot and drug zero-shot settings further increase the generalization requirement by holding out an entire cell line or drug, respectively. For the shared-label setting, pair zero-shot splitting additionally enforces barcode-level disjointness between training and test data to prevent reuse of the same cells across folds. Detailed split construction is provided in the Supplementary Methods Section D.2.

Baseline configuration. We use PCA cell representations, MACCS drug fingerprints, and LightGBM under pair zero-shot as a classical multimodal reference configuration. PCA provides a linear, dataset-fitted cell representation, while MACCS uses fixed structural fingerprints. Pair zero-shot serves as the primary reference because it tests unseen cell line–drug combinations while retaining both entities in training, bridging random splitting and stricter cell-line- or drug-level zero-shot settings.

4 Experiments

The benchmark highlights consistent trends in the determinants of predictive performance and generalization. Predictive accuracy is driven primarily by how cellular features are encoded and by the architecture of the prediction model, whereas the representation of compounds has a comparatively smaller impact, with sophisticated pretrained encoders showing no systematic advantage over simpler methods. Performance drops markedly under zero-shot evaluation, although the difficulty depends on whether generalization is assessed across cell lines or drugs. Finally, combining cell and drug modalities does not consistently outperform single-modality models, suggesting that much of the predictive signal may already be captured by one modality.

4.1 Cell representation and predictive model dominate prediction of unseen drug-cell combinations

For the baseline configuration, combining PCA cell representations, MACCS drug fingerprints, and LightGBM, within-drug MCC reaches 0.36 ± 0.02 in the shared-label setting and 0.07 ± 0.03 in the cell-specific-label setting (Fig. 2). Notably, the shared-label value is the highest observed across the full representation–model grid.

No drug encoder shows a consistent advantage: Morgan and MACCS fingerprints remain competitive with pretrained molecular embeddings. For example, PCA–LightGBM ranges from MCC 0.31 to 0.36 across drug representations in the shared-label setting, while Geneformer–LightGBM ranges from 0.15 to 0.17 in the cell-specific-label setting. This is consistent with previous studies showing that learned molecular representations do not uniformly outperform established fingerprints and that gains depend on the task and evaluation setting (Baptista et al., 2022; Zagidullin et al., 2021; Duroux et al., 2026). Corresponding AUROC results are provided in Supplementary Figs. 11 and 12.

In both settings, the best-performing configurations use LightGBM, although the optimal cell representation differs: PCA–MACCS–LightGBM achieves 0.36 in the shared-label setting (base-

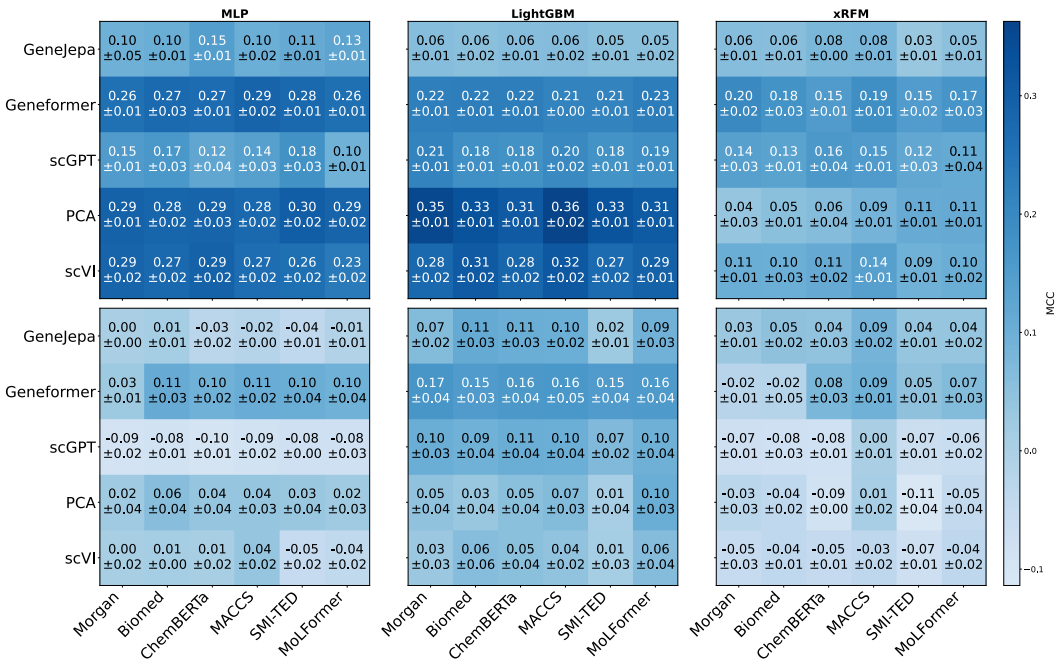

Figure 2: **Within-drug MCC across the full cell-representation × drug-representation grid under the pair zero-shot split.** Rows denote cell representations, columns denote drug representations, and panels correspond to the prediction models (MLP, LightGBM, and xRFM). Top: shared-label setting. Bottom: cell-specific-label setting.

line), whereas Geneformer–Morgan–LightGBM reaches 0.17 in the cell-specific-label setting, corresponding to a moderate and a weak positive predictive signal, respectively. The main difference between the two settings is the cell representation carrying the most predictive information. In the shared-label setting, PCA and scVI perform particularly well: with LightGBM, PCA achieves MCC values of 0.31–0.36 and scVI reaches 0.32. Geneformer is also competitive, especially with the MLP (MCC 0.26–0.29), but pretrained cell representations do not systematically outperform the simpler dataset-fitted embeddings. In contrast, overall performance is lower in the cell-specific-label setting, with many configurations near or below zero.

These differences are consistent with the granularity of the supervised signal. In the shared-label setting, all cells from a cell line–drug pair share the same label, making representations that capture stable between-cell-line differences particularly informative. This is also reflected in the PCA and scVI UMAPs, which show clear separation by cell line (Fig. 10). In the cell-specific-label setting, labels vary between individual cells, so prediction requires finer-grained cellular-state information rather than primarily cell-line identity. Indeed, PCA still separates cell lines clearly in this setting but no longer performs best, suggesting that cell-line identity alone is insufficient to explain response variation. The cell-specific-label setting is also substantially sparser, with many drugs observed in only one cell type, limiting the information available to learn transferable cell–drug interactions. Overall, the two settings therefore differ in the type of cellular signal that is most useful, while both indicate that cell representation and downstream model choice matter more for pair zero-shot generalization than the specific drug encoder.

4.2 Cell representation rankings are stable in the cell-specific-label setting but depend on split and evaluation regime in the shared-label setting

In the cell-specific-label setting, the conclusions about cell representations are highly robust across splitting strategies (Fig. 3). The configuration-level MCC values underlying these rank summaries are reported in Supplementary Section F.5. Geneformer ranks highest across the random, pair zero-shot, cell-line zero-shot, and drug zero-shot splits, under both within-drug and within-cell-line MCC. Across these four splitting strategies and the two evaluation conventions, configurations

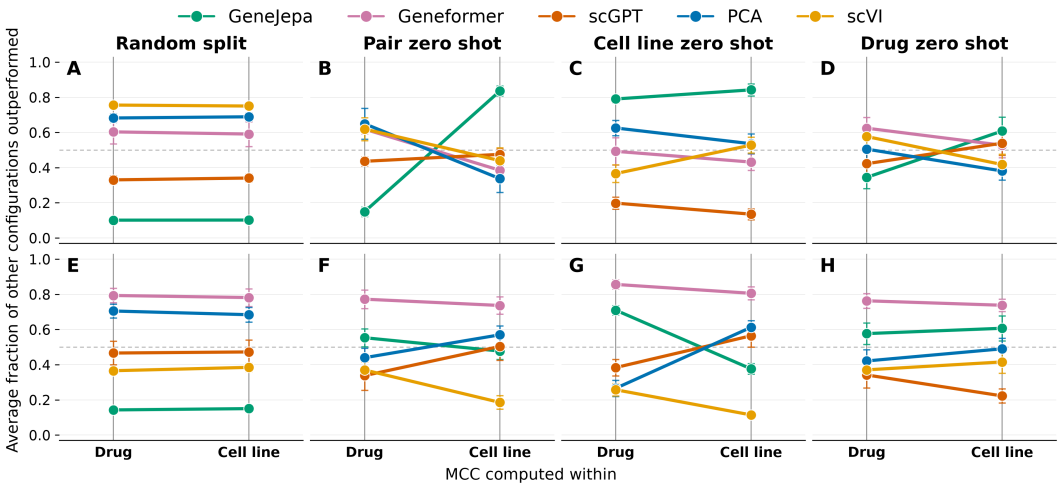

Figure 3: **Cell-representation rankings across splits and evaluation conventions.** Cell representations are ranked according to the average fraction of alternative configurations they outperform using within-drug MCC (left) or within-cell-line MCC (right). Panels correspond to random (A, E), pair zero-shot (B, F), cell-line zero-shot (C, G), and drug zero-shot (D, H) splits. Error bars show the standard error across 18 configurations (Section 3.3 Representation-level ranking). Top: shared-label setting. Bottom: cell-specific-label setting.

using Geneformer outperform, on average, 74–86% of the alternative configurations. The relative ordering of the remaining representations is less stable. For example, under cell-line zero-shot, GeneJEPA decreases from 0.71 under within-drug MCC to 0.38 under within-cell-line MCC, while PCA increases from 0.27 to 0.61.

The shared-label setting shows substantially greater sensitivity to both the splitting strategy and the evaluation convention. Under the random split, PCA and scVI rank highest under both metrics, consistent with their strong performance observed previously. A similar ordering is obtained for within-drug MCC under pair zero-shot, where PCA, scVI, and Geneformer rank above GeneJEPA. However, this pattern does not generalize to the more restrictive entity-level splits. Under cell-line zero-shot, GeneJEPA becomes the highest-ranked representation under both evaluation conventions, while under drug zero-shot, the ranking depends strongly on the evaluation axis: Geneformer and scVI rank highest by within-drug MCC, whereas GeneJEPA performs best by within-cell-line MCC.

Together, these results show that representation rankings depend on the evaluation axis, particularly in the shared-label setting, where within-drug evaluation favors representations capturing cell-line differences while within-cell-line evaluation reduces this advantage. Geneformer remains consistently highly ranked across splits and evaluation axes in the cell-specific-label setting. Thus, representation quality is more context-dependent in the shared-label setting, where no single representation consistently dominates.

4.3 Zero-shot generalization is substantially harder and depends on the evaluation axis

Moving from the random split to the zero-shot settings causes a substantial drop in performance in both labeling settings, but the magnitude of this drop differs markedly (Fig. 4). In the shared-label setting, average MCC decreases from 0.88–0.91 under the random split to 0.18 and 0.40 under pair zero-shot for within-drug and within-cell-line evaluation, respectively. This large reduction is expected because the random split allows cells from the same cell line–drug pair, which share a common label, to occur in both training and test sets, whereas pair zero-shot requires prediction for an unseen combination. We therefore regard the random split as a positive-control regime rather than a realistic test of drug-response generalization. The reduction in performance is even more pronounced when the entity relevant to the evaluation axis is held out: cell-line zero-shot reduces within-drug MCC to zero, while drug zero-shot reduces within-cell-line MCC to about 0.14. By

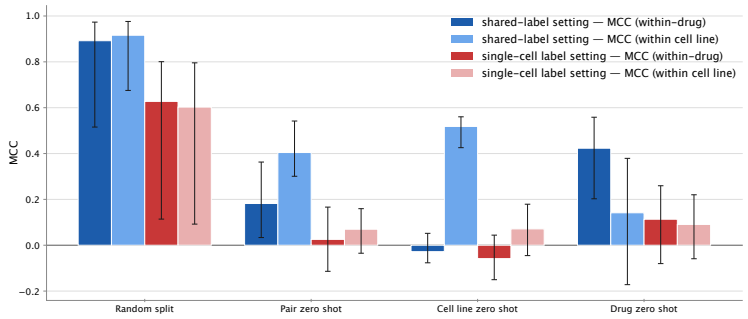

Figure 4: **Performance across data-splitting strategies under the two evaluation conventions.** Bars show the mean MCC across all cell-embedding, drug-embedding, and predictive-model configurations for each split, evaluated using within-drug MCC (dark tone) or within-cell-line MCC (light tone) for the shared-label setting (blue) and single-cell label setting (red). Error bars indicate the minimum–maximum range across configurations.

