Widespread data leakage inflates accuracy and corrupts biomarker discovery in cancer drug response prediction

Amir Asiaee[1*], Jared Strauch[2], Leila Azinfar[1], Samhita Pal[1], Heather H. Pua[3], James P. Long[4], Kevin R. Coombes[5]

[1*]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, 37232, USA.
[2]Department of Biostatistics, Vanderbilt University, Nashville, TN, 37232, USA.
[3]Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN, 37232, USA.
[4]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA.
[5]Department of Biostatistics, Data Science, and Epidemiology, Georgia Cancer Center, Augusta University, Augusta, GA, 30912, USA.

*Corresponding author(s). E-mail(s): amir.asiaeetaheri@vumc.org;

Abstract

Drug response prediction models guide biomarker discovery, yet their reported accuracy depends on rigorous cross-validation (CV). We show that supervised feature screening applied to all samples before CV, a widespread practice, introduces data leakage that systematically underestimates prediction error. Across 265 drugs and 1,462 cancer cell lines, leakage-free CV raises mean squared error by 16.6% on average, with near-zero feature overlap between leaked and corrected pipelines (mean Jaccard 0.18; 36.2% of drugs share no features). Despite selecting five times more features, the leaked pipeline recovers known drug targets at nearly the same rate as the corrected pipeline, indicating that the inflated feature sets capture statistical artifacts rather than biological signal. A code-level audit of 32 published methods (2017–2024) confirms leakage in 23 (72%), spanning five distinct modes and cited over 3,000 times. The magnitude of inflation from this single leakage mode is comparable to improvement margins typically claimed over elastic net baselines, raising the possibility that some reported advances reflect

evaluation artifacts. We provide a leakage taxonomy, audit guide and reference implementation for leakage-free evaluation.

Keywords: data leakage, cross-validation, drug response prediction, pharmacogenomics, elastic net, reproducibility

Genomic markers that predict drug sensitivity in cancer cells are routinely nominated from large-scale pharmacogenomic screens such as the Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC) [1–4]. Cross-validation (CV) is the standard approach for estimating prediction performance in these high-dimensional settings. However, subtle errors in CV implementation can introduce data leakage—the inadvertent use of test-set information during model training—leading to optimistic performance estimates and unreliable feature rankings [5–8]. Here we quantify one prevalent form of leakage, supervised feature screening performed on the full dataset before CV splits, and document its widespread occurrence across the drug response prediction literature.

We compared two elastic-net pipelines [9, 10] using GDSC drug sensitivity for 265 compounds matched to CCLE molecular features for 1,462 cell lines spanning 31 tissue lineages: (i) an *incorrect* pipeline that applies variance filtering, correlation screening and scaling to the full dataset before CV—a "screen-then-validate" pattern used in the seminal CCLE and GDSC studies [1, 2] that implicitly exposes test-fold labels to the feature selection step, violating the assumption that CV folds are held out [5]; and (ii) a *leakage-free* pipeline that repeats these steps within each fold using training data only (Fig. 1b; Online Methods).

Leakage-free CV increased the estimated prediction error for most compounds. Across 265 drugs, the corrected pipeline raised mean MSE by 16.6% (median 14.0%), and 83.0% of drugs showed inflated performance under the leaked pipeline (Fig. 1c). The effect was heterogeneous: 35.8% of drugs showed $\geq$20% inflation, 10.9% showed $\geq$40%, and the maximum reached 70.3%. Leakage also destabilized biomarker discovery. The Jaccard overlap between stable feature sets (features selected in $\geq$80% of bootstrap resamples) was low: mean 0.181, median 0.044, with 36.2% of drugs having zero overlap (Fig. 1d). The leaked pipeline produced much larger signatures—on average 18.1 stable features versus 2.2 under the correct pipeline (Fig. 1e). Critically, the inflated feature sets did not translate into proportionately better biological signal: across 219 drugs with known gene targets, the incorrect pipeline recovered any target for only 16.4% of drugs despite selecting five times more features, compared with 15.5% under the correct pipeline (Fig. 1f). These results indicate that leakage simultaneously inflates accuracy and overstates confidence in derived biomarker lists.

