DrugDis: Disentangling general and context-specific effects in drug-response prediction

Bo Li[1,2,†], Chengliang Liu[1,†], Yuzhong Peng[3,†], Bob Zhang[1,4,*], Qing Wang[5], Pinxian Zeng[1], Mengran Li[6], Shenghui Huang[7], Chuxia Deng[3,*], Yang Zhang[2,8,9,*]

[1]PAMI Research Group, Department of Artificial Intelligence, University of Macau, Macau SAR.
[2]Department of Computer Science, School of Computing, National University of Singapore, Singapore, Singapore.
[3]Faculty of Health Sciences, University of Macau, Macau SAR.
[4]Centre for Artificial Intelligence and Robotics, Institute of Collaborative Innovation, University of Macau, Macau SAR.
[5]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Florida, USA.
[6]Shenzhen Loop Area Institute, Shenzhen, Guangdong, China.
[7]Department of Molecular Biotechnology and Health Sciences, University of Turin, Turin, Italy.
[8]Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore.
[9]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore.

[†]These authors contributed equally to this work.
[*]Corresponding authors: Bob Zhang, PhD (bobzhang@um.edu.mo); Chuxia Deng, PhD (cxdeng@um.edu.mo); Yang Zhang, PhD (zhang@nus.edu.sg).

Abstract

Predicting cellular responses across perturbations and biological contexts is a central goal of AI virtual cells, yet aggregate prediction scores do not distinguish general response tendencies from context-specific effects. Here we introduce DrugDis, a component-resolved framework that applies an exact orthogonal decomposition to measured and predicted responses on the same observed pairs, separating additive effects from drug–sample interactions. It evaluates interaction direction, amplitude and prediction error separately, using cross-assay reproducibility as an empirical reference where repeated measurements are available. Across 3.14 million drug–sample pairs, additive effects accounted for 75.8% of response variance, and 59–83% within each of the five largest response resources. Under held-out cell lines, a model with an aggregate correlation of 0.86 recovered interactions at a correlation of only 0.32. Representation comparisons depended on the evaluated component, decoder and distribution shift, while holding out cell lines or compounds concentrated error in different components. Matched component supervision improved interaction direction for held-out compounds and in zero-shot organoid transfer, but consistently reduced interaction error only for held-out compounds. Different evaluation criteria changed model selection without improving selectivity error on independent GDSC2 measurements. DrugDis reveals which response structures underpin predictive performance, providing a clearer basis for comparing models, interpreting generalization and assessing model improvement.

Predicting how cells respond to new perturbations and in new biological contexts is a central ambition of AI virtual cells[1,2]. Drug-response prediction offers a well-established setting for examining this challenge, supported by large pharmacogenomic screens spanning diverse compounds and cellular backgrounds[3–5]. In continuous-response benchmarks, models are commonly compared by the overall agreement between predictions and measurements on held-out data[6,7]. These scores inform choices of architectures and molecular or transcriptomic representations, as well as claims about generalization[8,9]. However, they do not reveal which features of drug response a model has recovered.

Drug-response data form a partially observed matrix of compounds and biological samples. On the measured pairs, responses can be separated into drug-wide tendencies, sample-wide tendencies and deviations specific to each drug–sample combination. The general tendencies form an additive component; deviations from the best additive fit to the measured pairs define the interaction. Recovering a compound's general effect is therefore different from predicting its interaction with a particular cellular context. An aggregate correlation mixes recovery of both components and can be dominated by the component carrying more variance. Strong overall agreement can thus coexist with weak prediction

of context-specific effects[10]. Related evaluations of genetic perturbation models have also found that systematic variation can inflate apparent recovery of perturbation-specific effects[11].

This distinction matters when comparing representations, assessing generalization and judging model improvements. Cross-study analyses have shown that reported performance depends on the response dataset and validation design[12,13]. A representation may improve general response prediction without improving interaction recovery, and its apparent advantage can depend on the decoder that maps it to a response. Cell-line and compound holdouts also withhold measurements of different marginal effects, requiring models to infer sample-wide or drug-wide tendencies, respectively[7,10]. These tests may therefore assess different capabilities, not simply different levels of difficulty. Finally, better agreement with an interaction pattern need not imply correctly scaled effects or lower prediction error. Evaluation must identify both the response structure recovered and the source of remaining error.

Here we introduce DrugDis, a component-resolved framework for evaluating drug-response prediction. DrugDis applies the same orthogonal decomposition to measured and predicted responses on observed drug–sample pairs, separating additive effects from interactions and attributing squared prediction error exactly to these components. It reports interaction direction, amplitude and error separately, with empirical cross-assay reproducibility providing a reference for component recovery where repeated measurements are available. We analyse 3.14 million pairs spanning 986 cancer cell lines and 54,180 compounds under two holdout regimes, with a separate zero-shot evaluation in patient-derived organoids. The analyses reveal how general response structure can dominate aggregate performance and how representation comparisons and error patterns depend on the component and generalization regime examined. Matched supervision controls and an independent-assay model-selection test further distinguish improvements in interaction recovery from gains in overall accuracy or independently measured selectivity.

Results

Aggregate accuracy can hide context-specific recovery

We assembled 3,141,680 observed drug–sample pairs from 986 cancer cell lines and 54,180 compounds across eleven response resources (Fig. 1a,b; Supplementary Table 1). NCI60 supplied 68.5% of the source response rows, so pooled results reflect this uneven resource composition.

On this observed support, the response separates into an additive component containing drug-wide and sample-wide effects and a drug–sample interaction component. The decomposition is orthogonal even though the response matrix is incomplete. The additive subspace occupies only 1.76% of the degrees of freedom but contains 75.8% of response variance (Fig. 1c). Drug-wide effects uniquely account for 71.7% of total variance and sample-wide effects for 3.20%; another 0.97% is shared because the two marginals are not orthogonal on this unbalanced support (Supplementary Note 1; Supplementary Table 2). Additive dominance was not specific to NCI60. Decomposed separately within each of the five largest resources, on that resource's own drug–sample support, the additive component carried from 59.0% (GDSC1 and GDSC2) to 82.5% (NCI60) of response variance while occupying at most 2.44% of the degrees of freedom. It carried 66.5% in the benchmark rebuilt without NCI60 (Supplementary Table 3). The size of the additive share therefore varies with the screening resource.

This additive dominance changes how aggregate performance should be read. Under held-out cell lines, the single-head baseline M0 attained a mean total-response correlation of 0.86 across five seeds. On the same predictions, additive correlation was 0.96 and interaction correlation only 0.32 (Fig. 1d). The measured additive component retained most of the total response's width, whereas the interaction distribution was narrower and centred near zero (Fig. 1e). A high aggregate score could therefore be obtained while recovering relatively little drug–sample-specific variation.

The same projection separates prediction error exactly: $\mathrm{MSE} = E_{\mathrm{add}} + E_{\mathrm{int}}$. Under held-out cell lines, interaction error accounted for 77% of M0's squared error on average (Fig. 1f). We next evaluate this separation under held-out compounds and in zero-shot patient-derived organoids, alongside the held-out-cell-line setting (Fig. 1g).