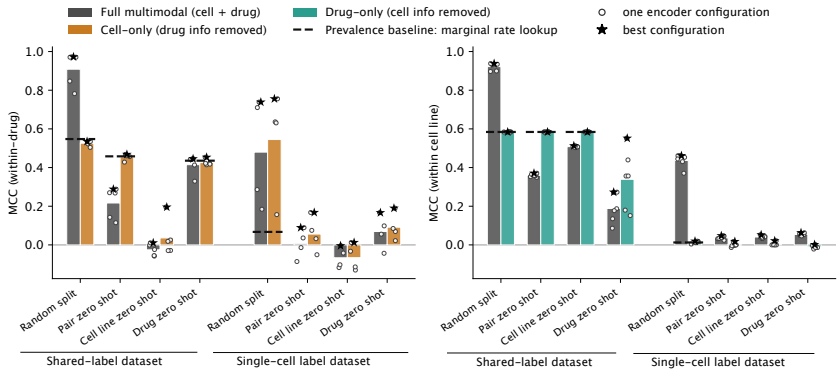

Figure 5: **Single-modality ablations and prevalence controls across splitting strategies.** Results are based on the MLP predictive model. Performance of the full multimodal model (gray) is compared with the cell-only (orange) and drug-only (turquoise) ablations under within-drug MCC (left) and within-cell-line MCC (right). Bars show the mean MCC across encoder configurations, open circles denote individual configurations, and stars indicate the best-performing configuration for each split. Dashed lines mark the corresponding prevalence controls. Where the dashed line is absent, the prevalence controls is undefined and cannot be computed.

contrast, the complementary combinations remain substantially more predictive, with a within-cell-line MCC of 0.52 under cell-line zero-shot and a within-drug MCC of 0.42 under drug zero-shot.

The cell-specific-label setting follows the same trend but at substantially lower performance: average MCC drops from 0.60–0.63 under random splitting to near zero under pair zero-shot, with similarly weak performance under entity-level holdouts. Across both settings, random splits therefore overestimate generalization, and zero-shot performance depends on the evaluation axis. The key difference is that moderate transferable signal remains in the shared-label setting, whereas it largely disappears in the cell-specific setting once cell–drug combinations are withheld.

4.4 Ablation analyses separate marginal entity effects from transferable cell–drug signal

The ablation analyses further clarify where the predictive signal originates in each setting (Fig. 5). In the shared-label setting, the random split produces a very high MCC for the full model (>0.9 under both evaluation conventions), clearly exceeding the corresponding single-modality models. However, this advantage largely disappears under the more restrictive splits. For within-drug MCC, the cell-only model matches or exceeds the full model under pair zero-shot (0.46 versus 0.22), whereas for within-cell-line MCC the drug-only model reaches 0.58 under several splits and often

exceeds the full model. These single-modality results are close to the corresponding marginal-prevalence controls, where defined, indicating that much of the predictive signal can be explained by stable cell-line- or drug-specific response propensities rather than transferable cell–drug interactions.

The cell-specific-label setting shows a more asymmetric pattern. For within-drug MCC, the cell-only model closely tracks, and sometimes exceeds, the full model, indicating little additional benefit from the drug representation. By contrast, within-cell-line performance collapses when cellular information is removed: the full model reaches 0.44 under the random split and retains small positive MCCs (0.04–0.05) under zero-shot evaluation, whereas the drug-only model remains near zero or slightly negative. Because many drugs are tested in only one cell type, however, drug identity is partially confounded with cellular context, limiting the ability to learn transferable drug effects.

Together, these controls reveal distinct sources of predictive signal. In the shared-label setting, much of the performance can be recovered from entity-specific marginal structure, consistent with all cells from a cell line–drug pair sharing the same label. In the cell-specific-label setting, predictive information is concentrated more strongly in cellular state, consistent with the larger effect of cell than drug representations observed above.

5 Discussion

Drug-response prediction depends jointly on cellular and chemical context, but we cannot assess the quality of their representations independently of how we evaluate generalization. We find that representation rankings change with the supervision structure, the held-out entities, and whether performance is measured across cell lines or across drugs. In particular, strong performance under random or pair-level splits does not necessarily translate to entity-level generalization, and pretrained representations do not consistently outperform simpler alternatives. Thus, evaluation design fundamentally shapes what conclusions can be drawn about representation quality.

Which evaluation is most informative depends on the intended generalization. Ranking drugs for a new cellular context is captured by within-cell-line performance under cell-line zero-shot, whereas stratifying new cellular contexts for a fixed drug requires within-drug evaluation under the same split. Generalization to novel compounds instead requires drug zero-shot. Reporting only one split or aggregation axis can therefore obscure important failure modes and lead to different conclusions about the same representation.

Our analysis is limited by the available supervision. Tahoe contains only 10 drugs and 416 independent cell line–drug response labels, while the cell-specific datasets provide finer supervision but much sparser coverage across drugs and cellular contexts. The fixed MCC threshold may also introduce sensitivity to calibration. More informative benchmarks will require larger, systematically crossed drug–cell datasets, realistic entity-level holdouts, complementary evaluation axes, and simple or single-modality controls alongside pretrained representations. These elements are necessary to distinguish transferable cell–drug interactions from performance driven by the structure of the benchmark itself. More broadly, the field is currently limited by the absence of large, chemically diverse, systematically crossed drug–cell datasets; generating such resources will be essential for distinguishing genuine transferable cell–drug interactions from marginal cell- or drug-specific effects.

AI Use Statement

Generative AI tools were used to assist with writing and optimizing the code for the project, as well as drafting and editing portions of this manuscript, including improving exposition and LaTeX formatting. They were not used to generate experimental data or experimental results. All AI-assisted text and technical content were reviewed and verified by the authors, who take responsibility for the final manuscript.

Reproducibility Statement

All code used to build the datasets, compute the cell and drug representations, train the predictive models, and generate the figures is available anonymously at https://github.com/anonymoussinglecell/iclr2027. The processed inputs, all precomputed cell and drug

representations for both settings, and the complete result tables are deposited on Zenodo under a CC BY 4.0 license at https://doi.org/10.5281/zenodo.22962632. The released tables regenerate every figure in the paper, and the released inputs allow any single configuration to be retrained. Dataset construction, preprocessing, model hyperparameters, split construction, and evaluation are detailed in Appendices B–D, and software versions in Appendix E.

REFERENCES

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8): 1657–1661, August 2025. doi: 10.1038/s41592-025-02772-6.

Delora Baptista, João Correia, Bruno Pereira, and Miguel Rocha. Evaluating molecular representations in machine learning models for drug response prediction and interpretability. *Journal of Integrative Bioinformatics*, 19(3):20220006, 2022.

Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, Joseph Lehár, Gregory V. Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F. Berger, John E. Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa A. Mapa, Justin Thibault, Eva Bric-Furlong, Purvesh Raman, Aaron Shipway, Ingerborg H. Engels, Jie Cheng, Guoying K. Yu, Jianjun Yu, Peter Aspesi, Melanie de Silva, Kalpana Jagtap, Michael D. Jones, Li Wang, Charles Hatton, Erica Palescandolo, Supriya Gupta, Scott Mahan, Carrie Sougnez, Robert C. Onofrio, Ted Liefeld, Laura MacConaill, Wendy Winckler, Michael Reich, Nanxin Li, Jill P. Mesirov, Stacey B. Gabriel, Gad Getz, Kristin Ardlie, Vivien Chan, Vaughan E. Myer, Barbara L. Weber, Jeff Porter, Markus Warmuth, Peter Finan, Jennifer L. Harris, Matthew Meyerson, Todd R. Golub, Michael P. Morrissey, William R. Sellers, Robert Schlegel, and Levi A. Garraway. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391): 603–607, March 2012. doi: 10.1038/nature11003.

Judith Bernett, David B Blumenthal, Dominik G Grimm, Florian Haselbeck, Roman Joeres, Olga V Kalinina, and Markus List. Guiding questions to avoid data leakage in biological machine learning applications. *Nature Methods*, 21(8):1444–1453, 2024.

Junyi Chen, Xiaoying Wang, Anjun Ma, Qi-En Wang, Bingqiang Liu, Lang Li, Dong Xu, and Qin Ma. Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. *Nature Communications*, 13(1):6494, October 2022. doi: 10.1038/s41467-022-34277-7.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, August 2024. doi: 10.1038/s41592-024-02201-0.

Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, 2002. doi: 10.1021/ci010132r.

Diane Duroux, Paul P Meyer, Giovanni Visonà, and Niko Beerenwinkel. Generalizable machine learning models for rapid antimicrobial resistance prediction in unseen health care settings. *GigaScience*, 15:giaf156, 2026.

Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mariano Gabitto, Mohammad Lotfollahi, Valentine Svensson, Eduardo da Veiga Beltrame, Vitalii Kleshchevnikov, Carlos Talavera-López, Lior Pachter, Fabian J. Theis, Aaron Streets, Michael I. Jordan, Jeffrey Regier, and Nir Yosef. A python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(2):163–166, February 2022. doi: 10.1038/s41587-021-01206-w.

Kaishun Huang and Hui Liu. Identification of drug-resistant individual cells within tumors by semi-supervised transfer learning from bulk to single-cell transcriptome. *Communications Biology*, 8 (1):530, March 2025. doi: 10.1038/s42003-025-07959-3.

Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, September 2023. doi: 10.1016/j.patter.2023.100804.

Gabriela S. Kinker, Alissa C. Greenwald, Rotem Tal, Zhanna Orlova, Michael S. Cuoco, James M. McFarland, Allison Warren, Christopher Rodman, Jennifer A. Roth, Samantha A. Bender, Bhavna Kumar, James W. Rocco, Pedro A. C. M. Fernandes, Christopher C. Mader, Hadas Keren-Shaul, Alexander Plotnikov, Haim Barr, Aviad Tsherniak, Orit Rozenblatt-Rosen, Valery Krizhanovsky, Sidharth V. Puram, Aviv Regev, and Itay Tirosh. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nature Genetics*, 52(11):1208–1218, 2020. doi: 10.1038/s41588-020-00726-6.

Kerstin Lenhof, Lea Eckhart, Lisa-Marie Rolli, and Hans-Peter Lenhof. Trust me if you can: A survey on reliability and interpretability of machine learning approaches for drug sensitivity prediction in cancer. *Briefings in Bioinformatics*, 25(5):bbae379, August 2024. doi: 10.1093/bib/bbae379.

Elon Litman, Tyler Myers, Vinayak Agarwal, Ekansh Mittal, Orion Li, Ashwin Gopinath, and Timothy Kassis. GeneJepa: A predictive world model of the transcriptome. bioRxiv, October 2025.

Danielle Maeser, Weijie Zhang, Yingbo Huang, and R. Stephanie Huang. A review of computational methods for predicting cancer drug response at the single-cell level through integration with bulk RNAseq data. *Current Opinion in Structural Biology*, 84:102745, February 2024. doi: 10.1016/j.sbi.2023.102745.

Nicholas McGranahan and Charles Swanton. Clonal heterogeneity and tumor evolution: Past, present, and the future. *Cell*, 168(4):613–628, February 2017. doi: 10.1016/j.cell.2017.01.018.

Anoop P. Patel, Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, Brian V. Nahed, William T. Curry, Robert L. Martuza, David N. Louis, Orit Rozenblatt-Rosen, Mario L. Suvà, Aviv Regev, and Bradley E. Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, June 2014. doi: 10.1126/science.1254257.

Simona Pellecchia, Gaetano Viscido, Melania Franchini, and Gennaro Gambardella. Predicting drug response from single-cell expression profiles of tumours. *BMC Medicine*, 21(1), December 2023. doi: 10.1186/s12916-023-03182-1.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022. doi: 10.1038/s42256-022-00580-7.

Hossein Sharifi-Noghabi, Soheil Jahangiri-Tazehkand, Petr Smirnov, Casey Hon, Anthony Mammoliti, Sisira Kadambat Nair, Arvind Singh Mer, Martin Ester, and Benjamin Haibe-Kains. Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. *Briefings in bioinformatics*, 22(6):bbab294, 2021.

Li Shen, Xinliang Sun, Shuyu Zheng, Ali Hashmi, Johanna Eriksson, Harri Mustonen, Hanna Seppänen, Bairong Shen, Min Li, Markus Vähä-Koskela, and Jing Tang. Predicting pre-treatment resistance or post-treatment effect? A systematic benchmarking of single-cell drug response models. *bioRxiv*, 2026. doi: 10.64898/2026.04.10.717709.

Riya Singh, Arjun Aggarwal Barsainyan, Ryan Irfan, et al. ChemBERTa-3: an open source training framework for chemical foundation models. *Digital Discovery*, 5:662–685, 2026.

Sanju Sinha, Rahulsimham Vegesna, Sumit Mukherjee, Ashwin V Kammula, Saugato Rahman Dhruba, Wei Wu, D Lucas Kerr, Nishanth Ulhas Nair, Matthew G Jones, Nir Yosef, et al. PERCEPTION predicts patient response and resistance to treatment using single-cell transcriptomics of their tumors. *Nature Cancer*, 5(6):938–952, 2024.

Eduardo Soares, Emilio Vital Brazil, Victor Shirasuna, Dmitry Zubarev, Renato Cerqueira, and Kristin Schmidt. A large encoder-decoder family of foundation models for chemical language. *arXiv preprint arXiv:2407.20267*, 2024.

Chayaporn Suphavilai, Shumei Chia, Ankur Sharma, Lorna Tu, Rafael Peres Da Silva, Aanchal Mongia, Ramanuj DasGupta, and Niranjan Nagarajan. Predicting heterogeneity in clone-specific therapeutic vulnerabilities using single-cell transcriptomic signatures. *Genome Medicine*, 13(1): 189, 2021.

Parthasarathy Suryanarayanan, Yunguang Qiu, Shreyans Sethi, Dhruv Mahajan, Hongyang Li, Yang Yang, Elif Eyigoz, Aldo Guzmán-Sáenz, Daniel E. Platt, and Timothy H. Rumbell. Multi-view biomedical foundation models for molecule-target and property prediction. *Advanced Science*, 2024.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, June 2023. doi: 10.1038/s41586-023-06139-9.