To assess how broadly leakage extends beyond elastic net, we audited 32 drug response prediction methods spanning classical machine learning and deep learning, published between 2017 and 2024 (Table 1). We categorized leakage into five recurring modes (Supplementary Note 1): (1) preprocessing on all samples before CV, (2) test data used for early stopping or model selection, (3) pair-level splits inconsistent with cold-start claims, (4) target-domain adaptation using test samples, and (5) post-hoc

selection of the best test metric. Across the audited methods, 23 of 32 (72%) contained confirmed leakage, three could not be verified due to incomplete public code, and six appeared clean (Table 1; Fig. 1a). The 23 confirmed-leaky methods have accumulated over 3,000 citations (OpenAlex, February 2026). Pre-CV preprocessing was the most prevalent mode, but several methods exhibited multiple leakage types simultaneously.

Of the 20 leaky methods for which we could extract a numeric performance margin (Supplementary Table 2), only two directly benchmark against elastic net: DrugCell (5.7% improvement) and TRANSACT (10.4%), both well within our mean leakage inflation of 16.6%. Across the broader set—which uses diverse baselines, metrics and evaluation protocols—the mean claimed improvement is 13.1% (median 9.2%), and 75% claim gains no larger than 16.6%. Because our inflation estimate derives from elastic net, these cross-baseline comparisons are approximate, but the order of magnitude suggests that evaluation artifacts may account for a substantial fraction of reported advances in drug response prediction. Analogous findings reinforce this concern: correcting data leakage in civil war prediction [11], neuroimaging [12] and connectomics [13] eliminated reported advantages of complex models. More broadly, rigorous re-evaluation of train/test methodology has shown that deep learning gene perturbation models do not outperform linear baselines [14] and that drug features contribute zero performance in deep response models [15]. Although we quantify inflation for elastic net, the leakage pattern—applying supervised screening to all samples before any CV split—biases any downstream model [5, 12], and our audit confirms its presence across both classical and deep learning architectures (Table 1).

Our findings do not imply that prior biological conclusions are necessarily invalid, but that many published performance numbers and feature rankings are likely overconfident. The consequences for biomarker discovery may be particularly severe: leakage inflates stable feature counts by nearly an order of magnitude (Fig. 1e) while providing no proportionate gain in drug target recovery (Fig. 1f), indicating that feature rankings from leaked pipelines are dominated by statistical artifacts rather than biological signal. This undermines downstream applications—drug repurposing, mechanism-of-action studies, patient stratification—where inflated feature lists can misdirect experimental follow-up and biological interpretation. To support improved practice, we provide three concrete deliverables: (1) a five-mode leakage taxonomy with code-level evidence for each audited method, (2) an audit checklist for identifying leakage in new methods, and (3) a reference implementation of leakage-free CV for drug response prediction, available at https://github.com/AsiaeeLab/drug-response-leakage. An independent benchmark effort, DrEval [16], has recently proposed standardized evaluation protocols that complement these resources.

Table 1: Code-level audit of 32 drug response prediction methods (2017–2024). Status reflects direct inspection of released source code: × confirmed leakage; — not independently verifiable (incomplete public code); ✓ no leakage found. The accompanying code repository provides specific commit hashes, file paths and line numbers for every finding.