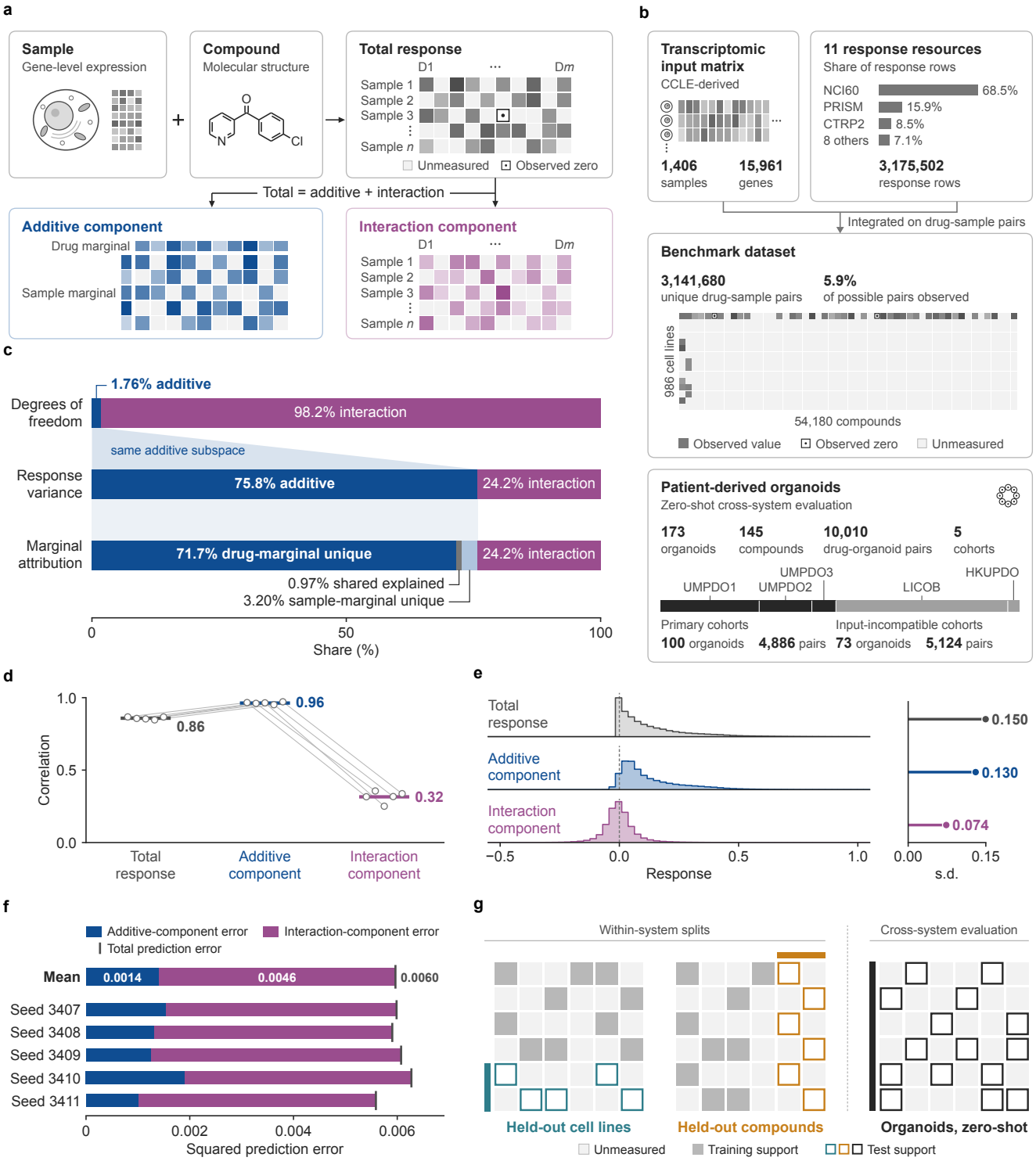

Figure 1: **General response structure dominates variance but not prediction error under held-out cell lines.** **a**, Schematic of the observed-pair decomposition. Grey cells are unmeasured; the outlined cell is a measured zero. Values are illustrative. **b**, Benchmark assembly from 1,406 profiled samples and eleven response resources. Source rows were collapsed to 3,141,680 unique pairs across 986 cell lines and 54,180 compounds. NCI60 supplied 68.5% of source rows. The organoid evaluation comprises 10,010 pairs from 173 organoids; the prespecified primary set contains 4,886 pairs from 100 organoids. **c**, Additive and interaction shares of degrees of freedom and response variance on all observed pairs. Marginal variance shares are drug unique, sample unique and shared explained variation; the last cannot be assigned to either marginal alone. **d**, Total, additive and interaction correlations for M0 under held-out cell lines; circles denote five seeds and bars their means. Test supports contain 474,279–570,863 pairs. **e**, Measured response and component distributions over all pairs; histograms use 60 equal-width bins and the right-hand marks show standard deviations. **f**, Exact additive and interaction shares of M0 squared error by seed; ticks are directly calculated total error. **g**, Split schematics for held-out cell lines, held-out compounds and zero-shot organoids. Shading distinguishes training, test and unmeasured pairs; the organoid test uses cell-line checkpoints.

DrugDis distinguishes interaction direction, amplitude and error

DrugDis projects measured and predicted responses onto the same additive subspace on the observed evaluation pairs. It then compares their interaction components. This differs from global correlation, correlations within a drug or cell line, and DrEval's normalization by training-panel mean effects[7] (Fig. 2a; Supplementary Table 4). The interaction correlation reports direction, the ratio of predicted to measured interaction standard deviations reports amplitude, and interaction R^2 evaluates their combined effect on squared error.

Controlled changes to predictions show why these readings are separate. Across 20 matched split–model–seed cases, adding either drug-only or cell-only offsets changed several comparator scores but left all three DrugDis interaction readings unchanged (Fig. 2b). Rescaling only the predicted interaction by 0.5 or 2 left its correlation unchanged, changed amplitude by the same factor, and generally reduced interaction R^2 (Fig. 2c). The response, pair support and checkpoints were fixed in both tests. Thus a model can preserve the interaction pattern while changing its magnitude and error.

A score also needs a meaningful null. Constant and training-marginal predictors emit no interaction, making their interaction correlation undefined; shuffled predictions emit interaction but score near zero (Fig. 2d; Supplementary Note 2; Supplementary Table 5). On 32,724 pairs measured in both GDSC1 and GDSC2, cross-assay reproducibility was 0.75 for the additive component and 0.50 for the interaction component. Their square roots, 0.87 and 0.71, are empirical references for interpreting recovery on matched supports, not attainable limits (Fig. 2e; Supplementary Table 6).

We also rescored the baseline M0 and a component-supervised variant M4, introduced below, by response resource. Their whole-modification change in interaction correlation depended on resource composition (Fig. 2f; Supplementary Table 7). Under held-out cell lines its mean changed from -0.0081 on all pairs to $+0.0078$ after excluding NCI60. These are rescored subsets of the same predictions; they do not test the matched M3–M4 supervision effect by resource.

Component-resolved evaluation quantifies score separation relative to seed variation

We next applied component recovery to a benchmark of twelve molecular[14–25] and twelve transcriptomic[26–36] representations, eleven of them pretrained and the twelfth the untransformed gene-level expression, under a fixed nonlinear dual-tower decoder and three seeds (Fig. 3a-d; Supplementary Table 8 and Supplementary Table 9). Among the ten full-panel molecular representations, ECFP4 had the highest seed-mean score on all three components under both distribution shifts. GeminiMol and UniMolV2 were evaluated on partial compound supports and are shown descriptively, not included in the full-support molecular score-separation statistic. On the transcriptomic axis no pretrained embedding exceeded the untransformed gene-level expression by more than the seed-to-seed variation; the largest gap, SCimilarity under held-out cell lines, reaches $+0.011$ on the interaction component and less on the other two.

The spread of seed-mean scores relative to seed-to-seed variation is a descriptive property of a benchmark, rather than a consequence of reporting a ranked list, and the two axes differ (Fig. 3e; Supplementary Table 10). Dividing the spread between representations by the median seed-to-seed standard deviation gives 23.6 on total response and 33.0 on the additive component for the molecular axis under held-out cell lines, but only 3.5 on the interaction component; under held-out compounds it is 31.9 to 54.0 on all three. The same ratio for the transcriptomic axis is 3.9 to 5.9 under held-out cell lines. Under held-out compounds, the transcriptomic score spread is 1.8 times the median seed-to-seed standard deviation for total response and for the additive component, and 2.8 times for the interaction component. These ratios describe score separation on the reported runs. They neither test pairwise differences nor establish ranking stability, because they do not estimate the paired covariance of representation scores across seeds.