Itay Tirosh, Benjamin Izar, Sanjay M. Prakadan, Marc H. Wadsworth, Daniel Treacy, John J. Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, Mohammad Fallahi-Sichani, Ken Dutton-Regester, Jia-Ren Lin, Ofir Cohen, Parin Shah, Diana Lu, Alex S. Gennert, Anat Shefler, Paulina E. Ghosh, Allison Hanker, Ronald Mosbergen, Phillip Lovelace, David A. Barbie, Genevieve M. Boland, Charles H. Yoon, Min Jiang, F. Stephen Hodi, George F. Murphy, Paul Russell, Kimberly Glass, Levi A. Garraway, Jon C. Aster, Cyril H. Benes, Peter K. Sorger, Sandro Santagata, and Aviv Regev. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196, April 2016. doi: 10.1126/science.aad0501.

Qing Wang, Yining Pan, Minghao Zhou, Zijia Tang, Yanfei Wang, Guangyu Wang, and Qianqian Song. scDrugMap: benchmarking large foundation models for drug response prediction. *Nature Communications*, 17(1):730, 2025. doi: 10.1038/s41467-025-67481-2.

Meisheng Xiao, Qianhui Zheng, Paul Popa, Xinlei Mi, Jianhua Hu, Fei Zou, and Baiming Zou. Drug molecular representations for drug response predictions: a comprehensive investigation via machine learning methods. *Scientific Reports*, 15(1):20, 2025. doi: 10.1038/s41598-024-84711-7.

Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A. Smith, I. Richard Thompson, Sridhar Ramaswamy, P. Andrew Futreal, Daniel A. Haber, Michael R. Stratton, Cyril Benes, Ultan McDermott, and Mathew J. Garnett. Genomics of drug sensitivity in cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1):D955–D961, January 2013. doi: 10.1093/nar/gks1111.

Bulat Zagidullin, Ziyan Wang, Yuanfang Guan, Esa Pitkänen, and Jing Tang. Comparative analysis of molecular fingerprints in prediction of drug combination effects. *Briefings in bioinformatics*, 22(6):bbab291, 2021.

Jesse Zhang, Airol A Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G Jones, Daniele Merico, Nima Alidoust, Hani Goodarzi, and Johnny Yu. Tahoe-100M: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025. doi: 10.1101/2025.02.20.639398.

Zetian Zheng, Junyi Chen, Xingjian Chen, Lei Huang, Weidun Xie, Qiuzhen Lin, Xiangtao Li, and Ka-Chun Wong. Enabling Single-Cell Drug Response Annotations from Bulk RNA-Seq Using SCAD. *Advanced Science*, 10(11):2204113, February 2023. doi: 10.1002/advs.202204113.

Murong Zhou, Zeyu Luo, Yu-Hang Yin, Qiaoming Liu, Guohua Wang, and Yuming Zhao. scATD: A high-throughput and interpretable framework for single-cell cancer drug resistance prediction and biomarker identification. *Briefings in Bioinformatics*, 26(3):bbaf268, May 2025. doi: 10.1093/bib/bbaf268.

A Supplementary Material

B Data characterization

B.1 Shared-label setting

Single-cell data. Tahoe-100M (Zhang et al., 2025) is a single-cell drug perturbation atlas containing more than 1,100 small-molecule perturbations across 50 cancer cell lines derived from multiple organs, including lung (26%), bowel (19%), pancreas (15%), and skin (9%). Cells were profiled from pooled mixtures of cancer cell lines after 24 hours of treatment using combinatorial-barcoding single-cell RNA sequencing. We used the public Tahoe-100M release distributed through Hugging Face (`tahoebio/Tahoe-100M`), specifically the 14 quality-filtered per-plate AnnData files. Only vehicle-control (DMSO) cells are used as model input in this study, so that the cellular representation describes the baseline transcriptional state of each cell line rather than a treatment-induced response. From the per-plate files, we pooled all DMSO control cells of each cell line across plates and drew a uniformly random subset of at most 2,000 cells per cell line, where cell lines with fewer control cells kept all of their cells. Raw counts were library-size normalized and log-transformed (`log1p`), and the gene space was restricted to the 5,000 highly variable genes (`seurat_v3` flavor). The raw count layer alongside the log-normalized layer is the common input to all five cell representations in the shared-label setting: PCA and scVI are fitted on it per fold, and the frozen foundation-model embeddings were computed from the raw counts of the same genes, so that no representation has access to genes the others do not see (Section C.1).

Response labels. Drug-response labels were derived from the Genomics of Drug Sensitivity in Cancer database, GDSC2 (Yang et al., 2013; Iorio et al., 2016). GDSC response measurements are obtained from bulk cell populations using fluorescence-based viability assays after 72 hours of treatment across nine drug concentrations. For each cell line–drug pair we used the viability at the clinically reported maximum plasma concentration (Cmax): the Cmax value of each drug was taken from the literature, and the viability at that concentration was interpolated from the dose–response curve fitted by GDSC2. Cmax viability was binarized per drug into a sensitive class ($y = 1$, low Cmax viability) and a resistant class ($y = 0$) using a two-cluster partition of the drug's Cmax viability values (partitioning around medoids); for 22 of 23 drugs, the resulting classes are separated by a drug-specific threshold close to 0.5, and a global threshold of 0.5 reproduces 946 of 963 labels (the exceptions are concentrated in Dactinomycin). In total, 46 cell lines and 65 drugs are shared between Tahoe-100M and GDSC2; Cmax-based labels were available for 23 of these drugs, giving 963 labeled cell line–drug pairs (class-1 fraction 0.38). Tahoe cell-line names were mapped to GDSC2 names through Cellosaurus accessions (CVCL identifiers). Of the 50 cell lines in the Tahoe control set, 43 have a GDSC2 label; the remaining seven (CVCL_0293, CVCL_0459, CVCL_0480, CVCL_0546, CVCL_1098, CVCL_1517, CVCL_C466) are dropped by the inner join between cells and labels. The full mapping is given in Table 1.

Drug panel. The drug universe is the set of drugs present in every drug-embedding file (23 drugs). For the primary analysis, we further restricted the panel to drugs whose class-1 fraction, measured over the cell lines on which the drug was tested, lies between 0.15 and 0.85. Highly imbalanced drugs contain few observations from one response class and can therefore produce unstable or undefined within-drug MCC estimates. This filter retains 10 drugs (Table 2) and removes 13: Bortezomib (0.98), Epirubicin (0.95), Irinotecan (0.14), Dabrafenib (0.12), Dasatinib (0.12), Erlotinib (0.10), Trametinib (0.10), Venetoclax (0.07), Sirolimus (0.05), Lapatinib (0.05), Olaparib (0.02), Vorinostat (0.02), and Oxaliplatin (0.02), where the parenthetical value is the class-1 fraction. Conclusions from the primary analysis apply to drugs with sufficient representation of both response classes and should not be interpreted as establishing performance for highly imbalanced response distributions. A sensitivity analysis over alternative balance windows (0–1, 0.05–0.95, 0.15–0.85, and 0.25–0.75) motivated the chosen window.

Joining every control cell of a cell line with the labeled drugs of that cell line yields the training table summarized in Table 3. The benchmark comprises 43 cell lines, 10 drugs, 83403 unique cells, and 806030 cell–drug rows corresponding to 416 distinct cell line–drug pairs (416 of the 430 possible combinations, 96.7% coverage). Because all cells of a pair share the same response label, the effective number of independent supervised labels is 416, not the number of cells or rows. Five

Table 1: **Cell-line name harmonisation between Tahoe-100M and GDSC2.** The 43 cell lines retained in the shared-label benchmark, with their Tahoe identifier, GDSC2 identifier, and Cellosaurus accession.

Tahoe	GDSC2	CVCL	Tahoe	GDSC2	CVCL
A172	A-172	CVCL_0131	LOXIMVI	LOX-IMVI	CVCL_1381
A427	A-427	CVCL_1055	LS180	LS180	CVCL_0397
A498	A-498	CVCL_1056	MIAPACA2	MIA PaCa-2	CVCL_0428
A549	A-549	CVCL_0023	NCIH1573	NCI-H1573	CVCL_1478
AN3CA	AN3-CA	CVCL_0028	NCIH1792	NCI-H1792	CVCL_1495
ASPC1	AsPC-1	CVCL_0152	NCIH2122	NCI-H2122	CVCL_1531
BT474	BT-474	CVCL_0179	NCIH23	NCI-H23	CVCL_1547
C32	C32	CVCL_1097	NCIH2347	NCI-H2347	CVCL_1550
C33A	C-33 A	CVCL_1094	NCIH596	NCI-H596	CVCL_1571
CFPAC1	CFPAC-1	CVCL_1119	NCIH661	NCI-H661	CVCL_1577
CHP212	CHP-212	CVCL_1125	PANC0327	Panc 03.27	CVCL_1635
COLO205	COLO 205	CVCL_0218	RKO	RKO	CVCL_0504
H4	H4	CVCL_1239	RPMI7951	RPMI-7951	CVCL_1666
HCT15	HCT 15	CVCL_0292	SHP77	SHP-77	CVCL_1693
HOP62	HOP-62	CVCL_1285	SKMEL2	SK-MEL-2	CVCL_0069
HS578T	Hs 578T	CVCL_0332	SNU1	SNU-1	CVCL_0099
HS766T	Hs 766T	CVCL_0334	SNU423	SNU-423	CVCL_0366
HT29	HT-29	CVCL_0320	SW1088	SW1088	CVCL_1715
J82	J82	CVCL_0359	SW1271	SW1271	CVCL_1716
KATOIII	KATO III	CVCL_0371	SW1417	SW1417	CVCL_1717
LOVO	LoVo	CVCL_0399	SW48	SW48	CVCL_1724
			SW900	SW900	CVCL_1731

Table 2: **The 10 drugs of the shared-label benchmark.** Number of cell lines tested, class-1 (sensitive) fraction over those cell lines, and number of DMSO control cells contributing observations for each drug.

Drug	Cell lines	Class-1 fraction	Cells
5-Fluorouracil	43	0.698	83,403
Cisplatin	42	0.238	81,403
Cyclophosphamide	42	0.214	81,403
Cytarabine	42	0.738	81,403
Dactinomycin	43	0.698	83,403
Gemcitabine	42	0.810	81,403
Mitoxantrone	39	0.333	75,403
Sorafenib	42	0.595	81,403
Teniposide	39	0.846	75,403
Vinorelbine	42	0.833	81,403

cell lines (CVCL_0334, CVCL_1094, CVCL_1239, CVCL_1495, CVCL_1550) are single-class across the 10 retained drugs; their within-cell-line MCC is undefined and they are excluded from the corresponding macro-average.

Table 3: **Size of the shared-label benchmark before and after the drug-balance filter.**

Table	Rows	Cells	Lines	Drugs	Pairs	Class-1 frac.
All 23 labeled drugs	1,866,269	83,403	43	23	963	0.379
Balanced panel (used)	806,030	83,403	43	10	416	0.601

Label structure and chemical diversity. Of the 416 evaluated pairs, 60.1% are labeled sensitive (class 1), though class balance varies substantially across both compounds and cell lines (Fig. 6). At the compound level, the proportion of class-1 responses ranges from 21.4% for Cyclophosphamide ($n = 42$ cell lines) to 84.6% for Teniposide ($n = 39$) (Fig. 7A). At the cell-line level, response profiles range from exclusively class 0 to exclusively class 1 (median 0.7). The 14 missing combinations are concentrated in five cell lines: four lack measurements for two of the ten compounds,

while A-427 lacks measurements for six. The compound panel is chemically diverse (Fig. 7B): pairwise Tanimoto similarities computed from the same 1,024-bit radius-2 Morgan fingerprints used as model input have a median off-diagonal value of 0.08 (minimum 0.00, maximum 0.62). The one exception is the pair Cytarabine–Gemcitabine, two cytidine-analog antimetabolites, with a Tanimoto similarity of 0.62.

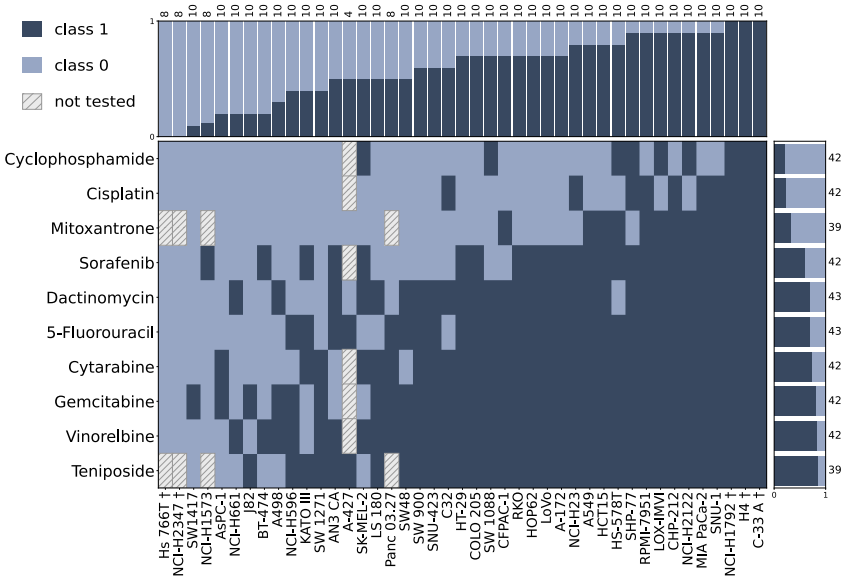

Figure 6: **Shared-label setting: label matrix of the drug panel.** Each cell of the grid is one drug–cell line pair, colored by its binary sensitivity label (class 1 = sensitive). Hatched cells are untested pairs. Rows and columns are ordered by their class-1 fraction, also indicated by the bars along the edges. The top bar chart gives that fraction for each cell line, with the number of drugs tested on it printed above; the right-hand bar chart gives the fraction for each drug, with the number of cell lines it was tested on. Cell lines marked with † are single-class, every tested pair carrying the same label.