Method	Year	Status	Confirmed leakage types
SRMF [17]	2017	×	No held-out split; in-sample evaluation
CaDRReS [18]	2018	×	Repo-default train=test; in-sample evaluation
CDRscan [19]	2018	×	Pair-level split
DeepDR [20]	2019	×	Pair-level split
DrVAE [21]	2019	✓	—
GraphDRP [22]	2019	×	Cell-level leakage
MOLI [23]	2019	×	Feature selection before CV
PaccMann [24]	2019	×	Pair-level split; global scaling
DeepCDR [25]	2020	×	Test-as-validation; pair-level split
DrugCell [26]	2020	×	Test monitoring / model selection
PANCDR [27]	2020	×	Test in training loop; domain adaptation
DualGCN [28]	2021	×	Test-as-validation; pair-level split
HiDRA [29]	2021	—	Incomplete public code
SWnet [30]	2021	×	Test-driven checkpointing; pair-level split
TRANSACT [31]	2021	×	Target-domain adaptation leakage
TransCDR [32]	2021	×	Global scaling; pair-level split
VAEN [33]	2021	—	Incomplete public code
AutoBorutaRF [34]	2022	×	Supervised feature selection before CV
DRPreter [35]	2022	×	Global preprocessing; test monitored each epoch
GADRP [36]	2022	—	Official repo empty
GraTransDRP [37]	2022	×	Pre-split PCA; test evaluated each epoch
NIHGCN [38]	2022	×	Test in training loop; pair-level split
TGSA [39]	2022	×	Default pair-level split; global preprocessing
Velodrome [40]	2022	✓	—
DeepResponse [41]	2023	✓	—
MTDRP [42]	2023	×	Train+test joint normalization
SubCDR [43]	2023	×	Test in training loop; pair-level split
DIPK [44]	2024	×	Test monitoring; post-hoc test-loss comparison
MMDRP [45]	2024	✓	—
PASO [46]	2024	×	Test-based model selection; global scaling
ScreenDL [47]	2024	✓	—
UNO [48]	2024	✓	—

1 Online Methods

1.1 Data sources

We analyzed drug sensitivity measurements from GDSC/Sanger release 6.0 and molecular features derived from CCLE/DepMap releases distributed with the CCLE update by Ghandi et al. [2–4]. The predictor matrix contained 86,546 features spanning gene expression, copy number, mutation and protein measurements for 1,462 cell lines, and drug response covered 265 compounds (AUC/activity-area values). Data provenance is documented in the accompanying code repository.

1.2 Correct and incorrect cross-validation pipelines

For each drug, we fit elastic-net regression models using `glmnet` [10] to predict response from molecular features. The **incorrect** pipeline applied the full preprocessing stack once using all samples: (i) variance filtering, (ii) correlation screening with the response, (iii) duplicate-feature removal, and (iv) scaling (excluding binary mutation features), and then performed 5-fold CV on the filtered matrix. The **leakage-free** pipeline repeated the same steps independently within each fold, fitting preprocessing on the training split and applying it to the held-out split.

Both approaches used the same hyperparameter grid over the elastic-net mixing parameter (α) and selected λ by minimum CV error along the regularization path. We report mean squared error (MSE) on held-out folds and compute standard errors as the fold-to-fold standard deviation divided by $\sqrt{K}$, following the standard definition of SE for the mean of K fold-level estimates.

1.3 Bootstrap feature selection

To quantify feature selection characteristics, we performed bootstrap resampling (200 replicates) and counted how often each feature received a non-zero coefficient. We report counts of selected features, distributions of selection frequencies, and overlap metrics (Jaccard index) at multiple bootstrap frequency thresholds.

1.4 Error bar calculation

We report the standard error of the mean ($\mathrm{SE} = \mathrm{SD}/\sqrt{K}$) of fold-level errors as the uncertainty measure for CV performance, following the standard definition for the mean of K fold-level estimates.

1.5 Code-level audit methodology

We audited 32 drug response prediction methods published between 2017 and 2024. Methods were identified through literature searches on PubMed and Google Scholar using terms related to drug response prediction, cell line sensitivity and pharmacogenomics, filtered to those with publicly available source code repositories on GitHub or similar platforms. The selection covers the major model families in the field (matrix factorization, random forests, deep neural networks, graph neural networks and variational autoencoders) but is not exhaustive. For each method, we identified the training

and evaluation scripts, traced data flow through preprocessing, splitting and model fitting stages, and classified any leakage patterns into the five-mode taxonomy (Supplementary Note 1). Specific commit hashes, file paths and line numbers supporting each assessment are documented in the accompanying audit guide.

1.6 Use of large language models

Large language models were used to assist with drafting and editing the manuscript text and with code-level auditing. All scientific claims, quantitative results and classifications were verified against the repository source code and result files by the authors, who take full responsibility for the content.

1.7 Reporting summary

No human participants or animal experiments were involved.

Data availability

Processed result tables and figures are available in the accompanying repository at https://github.com/AsiaeeLab/drug-response-leakage. The underlying CCLE/DepMap molecular data and GDSC drug response data are available from their original sources.

Code availability

All analysis code, including leakage-free CV implementations, figure reproduction scripts and the code-level audit guide, is available at https://github.com/AsiaeeLab/drug-response-leakage.