Decoder dependence was assessed separately in a four-representation comparison at seed 3407 under an additive ridge and a low-rank bilinear decoder (Fig. 3f; Supplementary Table 11 and Supplementary Table 12). On the molecular axis, rank correlations between decoder pairs run from 0.80 to 1.00 on total response and on the additive component, and ECFP4 ranks first under every decoder on all three components. On the molecular interaction component the bilinear and tower orderings agree only at 0.20, swapping the second and fourth places. On the transcriptomic axis, total-response rank correlations fall

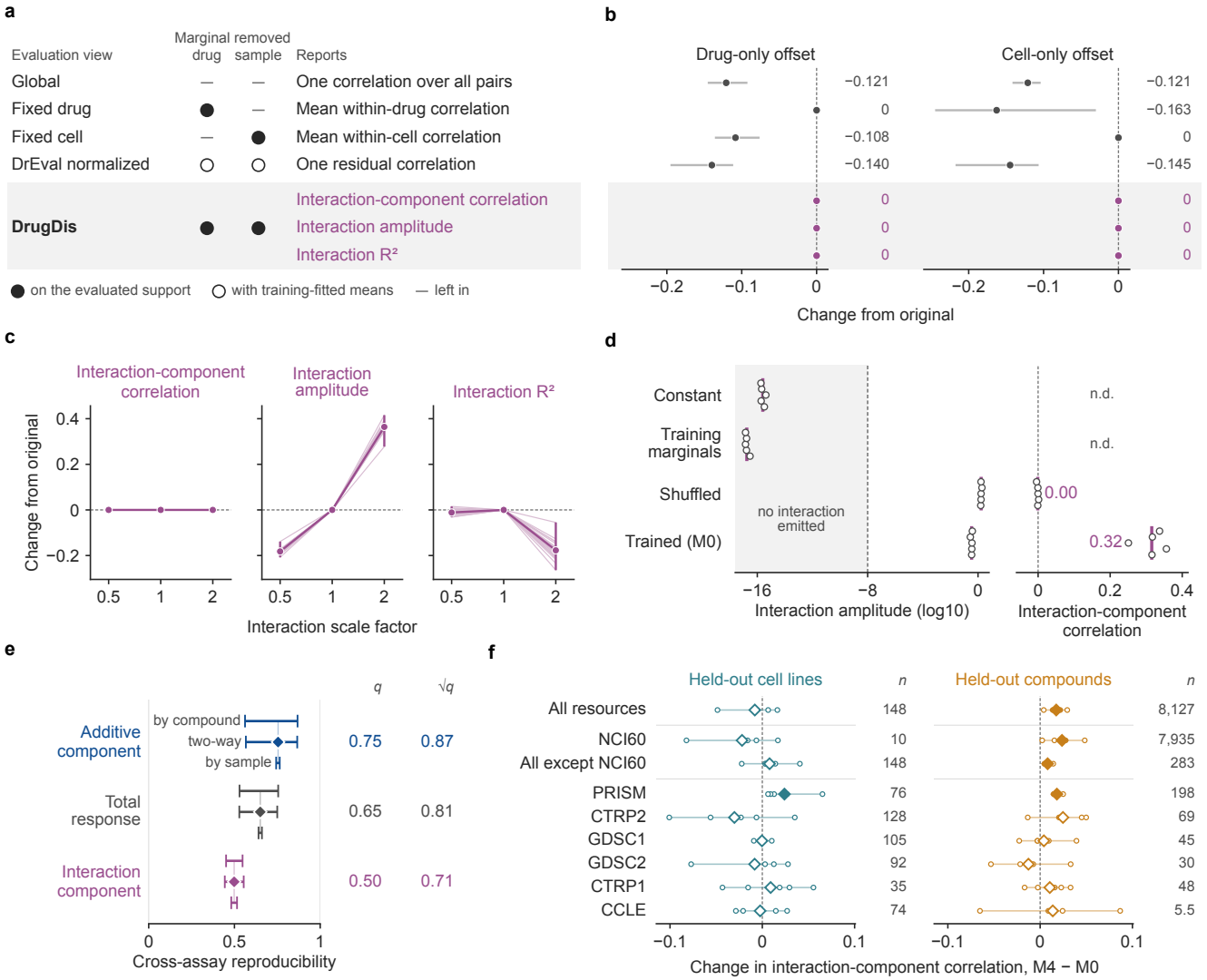

Figure 2: **DrugDis separates interaction direction, amplitude and error. a**, Evaluation views and the drug and sample marginals each removes. DrEval uses training-panel means; DrugDis projects both marginals on the observed evaluation support. **b**, Score changes after drug-only or cell-only offsets to frozen predictions, aligned with the rows of **a**. Points and ranges summarize 20 matched cases (two shifts, M0 and M4, five seeds). DrugDis readings are unchanged by either offset. **c**, Changes after scaling only predicted interactions by 0.5 or 2 in the same cases. Lines denote individual cases, points means and vertical bars ranges. Interaction correlation remains fixed while amplitude scales and interaction R^2 changes. **d**, Interaction amplitude and correlation for constant, training-marginal, shuffled and trained predictors under held-out cell lines (five seeds each). Correlation is undefined when amplitude is below the 10^{-8} non-emission threshold. **e**, Cross-assay reproducibility q_c on 32,724 GDSC1–GDSC2 pairs from 89 compounds and 600 samples. Intervals are 95% cluster bootstrap intervals by compound, by both axes or by sample; $\sqrt{q_c}$ is the empirical recovery reference. **f**, M4–M0 interaction-correlation differences by response resource and shift, rescored on the same predictions. Circles are seed differences, lines ranges and diamonds means; filled diamonds mark agreement in sign across available seeds. Most rows use five seeds; CCLE under held-out compounds uses four.

to 0.40 between the ridge and the tower and to 0.00 between the bilinear decoder and the tower, and the leading representation changes with decoder. On the interaction component the two orderings are nearly reversed, with a rank correlation of -0.80. On that axis the dependence is not confined to the decoder. Across three seeds the nonlinear tower does not reproduce its own ordering: the mean pairwise seed-to-seed rank correlations are -0.20, 0.40 and -0.33 on total response, the additive component and the interaction component, and on no component does the same representation lead at all three seeds. The low-rank bilinear decoder reproduces its own ordering at 0.87, 1.00 and 1.00, exactly on the additive and interaction components. These displayed orderings are therefore conditional on the tested representation set, decoder and seed; they are not a general ranking-stability test.

One decoder cannot be ranked on interaction at all. A ridge over concatenated inputs is additive by construction, and the standard deviation of its predicted interaction component is around 4×10^{-9} (Fig. 3g). A correlation computed there returns 0.0006, which a table of scores displays as a weak model; the emitted amplitude shows it as the undefined quantity it is. This structural classification is independent of the numerical-amplitude threshold used for generic prediction records.