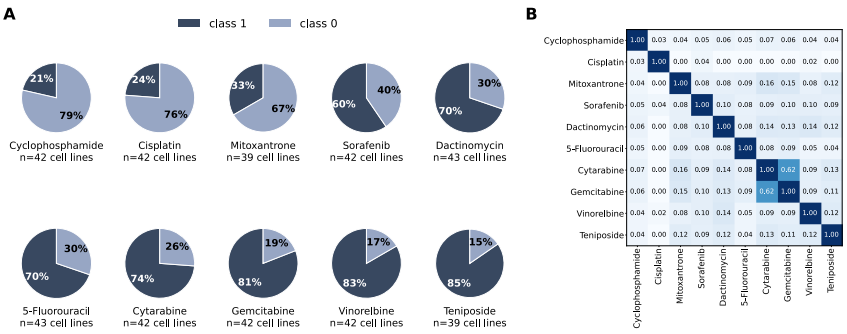

Figure 7: **Response-label balance and chemical diversity of the shared-label drug panel.** (A) Distribution of binary response labels across the ten compounds retained for analysis. Pie charts show the proportion of class 1 (sensitive) and class 0 (resistant) cell-line responses for each drug, with the number of evaluated cell lines indicated below each compound. (B) Pairwise Tanimoto similarity matrix calculated from 1,024-bit radius-2 Morgan fingerprints for the same compounds. Most drug pairs show low structural similarity, with Cytarabine and Gemcitabine forming the most similar pair (Tanimoto similarity 0.62).

B.2 CELL-SPECIFIC-LABEL SETTING

Source and composition. We assembled 19 single-cell RNA-seq drug-response datasets spanning 10 compounds and a diverse set of biological systems, including cancer cell lines, patient-derived cell lines, cell-line-derived xenografts, and primary chronic lymphocytic leukemia (CLL) samples (Table 4). The datasets were downloaded from the Gene Expression Omnibus (GEO) and the Broad Institute Single Cell Portal under the accessions listed in the table, and comprise the datasets used in the original publications of SCAD (Zheng et al., 2023), scDEAL (Chen et al., 2022), SSDA4Drug (Huang & Liu, 2025), and scATD (Zhou et al., 2025) for which the drug also has a GDSC-derived Cmax viability label. Unlike those studies, we do not use the bulk-to-single-cell transfer objective; the datasets serve only as single-cell inputs with binary per-cell sensitivity labels. Each dataset is used as an expression matrix and a binary per-cell label, with label derivations following the original studies as summarized below.

The datasets differ substantially in how drug sensitivity and resistance were defined. Most publicly available single-cell drug-response datasets, including those used in prior benchmarking studies, assign labels based on treatment status: untreated cells are labeled as sensitive, while cells surviving drug exposure are labeled as resistant. This design ignores pre-existing resistant subclones in the untreated population and conflates intrinsic resistance with treatment-induced transcriptional changes, so that treatment status, rather than drug susceptibility, often dominates the separation between the two classes. The SCC47 and JHU006 data (SCP542), used in the SCAD and SSDA4Drug publications, are instead labeled according to the epithelial senescence-related (EpiSen) program (Kinker et al., 2020): cells with EpiSen scores in the top 10% are labeled EpiSen-high and those in the bottom 10% EpiSen-low, and the remaining 80% of cells are excluded, which artificially amplifies the class separation. To counteract such biases, we also included datasets in which resistance was inferred independently of treatment survival. GSE223003 (Olaparib, KURAMOCHI) and GSE228154 (Afatinib, MDA-MB-468) use lineage tracing, which tags cells with DNA barcodes and links pre-treatment transcriptional states to eventual survival, so that labels are assigned to cells before drug exposure. We additionally included post-treatment xenograft data with sensitive and resistant samples for Cisplatin (GSE138267, SCLC) and Sorafenib (GSE175716, HCC), and pre-treatment PBMCs from CLL patients treated with Ibrutinib (GSE111014). GSE163836 pools the paclitaxel-sensitive parental line FCIBC02 with its resistant derivative FCIBC02PR. Across datasets, transcript abundance was represented predominantly by UMI counts; GSE112274 was quantified using full-length read counts, and the three Cisplatin datasets (GSE117872, GSE138267) are represented by TPM-like normalized values.

Table 4: **The 19 datasets of the cell-specific-label benchmark.** One row per (drug, dataset) file. "Unit" is the cell-line unit used for within-cell-line evaluation and cell-line zero-shot splitting; cell counts are after preprocessing. Labels: 1 = sensitive, 0 = resistant. Quantification: UMI counts, full-length read counts (reads), or TPM-like normalized values (TPM).

Drug	Accession	Unit	Model system	Cells	Sens.	Res.	Quant.
Afatinib	GSE228154	MDAMB468	MDA-MB-468 breast line (lineage tracing)	1,512	670	842	UMI
Afatinib	SCP542	SCC47	UM-SCC-47 HNSCC line (EpiSen)	120	60	60	UMI
Cisplatin	GSE117872	HN120	HN120 patient-derived OSCC line	360	181	179	TPM
Cisplatin	GSE117872	HN137	HN137 patient-derived OSCC line	403	328	75	TPM
Cisplatin	GSE138267	SC	SCLC xenografts (four CDX models pooled)	20,053	17,595	2,458	TPM
Erlotinib	GSE149383	PC9	PC9 NSCLC line	2,541	635	1,906	UMI
Etoposide	GSE149383	PC9	PC9 NSCLC line	1,064	638	426	UMI
Gefitinib	GSE112274	PC9	PC9 NSCLC line	495	37	458	reads
Gefitinib	GSE162045	PC9	PC9 NSCLC line	4,020	2,318	1,702	UMI
Gefitinib	GSE202234	H1975	NCI-H1975 NSCLC line	760	405	355	UMI
Gefitinib	GSE202234	PC9	PC9 NSCLC line	597	290	307	UMI
Gefitinib	SCP542	JHU006	JHU-006 HNSCC line (EpiSen)	66	33	33	UMI
Ibrutinib	GSE111014	CLL	PBMCs of four CLL patients (pre-treatment)	12,454	7,274	5,180	UMI
Olaparib	GSE223003	KURAMOCHI	KURAMOCHI ovarian line (lineage tracing)	9,215	8,313	902	UMI
Paclitaxel	GSE131984	SUM159	SUM159 breast line (all resistant)	734	0	734	UMI
Paclitaxel	GSE163836	FCIBC02	FCIBC02 + FCIBC02PR IBC lines	2,608	1,844	764	UMI
Sorafenib	GSE175716	HUH7	HCC xenograft (HUH-7)	4,321	4,010	311	UMI
Sorafenib	SCP542	SCC47	UM-SCC-47 HNSCC line (EpiSen)	120	60	60	UMI
Vorinostat	SCP542	JHU006	JHU-006 HNSCC line (EpiSen)	66	33	33	UMI
Total	16 sources	13 units	60,688 cells	61,509	44,724	16,785	

Preprocessing. Gene identifiers were harmonised across datasets by mapping gene symbols to Ensembl identifiers (and vice versa) with `mygene` (version 3.2.2); ambiguous mappings were resolved deterministically to the first sorted candidate, version suffixes were stripped from Ensembl identifiers, and pseudo-autosomal chrY identifiers were dropped. Four datasets were distributed in Ensembl space and 15 in symbol space; the Ensembl space gave the larger common panel, and the benchmark uses the intersection of **3,885 genes** present in all 19 datasets. Columns mapping to the same target gene were summed. All 19 files were distributed as counts-per-million normalized values (median row sum 10^6); we applied a `log1p` transform, so that the final matrix is $\log(1 + \mathrm{CPM})$. Raw integer counts are not recoverable from the distributed files, which is why scVI uses a Gaussian likelihood in this setting (Section C.1). No additional quality-control filter, scaling, highly-variable-gene subsetting, or batch correction was applied at the dataset-build stage: datasets are concatenated on the shared gene panel, and scVI is fitted without a batch covariate.

Cell identity across datasets and evaluation units. Each cell is identified by its source file and barcode. When the same barcode occurs under a second drug from the same source (Etoposide and Erlotinib in GSE149383; Afatinib and Sorafenib in SCC47; Gefitinib and Vorinostat in JHU006), it is treated as the same physical cell and collapsed to a single observation with two drug rows. The final dataset therefore contains 61,509 (cell, drug) rows over 60,688 unique cells, of which 821 appear under more than one drug. For the 635 cells shared between the Erlotinib and Etoposide files of GSE149383, whose distributed expression values differ slightly between the two files, the Erlotinib values are used. Datasets are grouped into 13 cell-line units by the biological system they profile: the four PC9 studies form a single PC9 unit, the four CLL patients form a single CLL unit, the four SCLC CDX models form a single SC unit, and FCIBC02 with FCIBC02PR a single FCIBC02 unit. These 13 units define the groups for within-cell-line MCC and for the cell-line zero-shot split, so that holding out PC9 removes all four PC9 studies from training. The drug–unit coverage is sparse: only three units are observed with more than one drug (PC9 with Erlotinib, Etoposide, and Gefitinib; SCC47 with Afatinib and Sorafenib; JHU006 with Gefitinib and Vorinostat), giving 17 (unit, drug) pairs, 16 of which contain both response classes. Because labels are cell-specific, a pair-level majority vote would discard the within-pair heterogeneity; predictions in this setting are therefore scored at the cell level (Section D.1).

Class balance. Overall, 72.7% of rows are labeled sensitive. Several units are heavily imbalanced: SUM159 has no sensitive cells (0%), HUH7 has 92.8% sensitive cells, KURAMOCHI 90.2%, and SC 87.7%. Vorinostat is represented by a single 66-cell dataset. The single-class SUM159 unit yields an undefined within-cell-line MCC and is excluded from that macro-average; it still contributes training cells and enters within-drug evaluation for Paclitaxel together with FCIBC02. Figures 8 and 9 summarize the label matrix and the drug panel.

C Representations and predictive models

C.1 Cell representations

Dataset-fitted representations (PCA and scVI). PCA and scVI are refit inside every cross-validation fold using only the training cells of that fold; the fitted transformation is then applied to all cells, so that no held-out cell influences gene selection or model fitting. Fits are cached on disk keyed by the exact set of training cells, so that the same fold shares one fit across drug-representation and predictive-model configurations. Settings are summarized in Table 5. For PCA, the 3,000 most variable genes (by variance of the log-normalized values over the training cells) are selected, and 50 components are fitted with randomized SVD (`scikit-learn` 1.8.0); no scaling is applied before PCA. For scVI (Gayoso et al., 2022) (`scvi-tools` 1.4.2), the input differs between settings because raw counts exist only for Tahoe: in the shared-label setting, the model uses raw counts of 2,000 `seurat_v3` highly variable genes selected on at most 5,000 sampled training cells with a zero-inflated negative binomial likelihood, whereas in the cell-specific-label setting it uses the $\log(1 + \mathrm{CPM})$ values of all 3,885 genes with a Gaussian likelihood and no additional log transform. In both settings, the latent dimension is 50, the encoder and decoder use the `scvi-tools` defaults (one hidden layer of 128 units); training runs for at most 100 epochs with a batch size of 512 and early stopping on the validation ELBO (patience 10); no batch covariate is provided and the random seed of the fold is used for initialization.

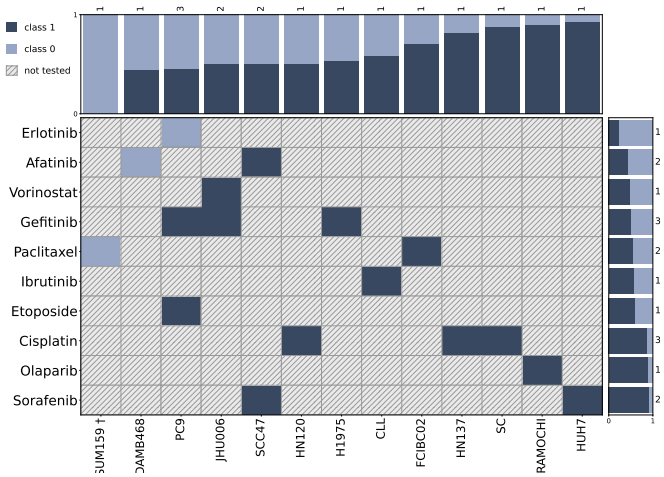

Figure 8: **Cell-specific-label setting: label matrix of the drug panel.** Each cell of the grid is one drug–cell-line-unit pair, colored by the majority label of its cells (class 1 = sensitive). Hatched cells are untested pairs. Rows and columns are ordered by their class-1 fraction, also indicated by the bars along the edges. The top bar chart gives that fraction for each unit, with the number of drugs tested on it printed above; the right-hand bar chart gives the fraction for each drug, with the number of units it was tested on. Units marked with † are single-class.