Author contributions

A.A. conceived the study, discovered the leakage pattern, designed the evaluation framework, implemented the analysis pipeline, conducted the code-level audit and wrote the manuscript. J.S. contributed to discovering the leakage in the original CCLE study and to data processing and pipeline implementation. L.A. crafted parts of the manuscript. S.P. edited the manuscript. H.H.P. provided biological interpretation and edited the manuscript. J.P.L. contributed statistical methodology guidance and edited the manuscript. K.R.C. supervised early stages of the project and edited the manuscript.

Competing interests

The authors declare no competing interests.

References

[1] Barretina, J. *et al.* The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012). URL http://dx.doi.org/10.1038/nature11003.

[2] Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012). URL http://dx.doi.org/10.1038/nature11005.

[3] Iorio, F. *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016). URL http://dx.doi.org/10.1016/j.cell.2016.06.017.

[4] Ghandi, M. *et al.* Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503–508 (2019). URL http://dx.doi.org/10.1038/s41586-019-1186-3.

[5] Ambroise, C. & McLachlan, G. J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences* **99**, 6562–6566 (2002). URL http://dx.doi.org/10.1073/pnas.102102699.

[6] Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7** (2006). URL http://dx.doi.org/10.1186/1471-2105-7-91.

[7] Cawley, G. C. & Talbot, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* **11**, 2079–2107 (2010). URL http://jmlr.org/papers/v11/cawley10a.html.

[8] Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews Genetics* **23**, 169–181 (2022).

[9] Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**, 301–320 (2005). URL http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.

[10] Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** (2010). URL http://dx.doi.org/10.18637/jss.v033.i01.

[11] Kapoor, S. & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **4**, 100804 (2023).

[12] Pulini, A. A., Kerr, W. T., Loo, S. K. & Lenartowicz, A. Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: effects of sample size and circular analysis. *Biological Psychiatry: Cognitive Neuroscience*

and Neuroimaging **4**, 108–120 (2019).

[13] Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S. & Scheinost, D. Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications* **15**, 1829 (2024).

[14] Kedzierska, K. Z., Crawford, L. & Amariuta, T. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods* (2025).

[15] Branson, O. E., Cutillas, P. R. & Bessant, C. Understanding the sources of performance in deep drug response models reveals insights and improvements. *Bioinformatics* **41**, i197–i205 (2025).

[16] Bernett, J., Gao, Y., Keilwagen, J., Baumbach, J. & Zolotareva, O. From hype to health check: Critical evaluation of drug response prediction models with DrEval. *bioRxiv* (2025). Preprint.

[17] Wang, L., Li, X., Zhang, L. & Gao, Q. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC Cancer* **17**, 513 (2017).

[18] Suphavilai, C., Bertrand, D. & Nagarajan, N. CaDRReS: A calculator for drug response score. *Bioinformatics* **34**, 3907–3909 (2018).

[19] Chang, Y. *et al.* CDRscan: a drug response model using convolutional neural networks. *BMC Bioinformatics* **19**, 526 (2018).

[20] Zeng, X. *et al.* DeepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* **35**, 5191–5198 (2019).

[21] Rampasek, L., Hidru, D., Smirnov, P., Haibe-Kains, B. & Goldenberg, A. Dr.VAE: improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics* **35**, 3743–3751 (2019).

[22] Nguyen, T., Nguyen, G. T. T., Nguyen, T. & Le, D.-H. GraphDRP: a graph neural network for drug response prediction. *bioRxiv* (2021). Preprint.

[23] Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C. & Ester, M. Moli: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* **35**, i501–i509 (2019). URL http://dx.doi.org/10.1093/bioinformatics/btz318.

[24] Manica, M. *et al.* Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Molecular Pharmaceutics* **16**, 4797–4806 (2019).

[25] Liu, Q., Hu, Z., Jiang, R. & Zhou, M. Deepcdr: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics* **36**, i911–i918 (2020). URL http://dx.doi.org/10.1093/bioinformatics/btaa822.

[26] Kuenzi, B. M. *et al.* Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* **38**, 672–684.e6 (2020). URL http://dx.doi.org/10.1016/j.ccell.2020.09.014.