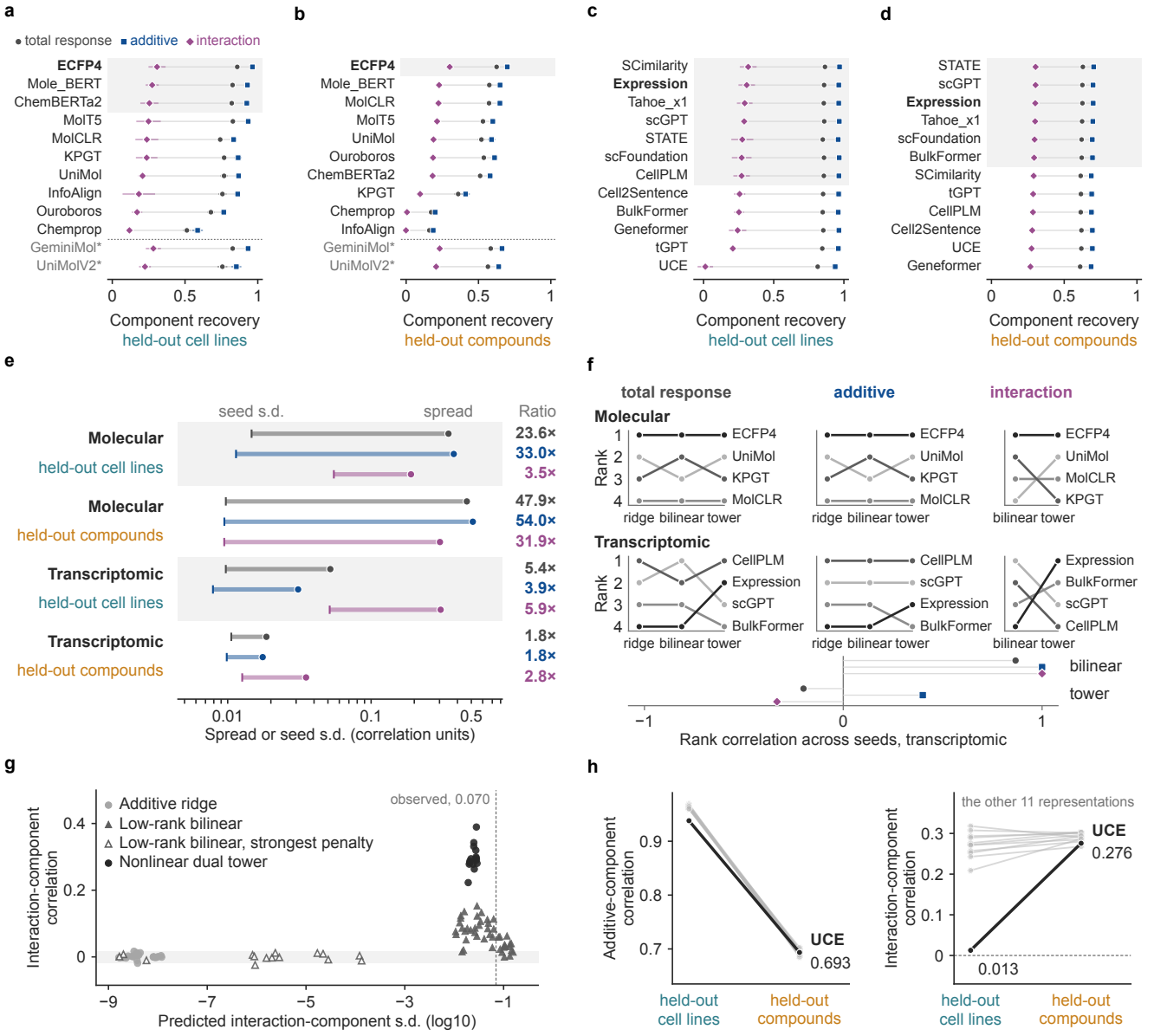

Figure 3: **Component-resolved evaluation distinguishes score separation, decoder dependence and interaction degeneracy. a–d**, Component recovery of twelve molecular (**a**,**b**) and twelve transcriptomic (**c**,**d**) representations under a fixed nonlinear dual-tower decoder, for held-out cell lines (**a**,**c**) and held-out compounds (**b**,**d**). Markers are means over three seeds, bars seed standard deviations and the hairline the span of the three components; rows are ordered by interaction recovery and bold marks the anchor configuration. Expression is gene-level expression, `raw` in the Supplementary Tables. Shading marks the representations within one median seed standard deviation of the leader in interaction recovery. GeminiMol and UniMolV2 (asterisks) cover part of the compounds and are shown below the comparable rows. **e**, Between-representation spread and median seed standard deviation for each input axis, shift and component, on a log axis, so that the bar length is their ratio, printed at the right. The ratio is descriptive; it does not test pairwise separation or ranking stability. **f**, Orderings of four molecular representations (gene-level expression fixed) and four transcriptomic representations (ECFP4 fixed) under three decoders at seed 3407. Lines follow one representation, so crossings are changes in rank. The interaction blocks omit the additive ridge, which emits no interaction (**g**). The strip gives each decoder's mean pairwise rank correlation of its own transcriptomic ordering over the three seed pairs. **g**, Predicted interaction-component standard deviation ($\log_{10}$) against interaction-component correlation for all 103 decoder configurations. The dashed line is the observed standard deviation, 0.070, and the band the correlation range of the 28 additive-ridge configurations. Open triangles mark the low-rank bilinear decoder at the strongest penalty of the selection grid; 14 of its 15 configurations fall in the band. **h**, Additive-component and interaction-component correlation of the twelve transcriptomic representations under both shifts; UCE in black, the other eleven in grey. UCE's additive recovery is 2% below the lowest of the others under held-out cell lines and within their range under held-out compounds; its interaction recovery is 94% below the lowest under held-out cell lines only. Every value is performance under the specified decoder.

Finally, a representation can fail under one shift and not the other. UCE recovers interaction at 0.013 under held-out cell lines and at 0.276 under held-out compounds, while every other transcriptomic representation moves far less between the two (Fig. 3h). A benchmark run in one regime cannot be quoted for another. That the two shifts disagree about a representation, and about score separation relative to seed variation, raises a question the recovery scores cannot answer on their own: whether the shifts differ only in difficulty, or in where the error they leave behind actually sits.