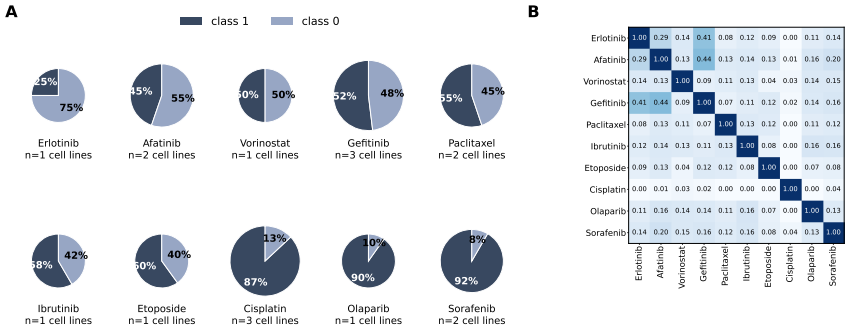

Figure 9: **Response-label balance and chemical diversity of the cell-specific-label drug panel.** (A) Distribution of binary response labels across the ten compounds. Pie charts show the proportion of class 1 (sensitive) and class 0 (resistant) cells for each drug, with the number of cell-line units indicated below each compound. (B) Pairwise Tanimoto similarity matrix calculated from 1,024-bit radius-2 Morgan fingerprints for the same compounds. Most drug pairs show low structural similarity; the exceptions are the three quinazoline EGFR inhibitors, with Afatinib–Gefitinib (0.44), Erlotinib–Gefitinib (0.41), and Erlotinib–Afatinib (0.29) forming the most similar pairs. Cisplatin, an inorganic platinum complex, shares essentially no substructure bits with any other compound.

Frozen pretrained representations (GeneJEPA, Geneformer, scGPT). The three foundation-model representations were computed once per setting with fixed pretrained weights, in inference mode on a single GPU, and are identical across folds and seeds. Embeddings were computed per cell, separately for the Tahoe control cells (from the raw counts of the 5,000-HVG matrix described in Section B.1) and for each cell-specific dataset (from the distributed $\log(1 + \mathrm{CPM})$ values), and the models were not fine-tuned. Gene identifiers were mapped to each model's vocabulary with `mygene` where the input namespace differed from the model's (symbols to Ensembl identifiers for Geneformer and GeneJEPA, Ensembl identifiers to symbols for scGPT), Ensembl version suffixes were stripped, columns mapping to the same vocabulary entry were summed for raw counts and averaged for normalized values, and genes absent from the model vocabulary were dropped. In the

Table 5: **Settings of the dataset-fitted cell representations.** Both are refit on the training cells of each fold.

	PCA	scVI (shared-label)	scVI (cell-specific)
Input layer	log-normalized	raw counts	$\log(1 + \mathrm{CPM})$
Gene selection	top-3,000 by variance (train cells)	2,000 `seurat_v3` HVGs (train cells)	all 3,885 genes
Likelihood	–	zero-inflated negative binomial	Gaussian
Output dimension	50	50	50
Solver / training	randomized SVD	$\leqslant$100 epochs, batch 512, early stopping	same
Batch covariate	–	none	none
Scaling before fit	none	none	none

shared-label setting, the foundation models therefore receive exactly the same 5,000 highly variable genes from which PCA and scVI select their genes, so that all five cell representations are computed from the same input matrix and the comparison is not confounded by differences in gene coverage.

Geneformer (Theodoris et al., 2023): we used the V2 104M-parameter model with the cancer-domain continual-pretraining checkpoint (`Geneformer-V2-104M_CLcancer`, the "pan-cancer" checkpoint referred to in the main text) and its 104M token dictionary. Each cell is encoded by rank-value encoding: genes with non-zero expression are sorted by decreasing expression (raw counts for Tahoe; per-cell min-shifted log-normalized values for the cell-specific datasets), truncated to the model's maximum input length of 4,096 genes, and padded. The 768-dimensional cell embedding is the attention-mask-aware mean of the last-layer hidden states over the gene tokens (batch size 16).

scGPT (Cui et al., 2024): we used the released pan-cancer checkpoint (pretrained on 5.7 million cancer cells) through the `scgpt.tasks.embed_data` interface with gene symbols as tokens (batch size 64). For Tahoe, raw counts were normalized to 10^4 counts per cell and log-transformed before tokenization, matching the model's pretraining preprocessing; for the cell-specific datasets the log-normalized values were min-shifted per cell. The 512-dimensional cell embedding is the model's `<cls>` cell token output.

GeneJEPA (Litman et al., 2025): we used the public release on Hugging Face (`elonlit/GeneJEPA`, checkpoint `genejepa-epoch=49.ckpt`, 62,710-gene Ensembl vocabulary). Only non-zero genes of each cell are passed to the model as a ragged (token, value) set; for Tahoe the values are $\log(1 + x)$ counts standardized with the model's global mean and standard deviation, whereas the cell-specific datasets' log-normalized values are passed unchanged. The 768-dimensional embedding is obtained from the teacher encoder (`get_embedding` with `use_teacher=True`, batch size 32), as recommended by the authors.

No cell representation is standardized before concatenation with the drug representation; the only standardization is performed inside xRFM (Section C.3).

C.2 Drug representations

All drug representations are computed once from canonical SMILES strings retrieved from PubChem by compound identifier, and kept fixed. Morgan fingerprints are 1,024-bit binary fingerprints of radius 2 computed with RDKit (2026.03.3) using the `rdFingerprintGenerator` Morgan generator, without chirality; MACCS keys are the standard 167-bit RDKit MACCS vector (Rogers & Hahn, 2010; Durant et al., 2002). The four pretrained molecular encoders are used for feature extraction only, in inference mode on a single GPU. ChemBERTa (Singh et al., 2026) is a RoBERTa-style transformer pretrained on SMILES; we used the `DeepChem/ChemBERTa-77M-MTR` checkpoint (77M-molecule pretraining corpus, multi-task-regression objective, 384-dimensional hidden size) and took the last-layer hidden state of the first (CLS) token as the molecule embedding. MoLFormer (Ross et al., 2022) is a linear-attention transformer with rotary positional embeddings pretrained on approximately 1.1 billion molecules from PubChem and ZINC; we used `ibm/MoLFormer-XL-both-10pct` in deterministic evaluation mode and took its pooler output (768 dimensions). BioMed (Suryanarayanan et al., 2024) is a multi-view model combining SMILES, molecular-graph, and molecular-image encoders; we used `ibm/biomed.sm.mv-te-84m`

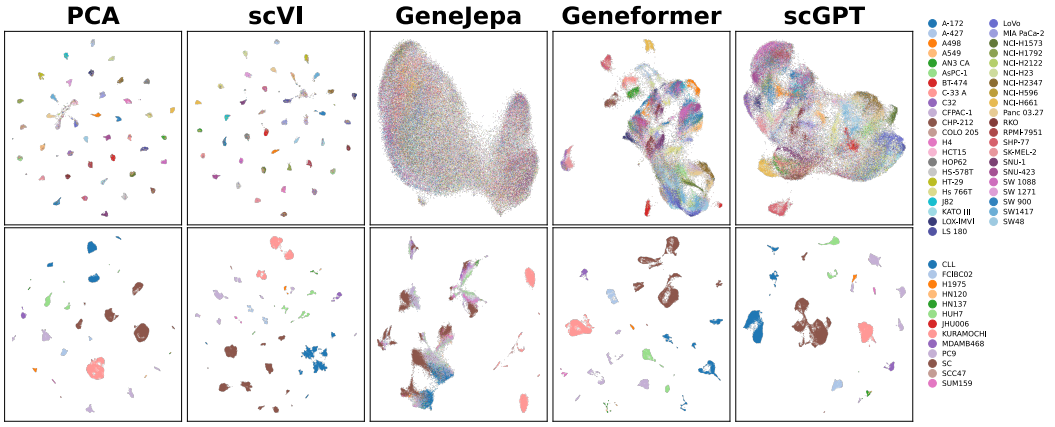

Figure 10: **UMAP projections of the five cell representations, colored by cell line of origin.** Top: shared-label setting. Bottom: cell-specific-label setting.

through the `SmallMoleculeMultiViewModel.get_embeddings` interface of the `biomed-multi-view` repository, which returns the fused 512-dimensional representation. SMI-TED is a transformer encoder–decoder pretrained on approximately 91 million curated PubChem SMILES; we used the SMI-TED-Light model from the IBM `materials` repository (checkpoint `smi-ted-Light_40.pt`) and its `encode` output (768 dimensions). In the result files and some figure sources the SMI-TED representation is labeled `materials`. Table 6 gives the resulting input dimensionality of every cell–drug combination.

Table 6: **Dimensionality of the concatenated input for each cell–drug representation pair.** Cell block first, drug block second.

Cell \ drug	Morgan (1,024)	BioMed (512)	ChemBERTa (384)	MACCS (167)	SMI-TED (768)	MoLFormer (768)
PCA (50)	1,074	562	434	217	818	818
scVI (50)	1,074	562	434	217	818	818
scGPT (512)	1,536	1,024	896	679	1,280	1,280
GeneJEPA (768)	1,792	1,280	1,152	935	1,536	1,536
Geneformer (768)	1,792	1,280	1,152	935	1,536	1,536

C.3 Predictive models

All three predictive models are trained on the concatenated [cell | drug] feature vector of each cell–drug row, jointly across all drugs and cell lines, with the binary sensitivity label as target. No hyperparameter search was performed: every configuration of a setting uses the same fixed settings listed in Table 7, so that differences between representations are not confounded by per-configuration tuning. If the training rows of a fold contain a single class, no model is fitted and that class is predicted as a constant for every held-out row. Three random seeds (0, 1, 42) are used per configuration; each seed re-draws the fold assignment of the shuffled splitters (random and pair zero-shot), the PCA/scVI fits, the model initialization, and xRFM's internal validation split. For the leave-one-group-out splits the folds are fixed and only the fits vary across seeds.

The MLP and LightGBM settings are identical in the two supervision settings. The xRFM tree settings differ: the shared-label grid was run with 2 AGOP iterations and a maximum leaf size of 5,000 to fit within the available compute for the 806,030-row training tables, whereas the cell-specific-label grid (61,509 rows) used 5 iterations and a maximum leaf size of 20,000. Because every xRFM configuration within a setting shares the same values, comparisons across representations within a setting are unaffected, but absolute xRFM scores should not be compared across the two settings.

Table 7: **Predictive-model hyperparameters.** Values are shared by all 90 representation configurations of a setting. Where a value differs between settings it is given as shared-label / cell-specific.

MLP (PyTorch 2.11.0)	
Architecture	input → 256 → 64 → 1 (logit); each hidden layer: Linear → BatchNorm1d → ReLU → Dropout(0.2)
Loss / optimizer	binary cross-entropy with logits, no class weighting; Adam, learning rate 10^{-3}, no weight decay
Training	batch size 4,096; 8 epochs, fixed; no validation set and no early stopping
Input scaling	none (batch normalization inside the network)
LightGBM (4.6.0, CPU)	
Trees	`n_estimators`=1,000, `learning_rate`=0.05, `num_leaves`=31, `max_depth`=−1, `min_child_samples`=20
Sampling / regularization	`subsample`=0.9 (`subsample_freq`=1), `colsample_bytree`=0.8, `max_bin`=63, `reg_lambda`=1.0, `reg_alpha`=0
Objective	binary; no class weighting; no early stopping (all 1,000 trees are grown)
Input scaling	none; zero-variance columns of the training fold are dropped
xRFM (`xrfm` 0.4.4, GPU)	
Kernel	ℓ_2 kernel, bandwidth 1.0 (adaptive mode), exponent 1.0, ridge regularization 10^{-3}, full (non-diagonal) feature matrix M
Tree / AGOP	split method `top_vector_agop_on_subset`, one tree; AGOP iterations 2 / 5; maximum leaf size 5,000 / 20,000; AGOP subset capped at 20,000 rows in the shared-label runs
Tuning	early stopping of the RFM iterations on an internal stratified 15% validation split of the training fold (`StratifiedShuffleSplit`, seeded), tuning metric AUROC; this internal split is not group-aware and is used only for model selection inside the training fold, never for reporting
Input scaling	joint per-feature z-scoring of the concatenated vector using training-fold means and standard deviations (zero-variance columns are given unit scale)

D Evaluation

D.1 Performance assessment

The evaluation unit differs between the two supervision settings because labels are defined at different levels.

Shared-label setting. All cells from a cell-line–drug pair share the same GDSC-derived label. We therefore aggregate held-out cell-level predictions to the pair level before scoring. For cell line c, drug d, and random seed s,

$$\bar{p}^{(s)}_{c,d} = \frac{1}{|\mathcal{B}^{(s)}_{c,d}|} \sum_{b \in \mathcal{B}^{(s)}_{c,d}} p^{(s)}_{b,d}, \qquad \hat{y}^{(s)}_{c,d} = \mathbf{1}\left[\bar{p}^{(s)}_{c,d} \geqslant 0.5\right], \tag{1}$$

where $\mathcal{B}^{(s)}_{c,d}$ denotes the held-out cells belonging to pair (c, d) pooled over all folds of seed s, $p^{(s)}_{b,d}$ is the predicted probability for cell b and drug d, and $y_{c,d}$ is the corresponding pair-level ground-truth label. Thus, each cell-line–drug pair contributes one prediction to evaluation. (The implementation defines the pair label as the majority vote of its cells' labels; in the shared-label setting every pair is label-pure, so this coincides with the GDSC label.)