[27] Kim, J., Park, S., Min, D. & Kim, W. PANCDR: precise prediction for anticancer drug response. *BMC Bioinformatics* **23**, 424 (2022).

[28] Ma, T. *et al.* DualGCN: a dual graph convolutional network model to predict cancer drug response. *BMC Bioinformatics* **23**, 129 (2022).

[29] Jin, I. & Nam, H. HiDRA: Hierarchical network for drug response prediction with attention. *Journal of Chemical Information and Modeling* **61**, 3858–3867 (2021).

[30] Zuo, Z. *et al.* SWnet: a deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures. *BMC Bioinformatics* **22**, 434 (2021).

[31] Mourragui, S., Loog, M., van de Wiel, M. A., Reinders, M. J. T. & Wessels, L. F. A. TRANSACT maps transcriptional states across species by transferring drug response predictions. *Nature Communications* **12**, 7025 (2021).

[32] Xia, X. *et al.* TransCDR: a deep learning model for enhancing the generalizability of drug activity prediction. *bioRxiv* (2022). Preprint.

[33] Simidjievski, N. *et al.* Variational autoencoders for cancer data integration: design principles and computational practice. *Frontiers in Genetics* **10**, 1205 (2019).

[34] Xu, B. *et al.* AutoBorutaRF: an automated feature selection method for drug response prediction. *Briefings in Bioinformatics* **23** (2022).

[35] Shin, J., Piao, Y., Bang, D., Kim, S. & Jo, K. DRPreter: Interpretable anticancer drug response prediction using knowledge-primed graph neural networks and multi-head attention. *Bioinformatics* **38**, 4626–4628 (2022).

[36] Peng, S.-L. *et al.* GADRP: graph-augmented drug response prediction. *Bioinformatics* (2022).

[37] Nguyen, D.-T. *et al.* GraTransDRP: multimodal drug response prediction using graph transformer. *bioRxiv* (2022). Preprint.

[38] Wei, B. *et al.* NIHGCN: drug response prediction by integrating heterogeneous information. *Briefings in Bioinformatics* **23** (2022).

[39] Zhu, Y. *et al.* TGSA: protein-protein association-based twin graph neural networks for drug response prediction with similarity augmentation. *Bioinformatics* **38**, 461–468 (2022).

[40] Sharifi-Noghabi, H., Meng, C.-H., Zhu, X., Collins, C. C. & Ester, M. Velodrome: Out-of-distribution generalization from labeled and unlabeled gene expression data for drug response prediction. *bioRxiv* (2022). Preprint.

[41] Izmirli, B. C. *et al.* DeepResponse: large-scale prediction of cancer cell line drug response. *Briefings in Bioinformatics* (2023).

[42] Shen, X. *et al.* Multi-task deep learning for drug response prediction. *bioRxiv* (2023).

[43] Liu, X. *et al.* SubCDR: a subcomponent-guided deep learning model for cancer drug response prediction. *bioRxiv* (2023).

[44] Ma, P. *et al.* Improving drug response prediction via integrating gene relationships with deep learning. *Briefings in Bioinformatics* **25** (2024).

[45] Steinberg, L. *et al.* Multi-modal drug response prediction. *bioRxiv* (2024).

[46] Yang, Q. *et al.* PASO: a transformer-based multi-scale drug response prediction model. *bioRxiv* (2024).

[47] Ederman, C. S. *et al.* ScreenDL: a transfer learning framework for drug response prediction. *bioRxiv* (2024).

[48] JDACS4C-IMPROVE Consortium. UNO: unified drug response predictor. *IMPROVE Benchmark* (2024).