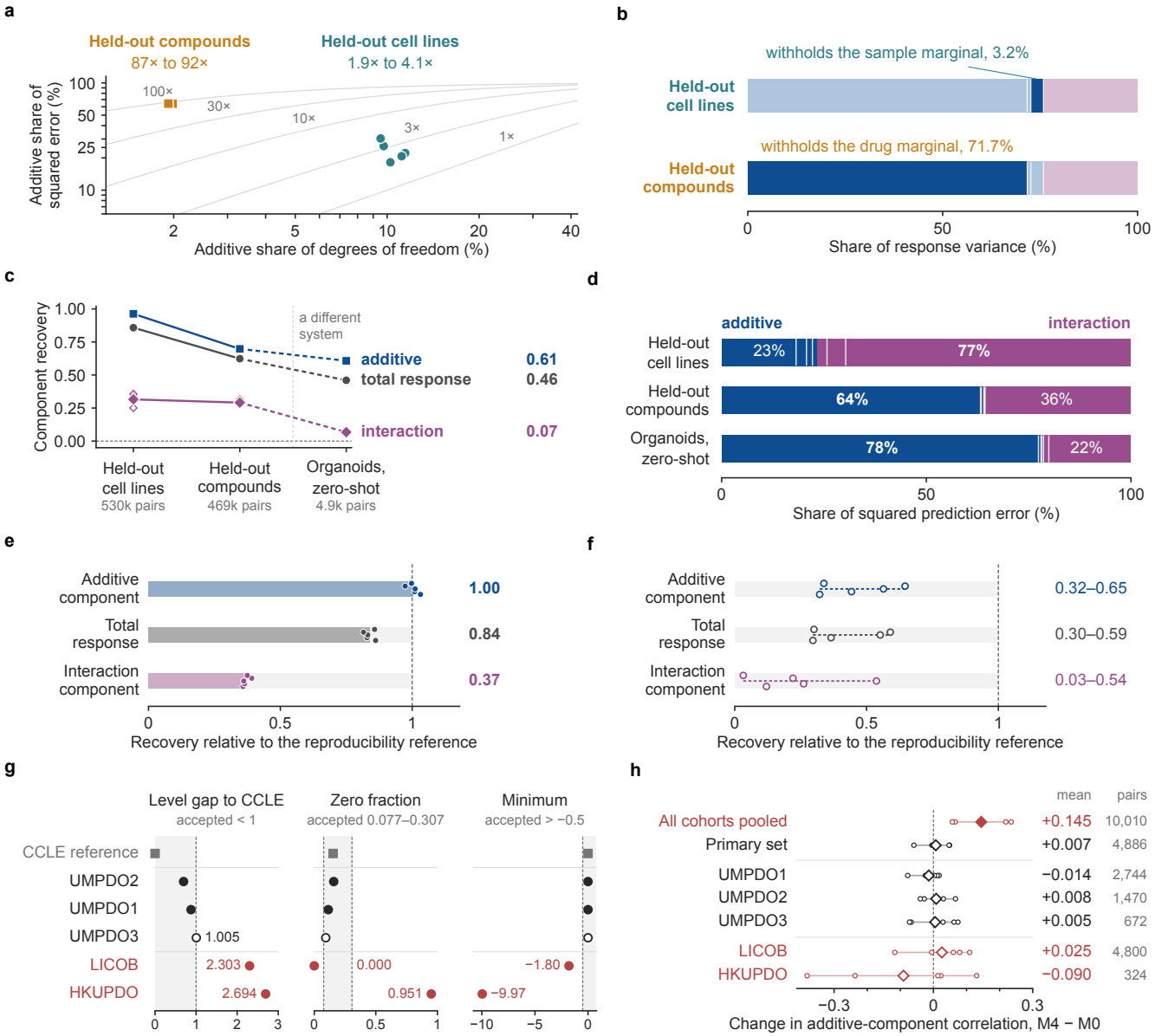

Figure 4: **Generalization regimes differ in component recovery and error allocation. a**, Additive shares of degrees of freedom and of squared error for M0; points are five seeds per regime. Contours mark constant additive-to-interaction error-density ratios; the printed ranges are the measured ratios over seeds. **b**, Share of response variance in the marginal each regime withholds (dark blue); pale blue, the rest of the additive component; pale violet, the interaction component. **c**, M0 total, additive and interaction correlations under held-out cell lines, held-out compounds and zero-shot organoids. Filled markers are seed means and open markers the five seeds; the dashed segment crosses into a different system. **d**, Additive and interaction shares of M0 squared prediction error in the same three regimes; white ticks mark the five seeds. **e**, Recovery relative to the empirical cross-assay reproducibility reference under held-out cell lines, on 2,876–3,051 matched pairs per seed. Bars are seed means and dots the five seeds; the dashed line marks the reference. **f**, The same ratios under held-out compounds, shown as seed ranges without a point estimate because only 7–19 compounds remain on the matched support per seed. **g**, Organoid transcriptomic inputs against the three prespecified CCLE-based conditions. Shading marks each accepted window, and values are printed where a cohort falls outside it. UMPDO1–3 (black) form the prespecified primary set; the open marker is UMPDO3, which misses the level condition by 0.005. LICOB and HKUPDO (red) are input-incompatible. **h**, Paired M4–M0 changes in additive-component correlation by cohort and after pooling. Circles show five seeds, lines their range and diamonds their means; filled diamonds mark a common sign across seeds. The primary set contains 4,886 pairs.

Held-out cell lines and compounds concentrate error in different components

Component correlations reveal what a model recovers, but not where its remaining error lies. We compared each component's share of squared prediction error with its share of degrees of freedom for the baseline M0 (Fig. 4a,d; Supplementary Table 13). Under held-out cell lines, interactions absorbed 77% of squared error. Under held-out compounds, the additive component absorbed 64%, despite occupying only about 2% of the evaluation degrees of freedom. Its error density was 87–92 times that of the interaction component across five held-out-compound seeds.

The regimes withhold different marginal effects (Fig. 4b). Sample-wide effects account for 3.20% of total response variance, whereas drug-wide effects account for 71.7%. The contrast in error allocation

is consistent with these unequal variance shares, but two regimes do not establish a causal relationship. M0's additive correlation fell from 0.96 for held-out cell lines to 0.70 for held-out compounds; interaction correlation was 0.32 and 0.29, respectively (Fig. 4c,d). Thus the holdouts exposed different predictive weaknesses, not one uniform change in accuracy.

The empirical cross-assay reference provides another comparison under held-out cell lines. On matched repeated-measurement supports, mean additive recovery was 1.00 times its reference and interaction recovery 0.37 times its own (Fig. 4e; Supplementary Table 14). Under held-out compounds, only 7–19 compounds per seed remained on the matched cross-assay support. We therefore do not interpret a measurement-normalized point estimate for that regime (Fig. 4f).

Zero-shot evaluation on patient-derived organoids introduced both a new biological system and new transcriptomic inputs. M0 correlations fell to 0.61 for additive effects and 0.07 for interactions (Fig. 4c). We assessed input comparability before interpreting transfer: two cohorts met all three prespecified expression-distribution criteria, and a third was a prespecified near-miss included in the primary analysis (Fig. 4g; Supplementary Table 15). Pooling all five cohorts, including two input-incompatible cohorts, inflated the apparent M4–M0 change in additive correlation from $+0.007$ to $+0.145$ (Fig. 4h). We therefore report the primary organoid effect on the prespecified three-cohort support.

Component supervision improves interaction direction in selected regimes

To test whether component-level diagnosis could guide a targeted model change, we compared three predictors on the benchmark (Fig. 5a). M0 has one prediction head. M3 and M4 have the same two-head architecture and 20,029,444 trainable parameters; only M4 receives separate additive and interaction training targets. The M3–M4 contrast therefore isolates component supervision, while M0–M4 includes the architectural change. All arms used the same split manifests, five seeds and validation-MSE checkpoint rule (Supplementary Table 16).

Relative to M3, M4 raised interaction correlation by $+0.029$ under held-out compounds and $+0.026$ in zero-shot organoids, with the same direction in all five seeds of each regime (Fig. 5b,c). The held-out-cell-line change was $+0.001$. Under held-out compounds, additive correlation fell by 0.015 across all seeds, and aggregate correlation did not improve. Component supervision thus improved the interaction pattern in two regimes without a general gain in measured response prediction.

Interaction error told a narrower story (Fig. 5d). M4–M3 interaction R^2 rose by $+0.020$ under held-out compounds across all five seeds, but changed by only $+0.002$ on average in organoids. Organoid interaction R^2 stayed below zero for every arm: none outperformed a zero-interaction prediction in squared error. Improved direction alone was insufficient to recover interaction magnitude accurately in that system.

We finally asked whether a component-specific selection rule improved an independently measured endpoint. Six criteria selected M0 or M4 from the same GDSC1 predictions, and each selected arm was evaluated on GDSC2 without refitting (Fig. 5e,f; Supplementary Table 17). DrugDis interaction correlation selected a different arm from aggregate correlation in two of five held-out-cell-line seeds and four of five held-out-compound seeds. Those changed choices did not reduce the prespecified GDSC2 double-difference selectivity error. Mean absolute error instead increased by 0.00245 and 0.00043, respectively, relative to aggregate-correlation selection (Fig. 5g). Secondary sign accuracy did not change that primary-endpoint verdict (Fig. 5h).

Discussion

High drug-response accuracy need not imply accurate prediction of context-specific effects. In this benchmark, and within each of its five largest response resources, most response variation was additive, allowing models to reproduce broad drug and sample tendencies while recovering substantially less of their interaction. Recent benchmarking also found strong performance from mean-effect baselines in drug-response prediction[7]. DrugDis separates these capabilities and attributes prediction error to the corresponding components. Both components are defined on the observed drug–sample pairs. The interaction component is the deviation from the best additive fit to those pairs, so measuring different compounds, samples or pairs changes it; it should not be read as an intrinsic biological interaction. Reporting interaction amplitude alongside correlation distinguishes weak recovery from predictions that