Cell-specific-label setting. Here, response labels are defined for individual cellular observations. Predictions are therefore evaluated directly without pair-level averaging:

$$\hat{y}^{(s)}_{b,d} = \mathbf{1}\left[p^{(s)}_{b,d} \geqslant 0.5\right], \tag{2}$$

with corresponding cell-specific label $y_{b,d}$. Pair-level aggregation is not meaningful here because only 17 (unit, drug) pairs exist and 16 of them contain both classes.

For both settings, Matthews correlation coefficient (MCC) is computed after thresholding predicted probabilities at 0.5, without test-set threshold optimization. We report two complementary aggregation schemes: *within-drug MCC*, measuring discrimination for a fixed drug, and *within-cell-line MCC*, measuring discrimination across drugs for a fixed cell line. The observations entering these metrics differ between settings:

$$\text{Shared-label:} \quad \mathrm{MCC}_d^{(s)} = \mathrm{MCC}\Big(\{y_{c,d}, \hat{y}_{c,d}^{(s)}\}_c\Big), \quad \mathrm{MCC}_c^{(s)} = \mathrm{MCC}\Big(\{y_{c,d}, \hat{y}_{c,d}^{(s)}\}_d\Big), \tag{3}$$

$$\text{Cell-specific:} \quad \mathrm{MCC}_d^{(s)} = \mathrm{MCC}\Big(\{y_{b,d}, \hat{y}_{b,d}^{(s)}\}_b\Big), \quad \mathrm{MCC}_c^{(s)} = \mathrm{MCC}\Big(\{y_{b,d}, \hat{y}_{b,d}^{(s)}\}_{b,d:\, b\in c}\Big). \tag{4}$$

Per-group scores are macro-averaged within each seed,

$$\mathrm{MCC}_{\text{within-drug}}^{(s)} = \frac{1}{|\mathcal{D}^*|} \sum_{d\in\mathcal{D}^*} \mathrm{MCC}_d^{(s)}, \qquad \mathrm{MCC}_{\text{within-cell}}^{(s)} = \frac{1}{|\mathcal{C}^*|} \sum_{c\in\mathcal{C}^*} \mathrm{MCC}_c^{(s)}, \tag{5}$$

where $\mathcal{D}^*$ and $\mathcal{C}^*$ contain only drugs and cell lines for which both response classes are present in the held-out set (a single-class group yields an undefined score, which is dropped). Final reported scores are the mean of the three per-seed macro-averages, and the reported standard error is the standard deviation across seeds divided by $\sqrt{3}$. Per-group scores are computed once per seed on the predictions pooled over all folds of that seed, rather than per fold, so that every group is scored on all of its held-out observations. AUROC is computed with the same grouping and pooling from the untresholded probabilities; area under the precision–recall curve and balanced accuracy are also stored for every run but are not reported. All metrics use `scikit-learn` 1.8.0.

Representation-level ranking summary. To summarize cell-representation performance across drug representations and predictive models, we rank all $5 \times 6 \times 3 = 90$ configurations within each split and evaluation convention using their MCC averaged across the three seeds. Ranks are converted to normalized scores $1 - r/90$, with higher values indicating better relative performance; ties receive the average rank. Each cell representation appears in 18 configurations (6 drug representations $\times$ 3 models), and the plotted value is the mean normalized rank across these 18 configurations, which equals the average fraction of the 90 configurations that the representation's configurations outperform (the best possible value is $1-1/90 \approx 0.989$ because a configuration does not outperform itself). Error bars show the standard error across the 18 values ($\mathrm{SD}/\sqrt{18}$) and therefore reflect consistency across drug-representation/model combinations rather than seed-to-seed variability. Configurations with undefined MCC for some groups still receive a rank from their remaining groups; all 90 configurations are always ranked. Rankings aggregated over splits exclude the random split.

D.2 Generalizability

We evaluate the models under splitting strategies designed to probe distinct forms of train–test overlap and generalization: a random observation-level split, pair-level splits that evaluate unseen combinations of previously observed entities with and without control for cell-level overlap, and entity-level splits that evaluate generalization to entirely unseen cell lines or drugs. All splits operate on the cell–drug row table, and the same code path is used in both settings. Table 8 summarizes the splitters.

Table 8: **Splitting strategies.** Folds are regenerated for each seed by the shuffled splitters; the leave-one-group-out splits are deterministic.

Split	Splitter	Grouping	Stratified	Folds (shared / cell-specific)
Random	`StratifiedKFold`	none	label	3 / 3
Pair ZS (with leakage)	`StratifiedGroupKFold`	(cell line, drug) pair	label	10 / 10
Pair ZS (leakage-free)	pair fold × cell fold	pair and cell	label (pair fold)	5 / 5
Cell-line ZS	`LeaveOneGroupOut`	cell line	–	43 / 13
Drug ZS	`LeaveOneGroupOut`	drug	–	10 / 10

- In the *random split*, individual cell–drug rows are assigned to training and test sets independently of cell line–drug pair identity, with a stratified 3-fold partition (stratified on the sensitivity label). Every fold trains on 66.7% of rows and evaluates on the remaining 33.3%. As a result, cells belonging to the same pair, and therefore sharing the same response label, can occur in both sets. This represents the least restrictive evaluation setting and primarily measures interpolation among previously observed cell line–drug combinations rather than generalization to unseen pairs or entities.
- In the *pair zero-shot split with leakage*, entire cell line–drug pairs are held out from training using a stratified group 10-fold partition over pairs (90%/10% train/test), such that the specific combination evaluated at test time has not been observed during training, while the constituent cell line and drug may still appear in the training set in other combinations. Because splitting is performed at the pair level only, the same cell can occur in training under one drug and in the test set under another drug. In the shared-label setting every cell appears under all of its cell line's drugs, so this overlap is pervasive; in the cell-specific-label setting it affects the 821 cells observed under more than one drug.
- The *leakage-free pair zero-shot split* additionally blocks individual cells. We construct two independent partitions: a stratified group 5-fold partition over cell line–drug pairs and a separate 5-fold partition over unique cells. For fold f, a row is assigned to the test set only when both its pair and its cell are assigned to f, and to the training set only when neither is assigned to f; rows that satisfy only one of these conditions are excluded from that fold. This two-way blocking ensures that both the evaluated pairs and the evaluated cells are disjoint from the training data, while the constituent cell lines and drugs remain represented through other pairs and cells. With $k = 5$, each fold trains on approximately $(4/5)^2 = 64\%$ of rows, tests on approximately 4%, and discards approximately 32%; pooled over the five folds, each pair is evaluated at most once and roughly one fifth of the pairs (about 83 of 416 in the shared-label setting) are evaluated. The fold count of 5 (rather than 10) keeps the number of evaluated pairs per drug large enough for within-drug MCC.
- The *cell-line zero-shot* setting uses leave-one-cell-line-out splitting (43 folds in the shared-label setting, 13 cell-line units in the cell-specific-label setting), such that all rows of the held-out cell line are excluded from training.
- The *drug zero-shot* setting uses leave-one-drug-out splitting (10 folds in both settings), such that all rows of the held-out drug are excluded from training.

Both pair zero-shot variants were run for all 90 configurations in both settings. In the shared-label setting the main text reports the leakage-free variant for the configuration-level comparisons including the cross-setting figure that shows the four shared splits (Fig. 4), whose shared-label pair zero-shot bar is the leakage-free variant; in the cell-specific-label setting, "pair zero-shot" refers to the with-leakage variant throughout, and the single-modality ablations were not run for the leakage-free variant. The effect of cell-level leakage is quantified in Section F.2.

Folds are constructed before any dataset-fitted representation is computed. Within each fold, the training-cell set is derived from the training rows, the gene selection and PCA or scVI fit are performed on those cells only, and the fitted transformation is applied to all cells. Frozen foundation-model and drug embeddings are loaded once and are fold-independent. The only validation data carved from a training fold is xRFM's internal 15% split (Table 7) and scVI's internal validation cells for early stopping; neither is group-aware, and neither is used for reporting. In the case of a fold whose training rows consist of a single class, no model is fitted and that class is predicted as a constant.

D.3 Ablations and prevalence baselines

In the cell-only ablation the drug feature block is removed from the input entirely (zero width, not zero-filled), so that the model receives only the cell representation; the drug-only ablation is symmetric. Model hyperparameters, splits, folds, and seeds are identical to the full runs. Because the removed block does not influence the model, cell-only runs are indexed by cell representation only (5 per split and model) and drug-only runs by drug representation only (6 per split and model). In the shared-label setting the ablations were run for all five splits; in the cell-specific-label setting for the random, pair zero-shot (with leakage), cell-line zero-shot, and drug zero-shot splits.

Prevalence baselines predict, for every held-out row, a constant probability equal to a positive rate estimated from the training rows of the same fold, without fitting any model: the *drug prevalence* baseline uses the positive rate of the row's drug, the *cell-line prevalence* baseline the positive rate of the row's cell line, and a *global prior* baseline the overall training positive rate. Baseline probabilities pass through the same pair-level aggregation, thresholding, and macro-averaging as model predictions. A baseline is uninformative by construction whenever its prediction is constant within the group being scored: the drug prevalence baseline under within-drug MCC, the cell-line prevalence baseline under within-cell-line MCC, and the global prior under both conventions all yield MCC equal to zero. When the held-out group is absent from training (the held-out cell line under cell-line zero-shot, the held-out drug under drug zero-shot), the group rate is undefined and the baseline falls back to the global prior; these cases are marked "null undefined" in the figures. The informative reference for within-drug MCC is therefore the cell-line prevalence baseline, and for within-cell-line MCC the drug prevalence baseline; Table 11 lists their values.

E Compute and software

All benchmark experiments were run in a single Python 3.11.6 environment: PyTorch 2.11.0 (CUDA 13.0), NumPy 2.2.6, pandas 2.3.3, scikit-learn 1.8.0, LightGBM 4.6.0, `xrfm` 0.4.4, scvi-tools 1.4.2, Lightning 2.6.5, scanpy 1.11.5, anndata 0.12.18, RDKit 2026.03.3, and umap-learn 0.5.12. Shared-label grid jobs ran on CPU nodes with 16 cores and 64 GB of memory (time limit 47 h); cell-specific-label grid jobs on 4 cores and 32 GB (8 h). xRFM training, the cell-specific scVI fits, and the rebuilt shared-label xRFM runs used one NVIDIA RTX 3090 (24 GB) GPU per job; MLP and LightGBM runs were CPU-only. Per-fold PCA and scVI fits are cached on disk and shared across configurations that use the same training cells. The full benchmark comprises 1,197 runs (900 grid runs of 5 splits $\times$ 90 configurations across the two settings, plus 165 shared-label and 132 cell-specific ablation runs), each with three seeds, corresponding to approximately 40,800 individual model fits (71 folds per configuration and seed in the shared-label setting, 41 in the cell-specific-label setting). Single xRFM folds on the shared-label data required roughly 1–3 h on CPU before the runs were moved to GPUs, which motivated the reduced AGOP settings in that setting (Section C.3). Foundation-model embeddings for cells and drugs were extracted beforehand in a separate PyTorch 2.11 environment containing the model-specific packages (`transformers` for Geneformer, ChemBERTa, and MoLFormer; `scgpt`; `genejepa`; the IBM `materials` and `biomed-multi-view` repositories), one job per model and dataset on a single NVIDIA GPU, and stored as fixed embedding tables that the benchmark only reads.

F Additional results

F.1 Best configuration per split and evaluation convention

Table 9 gives, for each setting, split, and evaluation convention, the mean MCC over the 90 configurations, its minimum and maximum, and the best configuration with its seed-level standard error. Under the random split, near-perfect scores are reached in the shared-label setting because cells of the same labeled pair occur in both training and test sets; we regard this split as a positive control. The leave-one-group-out splits use fixed folds, so the seed-level standard error of deterministic models (for example xRFM) can be zero.

F.2 Effect of cell-level leakage in the pair zero-shot split

Table 10 compares each configuration's MCC under the with-leakage and leakage-free pair zero-shot splits. In the shared-label setting, where every cell of a held-out pair is present in training under the other nine drugs, allowing this overlap improves within-drug MCC for 88 of 90 configurations (mean $+0.051$) and within-cell-line MCC for 65 of 90 (mean $+0.009$); the best achievable scores change little (0.370 versus 0.363 within-drug; 0.552 versus 0.542 within-cell-line). In the cell-specific-label setting, where only 821 cells occur under more than one drug, the two variants are statistically indistinguishable at the grid level. Note that the two variants also differ in the number of folds and hence in training-set size (90% versus 64% of rows per fold), so the difference is an upper bound on the effect of cell-level leakage alone.

Table 9: **Grid mean (min, max) MCC and best configuration for every setting, split, and evaluation convention.** Mean, minimum, and maximum are over the 90 cell-representation × drug-representation × model configurations; the best configuration is given with its mean ± SE over three seeds.