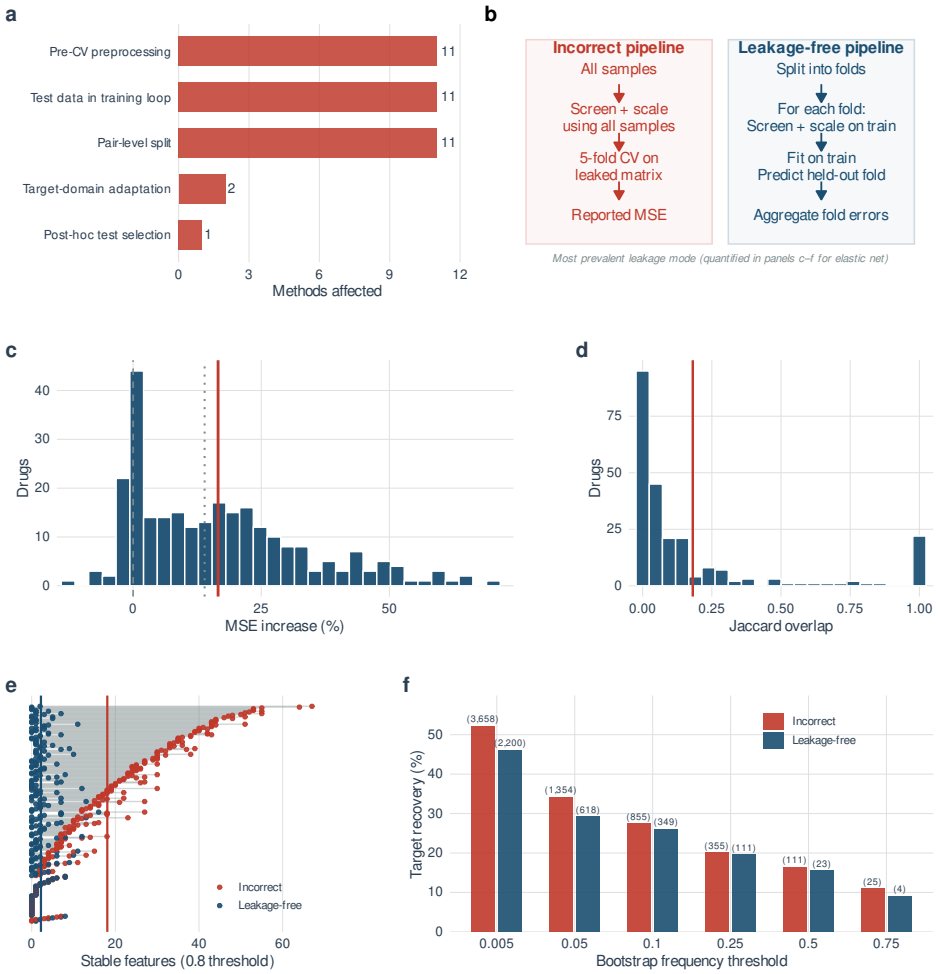

Fig. 1 Data leakage inflates performance estimates and destabilizes feature sets across 265 drugs. a, Leakage mode distribution across 23 confirmed-leaky methods (Table 1); methods may exhibit multiple modes. **b,** Schematic contrasting the incorrect pipeline (supervised screening on all samples before CV) with the leakage-free pipeline (screening repeated within each fold); this is the most prevalent leakage mode, quantified in panels c–f. **c,** Distribution of MSE inflation (%) under leakage-free CV; solid line, mean (16.6%); dotted line, median (14.0%). **d,** Jaccard overlap between stable feature sets (≥80% bootstrap threshold); solid line, mean (0.18); 36.2% of drugs have zero overlap. **e,** Stable feature counts per drug under incorrect (red) and leakage-free (blue) pipelines; drugs are ranked by the gap between incorrect and correct feature counts; vertical lines mark means (18.1 vs 2.2). **f,** Drug target recovery rates across bootstrap frequency thresholds for incorrect versus leakage-free pipelines (219 drugs with known targets); numbers in parentheses show mean selected feature counts per pipeline; the incorrect pipeline selects 5× more features yet recovers only marginally more targets (e.g., 16.4% vs 15.5% at the 0.50 threshold), indicating that leakage inflates feature counts without proportionate gains in biological signal.

Supplementary Information: Widespread data leakage inflates accuracy and corrupts biomarker discovery in cancer drug response prediction

Amir Asiaee[1*], Jared Strauch[2], Leila Azinfar[1], Samhita Pal[1], Heather H. Pua[3], James P. Long[4], Kevin R. Coombes[5]

[1*]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, 37232, USA.
[2]Department of Biostatistics, Vanderbilt University, Nashville, TN, 37232, USA.
[3]Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN, 37232, USA.
[4]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA.
[5]Department of Biostatistics, Data Science, and Epidemiology, Georgia Cancer Center, Augusta University, Augusta, GA, 30912, USA.