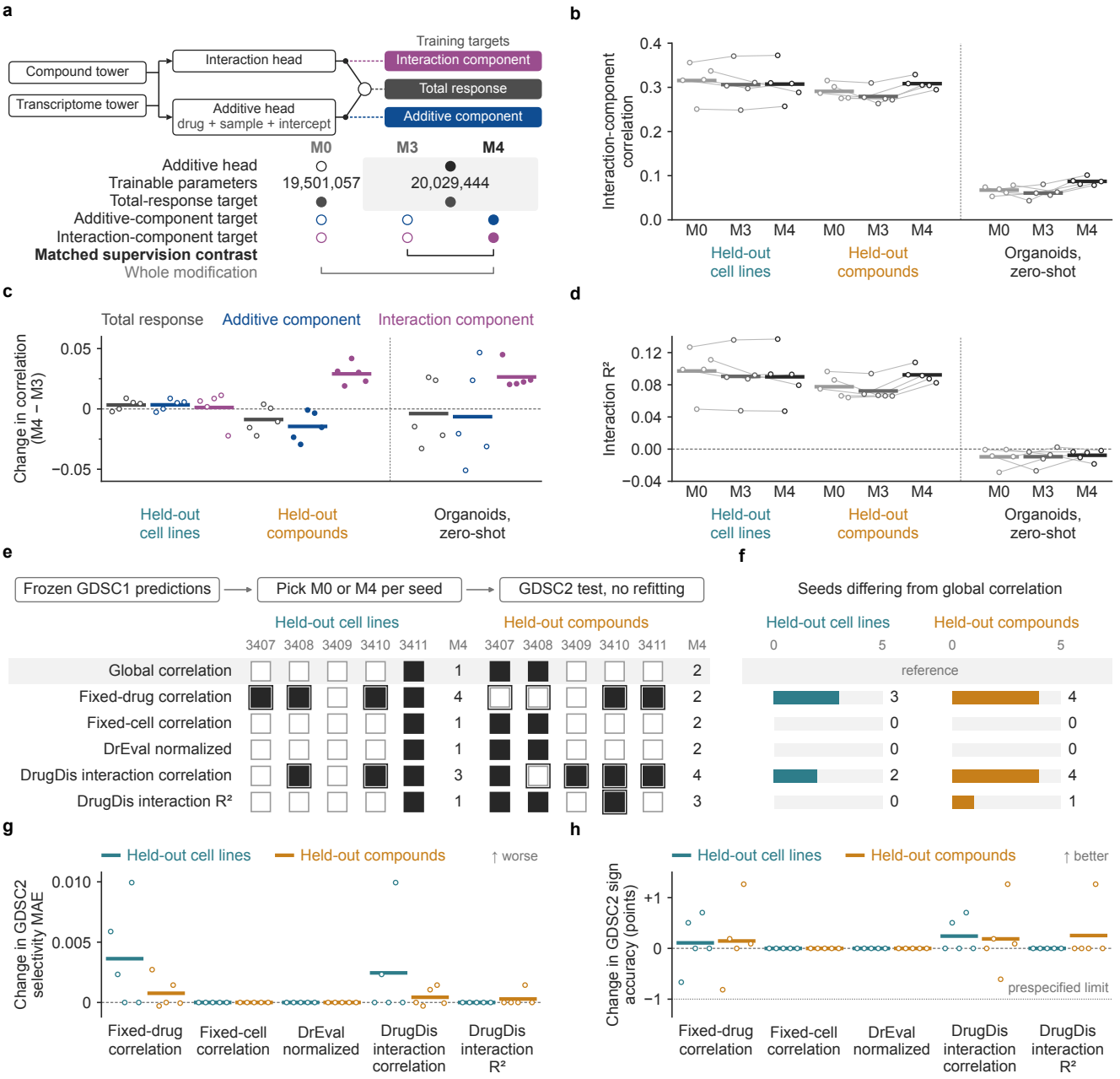

Figure 5: **Component supervision improves interaction direction without improving the independent selectivity endpoint. a**, Model arms. M0 has one head; M3 and M4 share a 20,029,444-parameter two-head architecture and differ only in component-specific training targets. **b**, Interaction correlations for M0, M3 and M4. Lines are means of five seeds; circles are seeds linked across arms. Organoids use held-out-cell-line checkpoints. **c**, Paired M4–M3 changes in total, additive and interaction correlation. Points are seeds and bars their means; filled points indicate the same sign across all five seeds. **d**, Interaction R^2 for the three arms, displayed as in **b**. The dashed zero line is the squared-error score of a prediction with no interaction. **e**, GDSC1 selection of M0 or M4 by six criteria for each of five seeds under two holdout regimes. Filled squares select M4; rings mark choices differing from aggregate correlation. Selected arms are tested on GDSC2 without refitting. **f**, Counts of differing choices from **e**. **g**, Change in prespecified GDSC2 double-difference selectivity mean absolute error relative to aggregate-correlation selection; positive values are worse. **h**, Secondary GDSC2 sign-accuracy change; positive values are better. In **g** and **h**, circles are seeds and lines are their means.

contain no interaction, while empirical cross-assay reproducibility places recovery in the context of measurement consistency. These distinctions sharpen the interpretation of representation benchmarks. Our comparisons showed that apparent advantages depended on the evaluated component, decoder and, in some cases, training run. A representation's performance therefore reflects both the information it provides and how that information is extracted and evaluated. Component-resolved evaluation identifies whether an apparent advantage reflects general response prediction, context-specific recovery or both.

Generalization is not simply a progression from easier to harder prediction tasks. Holding out cell lines and holding out compounds remove different sources of response information. In this benchmark, drug-level tendencies carried much more variation than sample-level tendencies. Prediction error was

dominated by interactions for held-out cell lines, but by additive effects for held-out compounds. The two regimes therefore exposed different weaknesses rather than a uniform change in predictive accuracy. Success under one holdout strategy does not establish the same capability under another. Cross-system evaluation adds the question of input comparability: performance in organoids must be interpreted in light of both the new biological system and the characteristics of the transcriptomic inputs supplied to the model. Generalization claims are more informative when they specify the unseen axis, the response structure being predicted and the conditions under which transfer was tested.

Model improvement likewise depends on which capability is being assessed. With architecture and parameter count matched, component supervision improved interaction direction for held-out compounds and in zero-shot organoid transfer, but consistently reduced interaction error only for held-out compounds. In organoids, interaction R^2 remained below zero: improved pattern agreement was still insufficient to outperform a zero-interaction prediction in squared error. Different evaluation criteria also changed which model was selected, without improving the prespecified GDSC2 selectivity-error endpoint relative to aggregate-correlation selection. Improving a component-specific score is therefore distinct from improving an independently assessed outcome.

The empirical reproducibility reference was estimable on a sufficiently repeated held-out-cell-line support, but the held-out-compound overlap was too limited for the same comparison. The benchmark was anchored to one transcriptomic input source, and NCI60 supplied most response rows, although additive dominance held within each of the five largest resources. The organoid analysis used a prespecified primary cohort that included one near-miss on the input-compatibility criteria. The matched supervision result concerns one architecture and objective, while the independent decision test compares two assay generations within one prediction pipeline. The variance shares and effect sizes observed here should therefore be examined across further response resources, biological systems and model families. Component-resolved evaluation shifts the emphasis from whether a score increased to which predictive capability improved, where that improvement generalizes and whether it benefits an independently measured outcome.

Methods

Benchmark dataset and transcriptomic input matrix

The benchmark dataset comprises 3,141,680 unique drug–sample pairs on 986 cancer cell lines and 54,180 compounds. Each pair carries a single scalar from the DROMA column `Sensitivity`, numerically in $[0, 1]$. The processing pipeline reshaped the per-resource response tables and applied no rescaling. The available source metadata do not supply a uniform per-resource assay-metric definition or biological direction for this harmonized field; we therefore call it DROMA `Sensitivity` throughout, rather than relabel it as AAC, viability or a response direction. Compounds are keyed by canonical SMILES string and samples by `Sample_ID`. Eleven drug-response resources contribute rows. NCI60 supplies 2,175,581 rows (68.51%), PRISM 504,493 (15.89%) and CTRP2 269,725 (8.49%). GDSC1, GDSC2 and CTRP1 supply 125,914 (3.97%), 44,797 (1.41%) and 39,784 (1.25%). CCLE supplies 10,730 (0.34%), and GRAY, gCSI, FIMM and UHNBreast together supply 4,478 (0.14%). These 3,175,502 rows exceed the 3,141,680 unique pairs because the response table is keyed by original drug name and sample. A compound recorded under more than one name therefore contributes one row per name. This affects 32,109 pairs, of which 17,044 draw their rows from a single resource. The processed response table holds one row per original drug name and sample. Where several resources measured the same name on the same sample, it keeps the value of the first resource in the fixed order CCLE, CTRP1, CTRP2, FIMM, GDSC1, GDSC2, GRAY, NCI60, PRISM, UHNBreast, gCSI and omits the others; the kept values equal the per-resource DROMA tables exactly. This overlap rule set aside 244,227 otherwise eligible rows, chiefly from PRISM (109,125), GDSC2 (43,710) and GDSC1 (41,972), and the row counts above are after it. Rows sharing a canonical pair were collapsed to their arithmetic mean when the dataset was assembled. The composition is reported rather than balanced: NCI60 contributes 68.51% of all rows on 67 cell lines (Fig. 1b).

The sample-side model input is a single transcriptomic input matrix derived from CCLE. It holds gene-level expression for 1,406 samples over a whitelist of 15,961 genes, in float32 and with no missing entries. The gene set is the sorted intersection of the CCLE and GDSC gene sets. Values range from 0.0000 to 16.5294, and 15.33% of entries are exactly zero. Per-sample mean expression spans 1.064 to

3.344, with a standard deviation of 0.221. No transformation was applied when the matrix was built, and none was applied afterwards. The 986 samples of the benchmark dataset are the subset of these 1,406 profiles that also meet the further criteria below.