Setting	Split	**Within-drug MCC**	**Within-cell-line MCC**
Shared-label	Random	0.883 (0.515, 0.973) Geneformer+SMI-TED+MLP, 0.973 ± 0.001	0.910 (0.675, 0.976) Geneformer+SMI-TED+MLP, 0.976 ± 0.001
	Pair ZS (with leakage)	0.233 (0.057, 0.370) PCA+MACCS+LightGBM, 0.370 ± 0.023	0.413 (0.283, 0.552) GeneJEPA+SMI-TED+xRFM, 0.552 ± 0.007
	Pair ZS (leakage-free)	0.182 (0.033, 0.363) PCA+MACCS+LightGBM, 0.363 ± 0.020	0.404 (0.301, 0.542) GeneJEPA+ChemBERTa+xRFM, 0.542 ± 0.019
	Cell-line ZS	-0.027 (-0.077, 0.052) Geneformer+BioMed+MLP, 0.052 ± 0.031	0.518 (0.426, 0.560) GeneJEPA+Morgan+xRFM, 0.560 ± 0.000
	Drug ZS	0.423 (0.203, 0.558) Geneformer+MACCS+xRFM, 0.558 ± 0.008	0.141 (-0.172, 0.379) GeneJEPA+MACCS+MLP, 0.379 ± 0.015
Cell-specific	Random	0.627 (0.114, 0.801) Geneformer+Morgan+LightGBM, 0.801 ± 0.002	0.603 (0.092, 0.796) Geneformer+Morgan+LightGBM, 0.796 ± 0.002
	Pair ZS (with leakage)	0.025 (-0.114, 0.166) Geneformer+Morgan+LightGBM, 0.166 ± 0.044	0.070 (-0.035, 0.160) scGPT+Morgan+LightGBM, 0.160 ± 0.035
	Pair ZS (leakage-free)	0.013 (-0.099, 0.195) Geneformer+BioMed+LightGBM, 0.195 ± 0.013	0.076 (-0.052, 0.232) Geneformer+BioMed+LightGBM, 0.232 ± 0.010
	Cell-line ZS	-0.058 (-0.150, 0.044) Geneformer+Morgan+LightGBM, 0.044 ± 0.008	0.071 (-0.045, 0.179) Geneformer+MACCS+LightGBM, 0.179 ± 0.007
	Drug ZS	0.113 (-0.080, 0.260) GeneJEPA+MoLFormer+LightGBM, 0.260 ± 0.000	0.091 (-0.059, 0.220) GeneJEPA+BioMed+LightGBM, 0.220 ± 0.005

Table 10: **Effect of allowing cell-level overlap in the pair zero-shot split.** Number of configurations (of 90) whose MCC is higher with leakage than without, and the mean difference (with leakage minus leakage-free).

Setting	Convention	Configurations improved	Mean ΔMCC
Shared-label	within-drug	88 / 90	$+0.051$
Shared-label	within-cell-line	65 / 90	$+0.009$
Cell-specific	within-drug	48 / 90	$+0.012$
Cell-specific	within-cell-line	43 / 90	-0.007

F.3 Single-modality ablations and prevalence baselines by model

Table 11 reports the mean MCC over configurations for the full model, the matching single-modality ablation, and the informative prevalence baseline, separately for each predictive model. For within-drug MCC the relevant ablation is the cell-only model and the relevant baseline is the cell-line prevalence; for within-cell-line MCC they are the drug-only model and the drug prevalence. In the shared-label setting, single-modality models track their prevalence baselines closely (for the MLP and xRFM the drug-only model reproduces the drug-prevalence baseline exactly), whereas the LightGBM drug-only model yields near-zero within-cell-line MCC despite an AUROC of 0.85, reflecting that its probabilities fall on one side of the fixed 0.5 threshold. In the cell-specific-label setting, values in the main text refer to the MLP.

F.4 AUROC under pair zero-shot generalization

Figures 11 and 12 report within-drug AUROC for the full representation–model grid under pair zero-shot generalization, complementing the MCC heatmaps in the main text (Fig. 2).

Table 11: **Mean MCC over configurations for the full model, the single-modality ablation, and the informative prevalence baseline.** "n/d" marks baselines that are undefined because the held-out group is absent from training (fallback to the global prior, MCC 0). Cell-specific ablations were not run for the leakage-free pair split.

		Within-drug MCC			Within-cell-line MCC		
Split	Model	Full	Cell-only	Line prev.	Full	Drug-only	Drug prev.
Shared-label setting							
Random	MLP	0.909	0.526	0.547	0.922	0.584	0.584
	LightGBM	0.843	0.531	0.547	0.883	0.000	0.584
	xRFM	0.897	0.522	0.547	0.926	0.580	0.584
Pair ZS (with leakage)	MLP	0.261	0.427	0.429	0.364	0.584	0.584
	LightGBM	0.276	0.430	0.429	0.399	0.014	0.584
	xRFM	0.161	0.441	0.429	0.476	0.583	0.584
Pair ZS (leakage-free)	MLP	0.217	0.455	0.458	0.359	0.584	0.584
	LightGBM	0.218	0.459	0.458	0.392	0.007	0.584
	xRFM	0.111	0.456	0.458	0.462	0.581	0.584
Cell-line ZS	MLP	−0.026	0.036	n/d	0.508	0.584	0.584
	LightGBM	−0.030	0.049	n/d	0.526	0.010	0.584
	xRFM	−0.027	0.058	n/d	0.521	0.581	0.584
Drug ZS	MLP	0.415	0.426	0.436	0.188	0.339	n/d
	LightGBM	0.441	0.430	0.436	0.045	0.000	n/d
	xRFM	0.414	0.438	0.436	0.191	0.015	n/d
Cell-specific-label setting							
Random	MLP	0.493	0.528	0.067	0.437	0.013	0.013
	LightGBM	0.700	0.680	0.067	0.689	0.013	0.013
	xRFM	0.689	0.697	0.067	0.682	0.007	0.013
Pair ZS (with leakage)	MLP	0.004	0.049	n/d	0.037	−0.001	n/d
	LightGBM	0.087	0.143	n/d	0.114	−0.000	n/d
	xRFM	−0.014	0.118	n/d	0.059	0.002	n/d
Cell-line ZS	MLP	−0.072	−0.078	n/d	0.041	0.004	n/d
	LightGBM	−0.047	−0.017	n/d	0.101	0.000	n/d
	xRFM	−0.057	0.003	n/d	0.072	0.006	n/d
Drug ZS	MLP	0.066	0.088	n/d	0.054	−0.010	n/d
	LightGBM	0.195	0.216	n/d	0.147	−0.010	n/d
	xRFM	0.080	0.194	n/d	0.071	−0.004	n/d

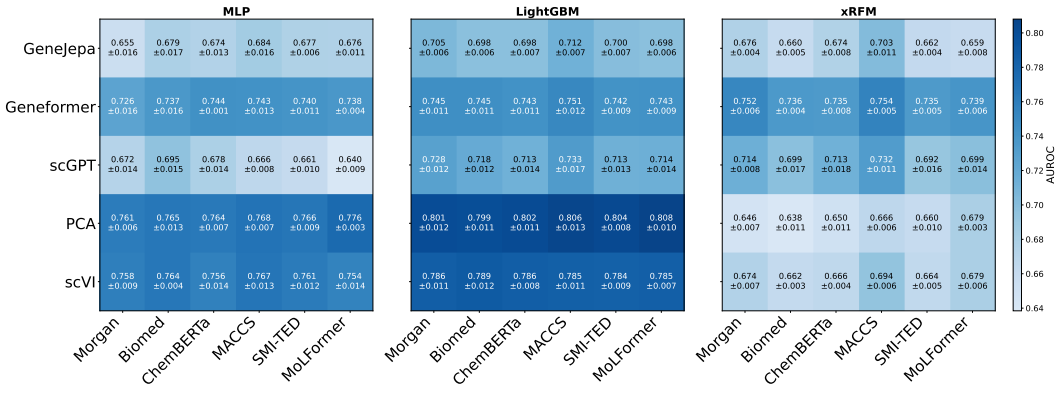

Figure 11: **Shared-label setting: within-drug AUROC across the full encoder × model grid under the leakage-free pair zero-shot split.** Each tile corresponds to one cell-representation, drug-representation, predictive-model combination. Rows are cell representations, columns are drug representations, and panels are predictive models. Scores are evaluated at the pair level for each drug and then macro-averaged over drugs. Standard errors reflect variability across the three seeds. The color scale is specific to this split.

F.5 Configuration-level performance across splits and evaluation axes

Detailed configuration-level MCC results for the shared-label setting are shown in Figs. 13 and 14, and the corresponding results for the cell-specific-label setting in Figs. 15 and 16. Color scales are specific to each split panel and are therefore not comparable across splits.

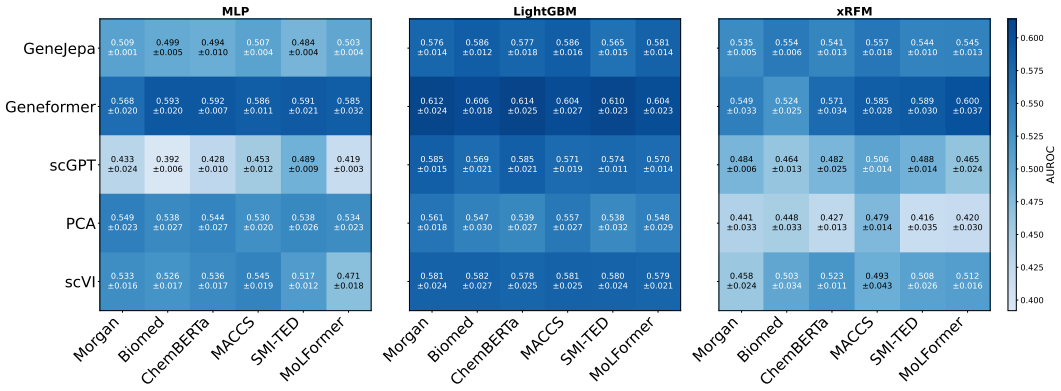

Figure 12: **Cell-specific-label setting: within-drug AUROC across the full encoder × model grid under the pair zero-shot split (with leakage).** Each tile corresponds to one cell-representation, drug-representation, predictive-model combination. Rows are cell representations, columns are drug representations, and panels are predictive models. Scores are evaluated at the cell level for each drug and then macro-averaged over drugs. Standard errors reflect variability across the three seeds. The color scale is specific to this split.

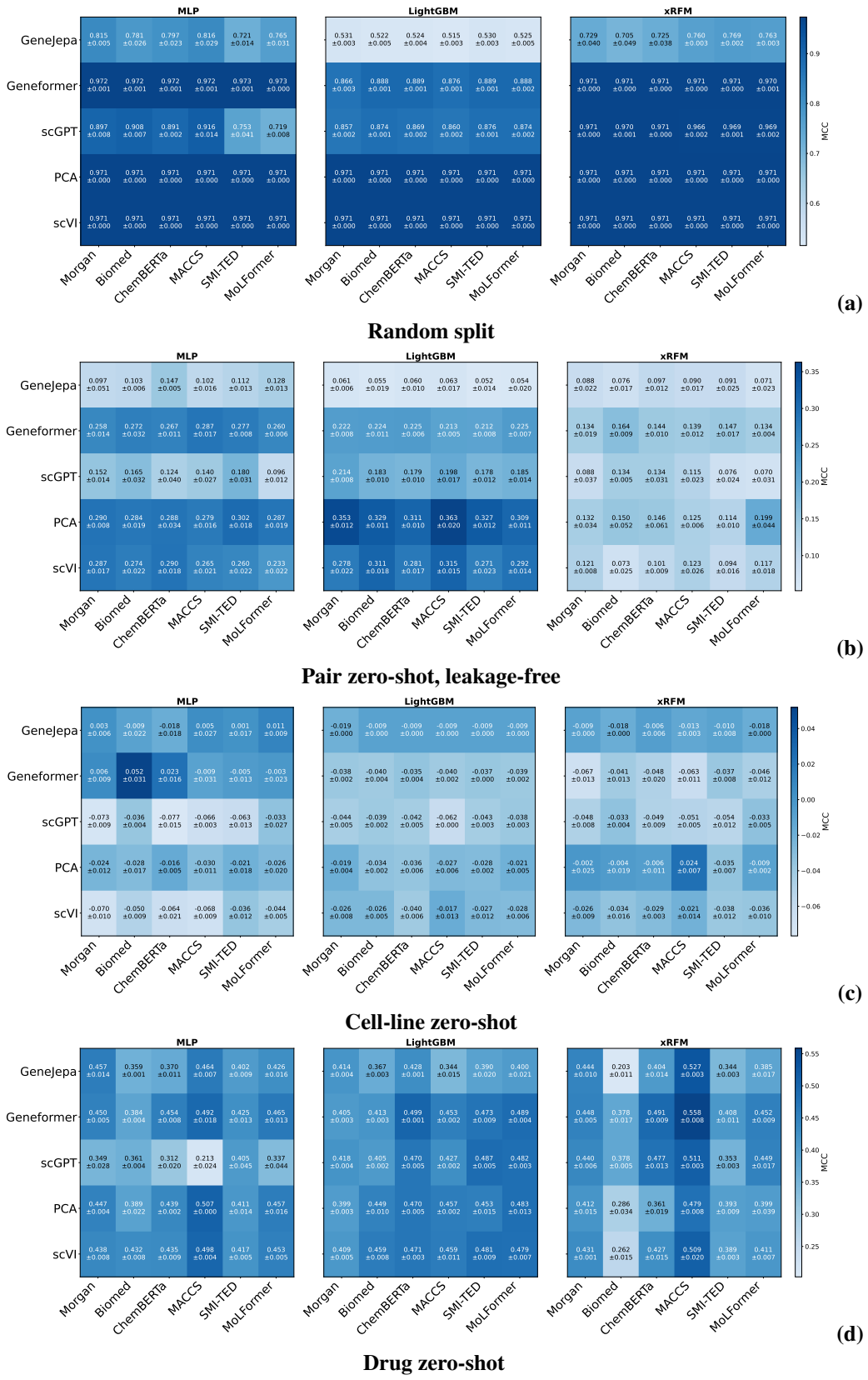

Figure 13: **Configuration-level within-drug MCC in the shared-label setting.** Heatmaps show mean ± SE over three seeds for each combination of cell representation (rows), drug representation (columns), and predictive model (panels).