*Corresponding author(s). E-mail(s): amir.asiaeetaheri@vumc.org;

1 Supplementary Note 1 — Leakage taxonomy

We identified five recurring modes of data leakage in drug response prediction methods. This taxonomy was developed through code-level inspection of 32 published methods and is intended as a practical audit guide.

1. **Pre-CV preprocessing leakage**: normalization, PCA, variance filtering or supervised feature selection fit on all samples before cross-validation. This is the most prevalent mode (Fig. 1a in the main text) and the one we quantify experimentally. When screening statistics (e.g., correlation with the response) are computed on the full dataset, test-fold labels influence which features enter the model, inflating apparent accuracy and producing non-replicable feature sets.
2. **Test set in training loop**: the test or validation split is used for early stopping, model selection or hyperparameter tuning during training. Even when a nominal

train/test split exists, monitoring test-set performance across epochs and selecting the best checkpoint introduces leakage that can bias final reported metrics.

3. **Pair-level split mismatch**: random splits on (cell line, drug) response pairs when the stated goal is generalization to new cell lines or new drugs. Because the same cell line (or drug) can appear in both training and test sets under pair-level splitting, the model may learn cell-line-specific or drug-specific patterns rather than generalizable drug–genomic relationships.
4. **Target-domain adaptation leakage**: domain adaptation or transfer learning methods that use unlabeled test-domain samples during training. When the test domain is the target domain and its samples (even without labels) are used to learn domain-invariant representations, information about the test distribution leaks into the model.
5. **Post-hoc test selection**: reporting the best test metric across epochs, hyperparameter configurations or random seeds without a separate held-out evaluation. This optimistic selection can inflate reported performance even when the training procedure itself is leak-free.

The full audit table is presented as Table 1 in the main text. The accompanying code repository includes direct links to specific commits, file paths and line numbers supporting each assessment.

2 Supplementary Note 2 — Additional methodological details

2.1 Variance filtering, supervised screening and scaling

For elastic net baselines, preprocessing consisted of:

- **Variance filter**: retain features with variance > 0.01.
- **Correlation screen**: retain features with $|\mathrm{cor}(\mathrm{feature}, \mathrm{response})| > 0.1$.
- **Duplicate removal**: remove duplicated columns after filtering.
- **Scaling**: z-score continuous features using training-set mean and standard deviation; binary mutation features (identified by the suffix pattern `_mut.`) were excluded from scaling.

In the incorrect pipeline, these steps were computed once using all samples (leakage). In the leakage-free pipeline, they were recomputed within each fold using training data only. Reference implementations are provided in the accompanying code repository.

2.2 Hyperparameter selection

Both pipelines used the same grid over the elastic-net mixing parameter $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ and selected λ by minimum 5-fold CV error along the regularization path. The hyperparameter combination yielding the lowest mean CV MSE was selected for each drug.

3 Supplementary Table 1 — Threshold sensitivity analysis

Jaccard and stable feature counts across bootstrap frequency thresholds (incorrect vs leakage-free CV pipeline):

Threshold	Mean Jacc.	Median Jacc.	Zero overlap (%)	Features (incorrect)	Features (leak-free)
0.50	0.147	0.103	16.5	68.6	9.8
0.60	0.128	0.071	21.3	49.4	5.9
0.70	0.107	0.053	28.0	33.9	3.2
0.80	0.094	0.034	42.9	21.5	1.7
0.90	0.091	0.000	65.8	10.9	0.8

Values derived from 265 drugs; stable features defined as those selected in $\geq$threshold fraction of 200 bootstrap resamples. Across all thresholds, the incorrect pipeline selects substantially more features with low overlap to the leakage-free pipeline.

4 Supplementary Table 2 — Margin-of-leakage comparison

Claimed performance margins of leaky methods. For each of the 20 leaky methods with an extractable numeric improvement from the original publication, we report the primary metric, strongest baseline comparator, and claimed relative improvement (%). Rows marked with † compare directly against elastic net—only these two rows permit a like-for-like comparison with our measured leakage inflation (mean 16.6%). All other baselines differ in model class, metric or evaluation protocol, so their improvement percentages serve as order-of-magnitude context rather than direct comparisons to our elastic-net-specific inflation estimate.