Dataset assembly, provenance and cohort inclusion criteria

A sample is eligible for the benchmark dataset only if a CCLE expression profile exists for it under the specified gene set. For eligible samples, all otherwise eligible response observations were retained across every resource, subject to the overlap rule above. The transcriptomic input source is standardised; the response source is not restricted. Two further requirements applied. The processed master sample table had to type the sample as a cell line. A compound had to carry a molecular representation, here a 2048-dimensional binary ECFP4 fingerprint. Together the rules admitted 986 of 1,303 samples, 54,180 of 54,181 compounds and 3,141,680 of 3,964,393 pairs.

We used CCLE as the sole source of cell-line expression to avoid combining profiles with different measurement scales. Tavor contains patient-derived cells and was excluded from the cell-line benchmark; none of its samples had an eligible CCLE expression profile.

Patient-derived organoid cohorts are not part of the benchmark dataset. They enter only the zero-shot cross-system evaluation, under the prespecified input-compatibility criteria given below.

Orthogonal response decomposition on an observed support

A drug-response dataset is a partially observed response surface over compounds and biological samples. Let Ω denote the set of observed drug–sample pairs and $N = |\Omega|$ its size. The additive design matrix on that support is $X = [\,\mathbf{1},\, X_{\text{drug}},\, X_{\text{sample}}\,]$. Its columns are an intercept and one indicator for each drug level and each sample level present in Ω. Write $H = X(X^\top X)^+ X^\top$ for the orthogonal projection onto the column space of X, and $M = I - H$ for the projection onto its orthogonal complement. Any response vector y on Ω then splits as $y = Hy + My$. We call Hy the additive component and My the drug–sample interaction component; the two are orthogonal by construction, so $\mathrm{Var}(y) = \mathrm{Var}(Hy) + \mathrm{Var}(My)$. Variance shares are computed on observed entries alone and need no value at an unobserved pair.

The operator is a function of the support alone. H depends on which pairs were observed and not on the responses measured at them, so the identical map applies to any vector defined on Ω, including a prediction. The interaction component is therefore support-indexed. It is the deviation from the best additive explanation on a stated drug and sample support. Changing that support changes the quantity, in the same way that the reference population fixes what any fixed-effect adjustment removes. The dimension of the additive subspace is $\mathrm{rank}(H) = n_{\text{drug}} + n_{\text{sample}} - c$, where c counts the connected components of the bipartite support graph. We obtained c from a sparse connected-components routine on that graph. The support was connected for the benchmark dataset and for every test set, so $c = 1$ in each case. On the full benchmark dataset this gives $\mathrm{rank}(H) = 55{,}165$ against 3,141,680 observations, or 1.76% of the available degrees of freedom.

Neither H nor the pseudo-inverse was formed. H is $N \times N$ with N up to 3,141,680, and X carries 55,167 columns on the full dataset, so a dense solution is not available at this scale. We computed My instead by the method of alternating projections, the standard device for absorbing high-dimensional fixed effects[37]. Each sweep subtracts the current level means of each factor in turn from the running vector, using grouped sums over integer factor codes. The intercept is absorbed automatically, because each factor's indicators span the constant vector. Sweeps continued until the largest per-observation change fell below a tolerance of 10^{-10}, under a cap of 2,000 sweeps. On the two-factor design the solver required 31 to 210 sweeps across the 91 model-by-evaluation-set records, so the cap did not bind. We checked the operator numerically with random vectors, on the two matched cross-assay supports of 5,280 and 4,453 rows. Linearity held to 8.0×10^{-11}, idempotence to 1.9×10^{-11}, and the cosine between Hv and Mv to 2.3×10^{-13}. Across the benchmark dataset and the ten test sets, the worst departure from the variance identity was 2.7×10^{-11} of $\mathrm{Var}(y)$, and the worst $|\mathrm{corr}(Hy, My)|$ was 3.2×10^{-11}.

We also applied the decomposition within each response resource. A resource's measurements were read from its DROMA table, so the overlap rule did not apply, and were restricted by the inclusion criteria of the benchmark dataset. Rows sharing a canonical pair were averaged within the resource, no value from another resource entered its support, and H was built on that resource's observed pairs. The

benchmark without NCI60 was rebuilt from the other ten resources by the rules that built the benchmark dataset. Every one of these supports was connected. Because the additive subspace contains the constant vector, replacing a resource's responses y by $ay + b$ with $a \neq 0$ leaves its variance shares unchanged, so the shares do not depend on the unit or direction of that resource's metric; they do depend on its screening design. Before any resource was decomposed, the same code rebuilt the processed response table from the DROMA tables row for row and the benchmark dataset pair for pair, and reproduced the benchmark decomposition to a relative difference of 3×10^{-15} (Supplementary Table 3).

Component-wise recovery and zero-amplitude predictions

Observed responses and predictions were decomposed by the same operator, each vector losing its own additive structure. For an evaluation set with observed responses y and predictions $\hat{y}$ on the support Ω, we formed Hy, My, $H\hat{y}$ and $M\hat{y}$. No single label-derived baseline vector was subtracted from both. This is the property that keeps the comparison symmetric, since no term in the decomposition of the prediction is estimated from the observed labels. Three Pearson correlations follow, and all three are reported for every model and every evaluation set:

$$r_{\text{total}} = \text{corr}(y, \hat{y}), \qquad r_{\text{add}} = \text{corr}(Hy,\, H\hat{y}), \qquad r_{\text{int}} = \text{corr}(My,\, M\hat{y}).$$

A correlation was returned only where both vectors had at least three entries and non-zero standard deviation. We also recorded the emitted interaction amplitude, $A_{\text{int}} = \text{sd}(M\hat{y})/\text{sd}(My)$, which a correlation does not carry. We additionally report the amplitude-sensitive fraction of interaction variance explained,

$$R^2_{\text{int}} = 1 - \frac{\|M\hat{y} - My\|^2}{\|My\|^2} = 2A_{\text{int}} r_{\text{int}} - A^2_{\text{int}},$$

where the equality holds after centring on the evaluated support. This quantity can be negative and is not the square of the correlation (Supplementary Table 18).

That amplitude separates two cases a correlation would otherwise score alike. A constant predictor, and any predictor whose output lies in the additive subspace, satisfies $M\hat{y} = 0$ up to floating point. Its interaction-component correlation is mathematically undefined, and a Pearson formula evaluated there returns the correlation of rounding noise rather than a calibrated zero. For generic prediction records, a record was marked as emitting no interaction when $A_{\text{int}} \leq 10^{-8}$. A decoder that is structurally additive by construction is instead declared non-emitting by construction, independently of its finite-precision residual, and its interaction-component correlation is reported as undefined. A predictor that varies within the additive subspace, such as one built from training-set marginals, still has a well-defined total-response and additive-component correlation. A constant predictor does not, because its predicted vector carries no variance in any subspace.

The threshold falls in an empty region of the measured distribution. The null-calibration records cover the test sets under both distribution-shift regimes and the primary organoid cohort. Predictors that emit no interaction reached A_{int} between 9.4×10^{-18} and 1.8×10^{-9}. The emitting records ranged from 0.2775 to 1.9275. The M0 model records ranged from 0.2775 to 0.3876, whereas shuffled predictions ranged from 1.3480 to 1.9275 times the observed interaction amplitude.

Formal comparison with existing evaluation views

The formal comparison used the checkpoints selected by minimum validation mean squared error for M0 and M4, both held-out-axis regimes and seeds 3407 to 3411. All metrics within a run were computed on identical pair-mean predictions and support. Global Pearson correlation, root mean squared error and R^2 were computed over all pairs. Following the fixed-entity views evaluated in the specification-game framework[10], we also computed Pearson correlation within each drug and within each cell line, retained groups with at least three pairs and reported the unweighted mean over groups. For DrEval normalized performance[7], the naive mean-effects predictor was fitted only on the corresponding training panel. Its prediction was subtracted from both the observed response and the model prediction before normalized Pearson correlation and R^2 were computed. The implementation followed the public DrEval source at commit `dec421b`. DrugDis quantities used the projection defined above on the same evaluation support.