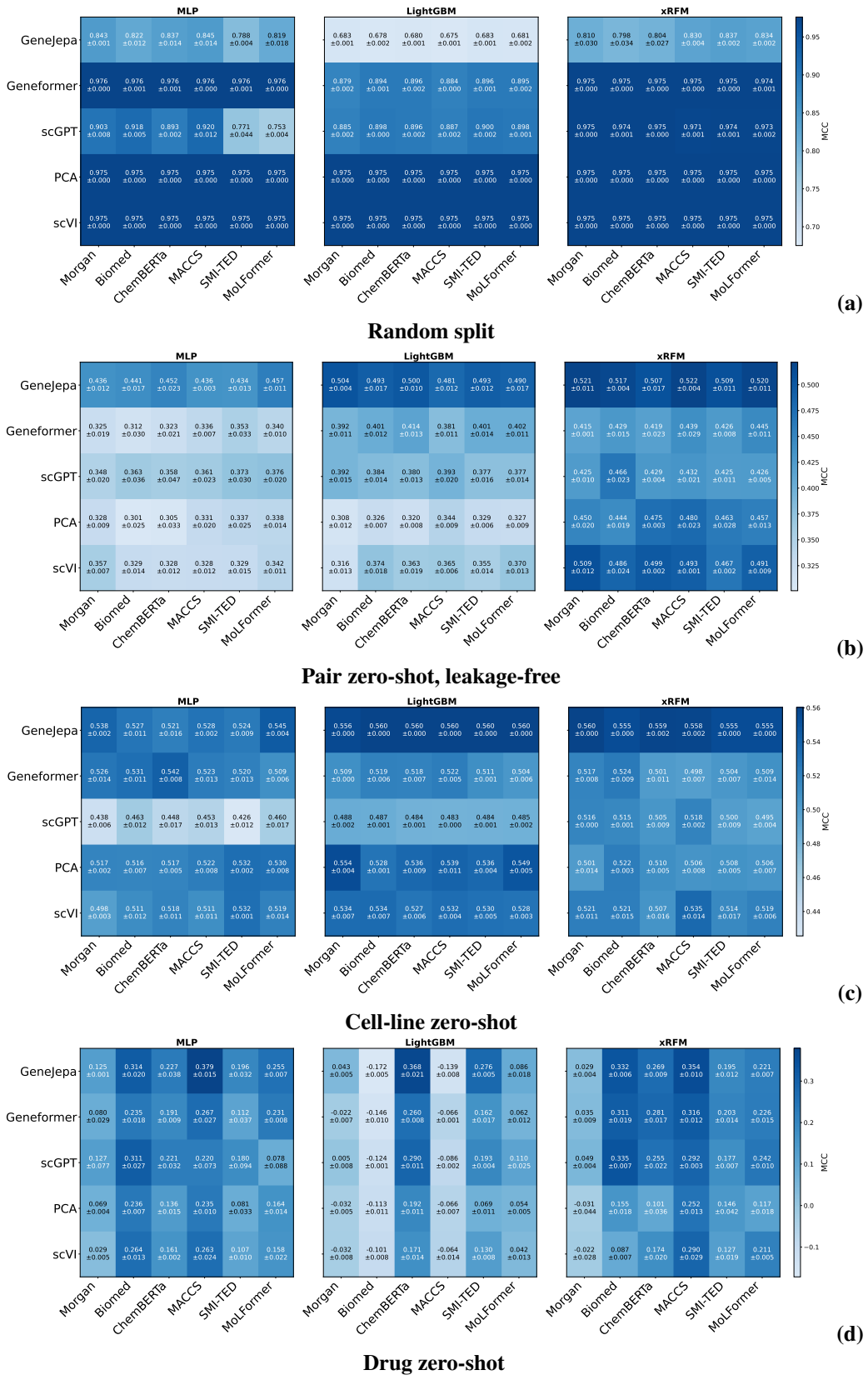

Figure 14: **Configuration-level within-cell-line MCC in the shared-label setting.** Heatmaps show mean ± SE over three seeds for each combination of cell representation (rows), drug representation (columns), and predictive model (panels).

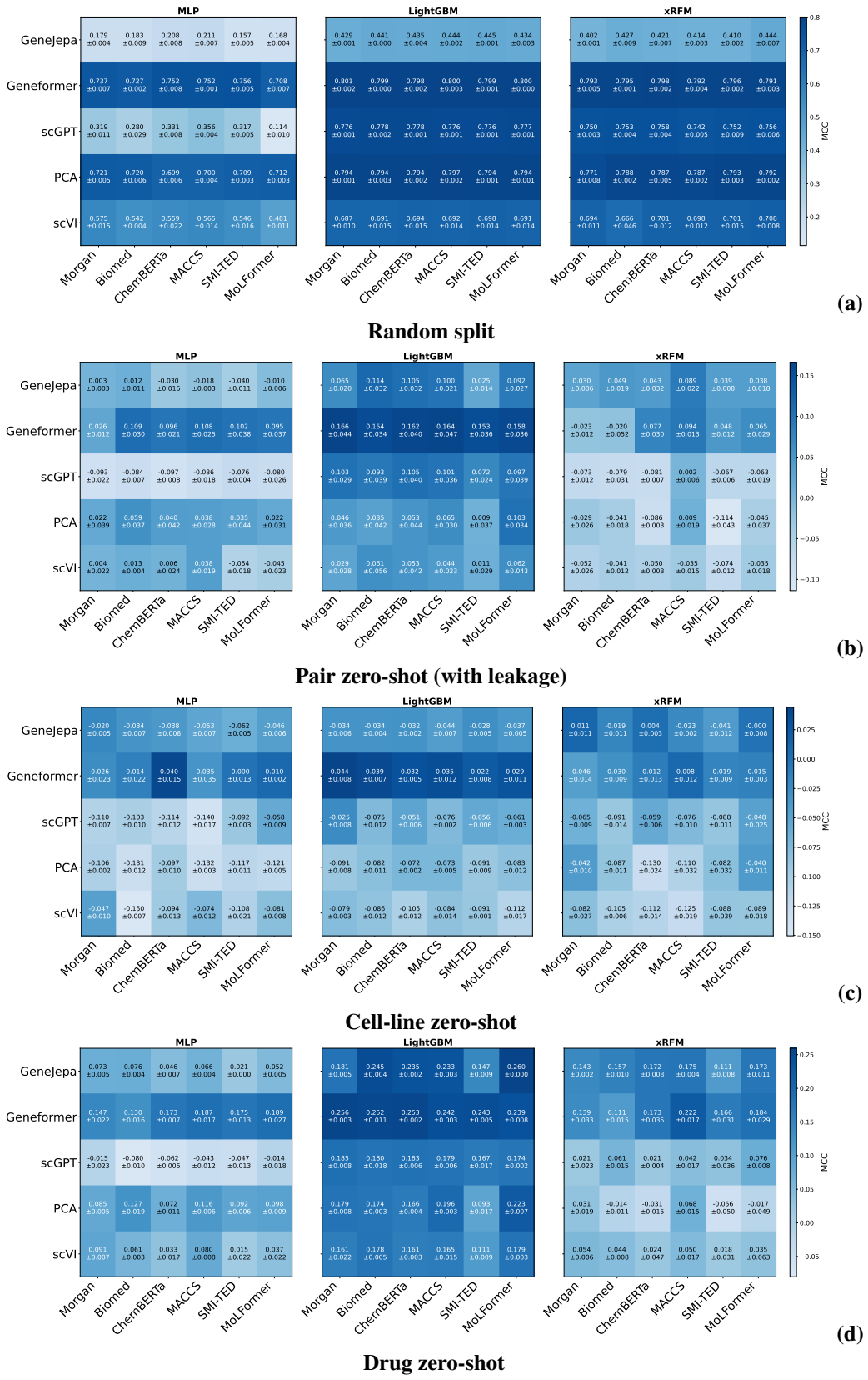

Figure 15: **Configuration-level within-drug MCC in the cell-specific-label setting.** Heatmaps show mean ± SE over three seeds for each combination of cell representation (rows), drug representation (columns), and predictive model (panels).

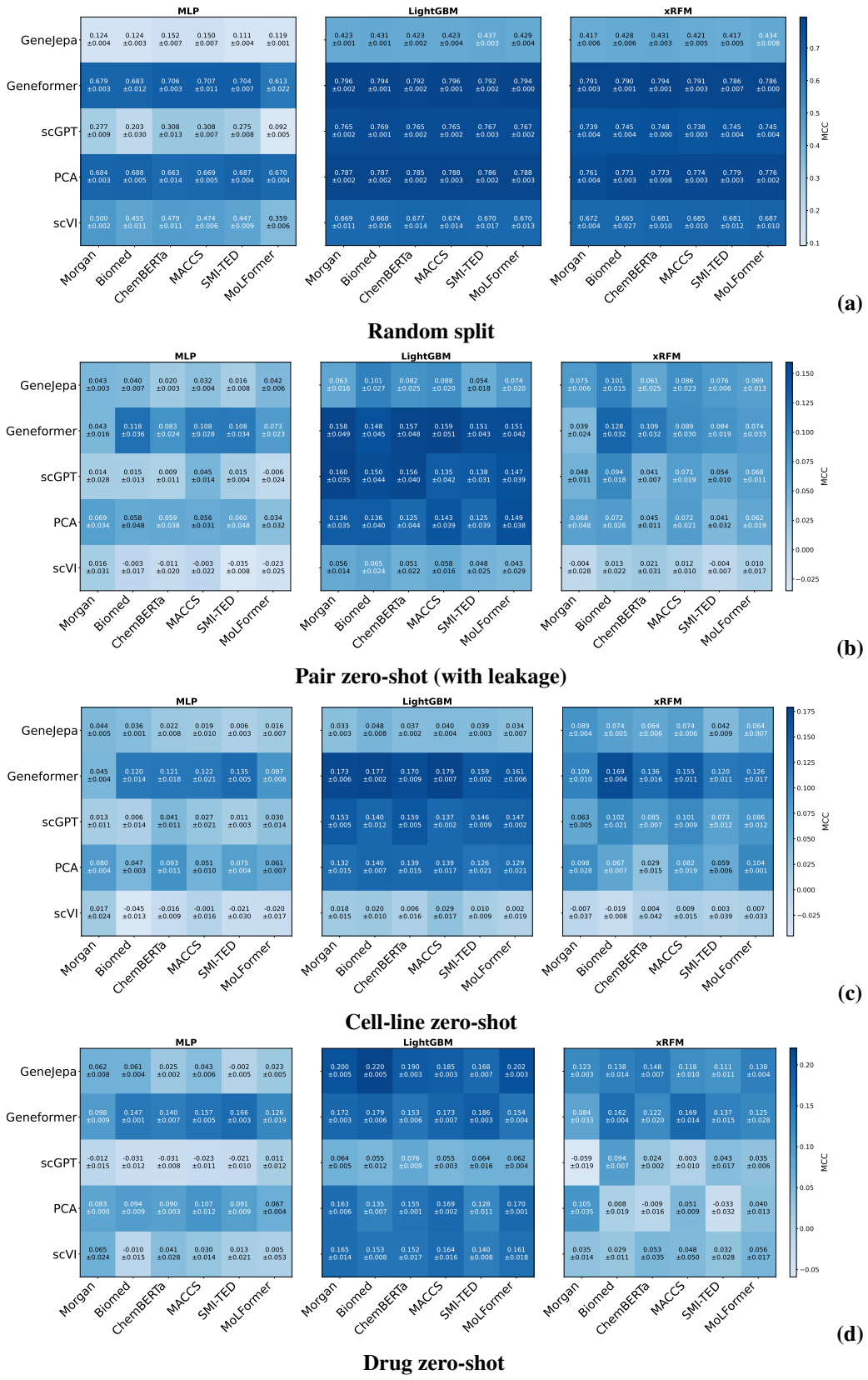

Figure 16: **Configuration-level within-cell-line MCC in the cell-specific-label setting.** Heatmaps show mean ± SE over three seeds for each combination of cell representation (rows), drug representation (columns), and predictive model (panels).