Method	Year	Metric	Baseline	Improvement (%)
SRMF	2017	RMSE	KBMF	13.5
CDRscan	2018	R^2	Random forest	20.8
MOLI	2019	AUC	Ridge	8.0
PaccMann	2019	RMSE	DNN	10.6
AutoBorutaRF	2019	AUC	Naive Bayes	4.8
DeepCDR	2020	RMSE	Random forest	53.4
DrugCell†	2020	Drug ρ	Elastic net	5.7
GraphDRP	2021	RMSE	tCNNs	14.4
SWnet	2021	MSE	WGRMF	4.7
TRANSACT†	2021	ROC-AUC	Elastic net	10.4
DualGCN	2022	RMSE	Ridge	14.9
DRPreter	2022	MSE	TGSA	7.9
GraTransDRP	2022	RMSE	GraOmicDRP	2.1
NIHGCN	2022	AUC	MOFGCN	0.9
TGSA	2022	RMSE	DeepCDR	4.1

Method	Year	Metric	Baseline	Improvement (%)
SubCDR	2023	RMSE	Bi-GNN	1.8
DIPK	2024	MSE	Precily	25.6
PANCDR	2024	AUC	DeepCDR	34.8
TransCDR	2024	RMSE	Ablation	1.4
PASO	2025	RMSE	Random forest	22.8

5 Supplementary Figure 1

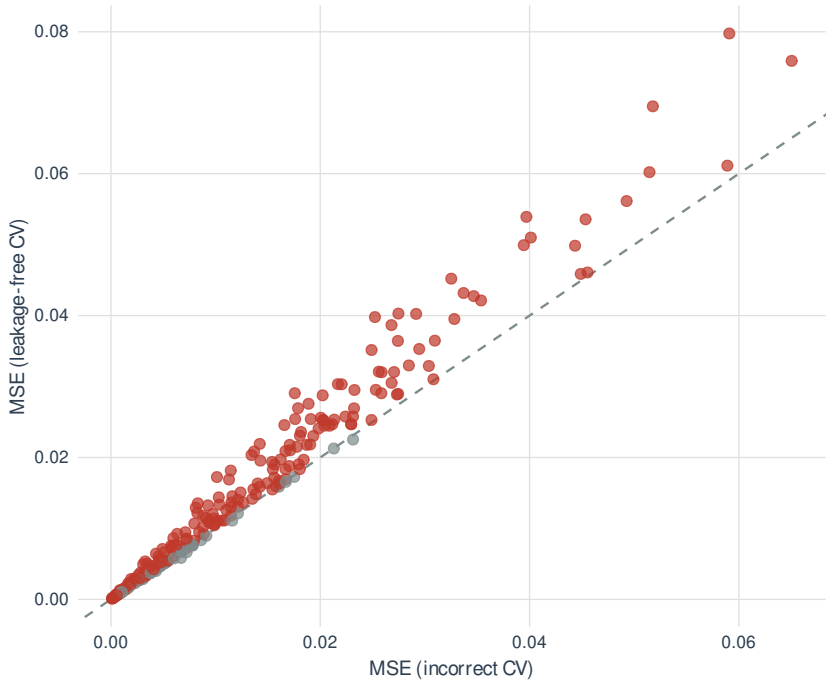

Fig. 1 Supplementary Fig. 1 | Drug-wise MSE under incorrect versus leakage-free CV. Each point represents one of 265 drugs. Points above the diagonal indicate higher MSE under leakage-free CV (i.e., the incorrect pipeline underestimated error). Of 265 drugs, 83.0% fall above the diagonal. Points are colored by direction: red indicates drugs where leakage-free CV reveals higher MSE; gray indicates drugs with no increase.

6 Code and Data Availability

All analysis code is available at https://github.com/AsiaeeLab/drug-response-leakage, including:

- Reference implementations of correct and incorrect CV pipelines

- Complete analysis notebooks reproducing all figures and tables
- Leakage audit guide with code links for each audited method
- Processed result tables and supplementary data files