Five prediction perturbations were fixed before the comparison. The original prediction was left unchanged. Drug-only and cell-only offsets were deterministic, centred effects scaled to a standard deviation of half the observed response standard deviation and added to the corresponding entity. For amplitude tests, the projected additive prediction was retained and the projected interaction prediction was multiplied by 0.5 or 2. The observed response, test support and model checkpoint never changed. Composition sensitivity was computed separately for every response resource carrying at least 1,000 matched rows. A second analysis removed NCI60 and recomputed pair means before all quantities were re-evaluated. Neither analysis is an external-generalization test because it rescored existing predictions.

Independent assay-generation model-selection test

The independent test was prespecified. GDSC1 was the selection arm and GDSC2 the independent test arm. The raw GDSC matching universe contains 43,622 keys with one measurement in each assay generation, across 89 drugs and 799 samples. This is not the per-seed prediction support. For each split and seed, M0 and M4 prediction rows were inner-joined to that raw matching universe, and their supports were required to match exactly. The resulting supports ranged from 4,927 to 5,411 pairs under leave-cell-line-out (mean 5,138) and from 2,723 to 7,164 under leave-small-molecule-out (mean 5,027). On each such support, GDSC1 supplied the selection criteria and GDSC2 supplied the independent endpoint. Selectivity was represented by the double difference $(y_{d_1s_1} - y_{d_1s_2}) - (y_{d_2s_1} - y_{d_2s_2})$, which cancels both additive marginals. For each split and seed, at most 50,000 unique observed-support rectangles were sampled using identifiers and support alone, with generator seed 20260911; neither response values nor predictions entered the sampling rule.

Six GDSC1 criteria independently selected M0 or M4 for each split and seed: global Pearson correlation, mean fixed-drug correlation, mean fixed-cell correlation, DrEval normalized correlation, DrugDis interaction correlation and DrugDis interaction R^2. Ties selected M0. The selected arm was then evaluated without refitting on the corresponding GDSC2 double differences. The primary endpoint was mean absolute error; sign accuracy, Pearson correlation and Spearman correlation were secondary. The endpoints for each candidate model are reported in Supplementary Table 19. The prespecified success rule required a DrugDis criterion to lower the mean paired primary error relative to global- correlation selection in at least one split without decreasing mean sign accuracy by more than 0.01. All criteria, seeds and endpoints were retained regardless of direction.

Exact attribution of squared prediction error

Let $e = \hat{y} - y$ be the prediction error on the evaluation support. Because H and M are complementary orthogonal projections, with $H + M = I$ and $HM = 0$, the error splits as $e = He + Me$ with no cross term, and $\|e\|^2 = \|He\|^2 + \|Me\|^2$. We report the two terms per observation,

$$E_{\text{add}} = \tfrac{1}{N}\|H(\hat{y} - y)\|^2, \qquad E_{\text{int}} = \tfrac{1}{N}\|M(\hat{y} - y)\|^2,$$

so that $\text{MSE} = E_{\text{add}} + E_{\text{int}}$ exactly. Alongside them we report the additive share of squared error, $E_{\text{add}}/(E_{\text{add}} + E_{\text{int}})$, which states where the error of a given predictor lies.

This is an algebraic identity and not an empirical approximation. It follows from the projection being orthogonal, so it holds for any predictor and any observed support. Nothing is fitted when the split is taken, and no distributional assumption enters. The same operator that defines the two response components defines the two error terms. That is what allows prediction error to be located in one subspace or the other without a modelling step.

Numerically, the relative departure $|\text{MSE} - E_{\text{add}} - E_{\text{int}}|/\text{MSE}$ was at most 7.3×10^{-11} across the 91 model-by-evaluation-set records spanning cell-line and organoid evaluations, below the solver tolerance of 10^{-10}. One consequence carries into training. Adding $\lambda\|M(\hat{y} - y)\|^2$ to a squared-error objective, with M built on the training support, reweights one half of an exact partition. It does not add an unrelated auxiliary task.

Cross-assay reproducibility estimation and matched supports

A repeated measurement is a drug-sample pair whose response was recorded twice by separate screening efforts. The reference used here comes from GDSC1 and GDSC2, two assay generations run by one consortium. Their measurements arrive as consecutive row blocks of the GDSC source file, GDSC1 below

row 229,420 and GDSC2 at or above it. Pairs carrying exactly two rows, one in each block, were kept, giving arm A (GDSC1) and arm B (GDSC2). The estimated quantity is therefore within-programme, cross-assay-generation reproducibility, not technical replication. Pairings across different resources were also estimated, but each spans only 389 to 3,174 pairs on fewer than 20 compounds. Each also adds protocol divergence to measurement noise, so none was adopted as the reference.

The full cross-assay reference first restricts the raw matching universe by benchmark *key* eligibility. A pair enters that key set only if its compound carries a molecular representation and its sample carries a CCLE baseline transcriptomic profile, is typed as a cell line and is not in the excluded project. Of the 43,622 raw matched pairs, 32,724 meet these key criteria, over the same 89 compounds and 600 samples, and the full-support reproducibility estimates were computed on them. Key eligibility alone does not assert that the benchmark target for every one of these pairs equals arm A. The distinct target-equality filter used for per-seed recovery is specified below; the complete data-flow and denominator map is in Supplementary Note 3.

The additive projection was built once on that support and applied identically to both arms. Cross-assay reproducibility is then $q_{\text{total}} = \text{corr}(y_A, y_B)$ for total response, $q_{\text{add}} = \text{corr}(Hy_A, Hy_B)$ for the additive component and $q_{\text{int}} = \text{corr}(My_A, My_B)$ for the interaction component. The point estimates are 0.650, 0.754 and 0.498. Observations sharing a compound or a sample are not independent, so a row-level interval would understate the uncertainty. We therefore ran three cluster bootstraps[38] of 600 replicates each, by compound, by sample and two-way. Repeated cluster draws were retained with multiplicity and relabelled as distinct entities. In the two-way scheme, compounds and samples were drawn independently, following the pigeonhole bootstrap[39]; an observed pair enters with multiplicity equal to the product of its compound and sample draw counts. The decomposition was rebuilt inside every replicate, and replicates below 300 rows were discarded. The compound axis dominates the uncertainty: for total response the sample bootstrap returns 0.640 to 0.660 across 600 clusters, against 0.529 to 0.755 for the compound bootstrap across 89.

Recovery relative to the reproducibility reference is $\eta_c = r_c / \sqrt{q_c}$, where r_c is one of r_{total}, r_{add} and r_{int}. Its interval comes from a joint two-way cluster bootstrap of 400 replicates. Each replicate resamples the two axes once and recomputes both the model correlation and the reproducibility reference on that resample, so numerator and denominator carry their uncertainty together. The quantity $\sqrt{q_c}$ is an empirical reference under the stated classical measurement model (Supplementary Note 3) and not a demonstrated limit. Under leave-cell-line-out evaluation the additive η averaged 1.00 over five seeds, with a range of 0.97 to 1.03, so it sits at the reference rather than below it. One reason the additive η sits at the reference is procedural: the benchmark target is arm A and compounds are not held out in that regime, so compound-level structure specific to that arm is available to the predictor. Two further properties argue against interpreting $\sqrt{q_c}$ as an attainable limit. The two arms differ in assay generation, so q_c mixes measurement noise with an assay change. And on the matched pairs q_{int} rises from 0.08 in the smallest quartile of interaction magnitude to 0.62 in the largest. A single scalar per component therefore summarizes rather than bounds.

The reference exists only where measurements repeat, so each regime carries its own matched support. A prediction row was scored against the reference only when it sits on a matched pair. Its stored target also had to match arm A to floating-point tolerance, with an absolute tolerance of 10^{-6}. That filter removes rows which sit on a matched pair but carry another resource's measurement. Under leave-cell-line-out evaluation the support holds 2,876 to 3,051 pairs per seed, over 79 to 86 compounds and 90 to 101 held-out cell lines, and we report η_c there. Under leave-small-molecule-out evaluation it holds 1,970 to 3,873 pairs over 588 to 598 samples, but only 7 to 19 held-out compounds. With so few compound clusters the estimate is unstable. The interaction ratio ranges from 0.03 to 0.54 across the five seeds, while the additive ratio ranges from 0.32 to 0.65. The per-seed intervals for total response and for the additive component include zero at three or more seeds in either training arm. We therefore report no measurement-normalized value in that regime. No repeated cross-assay measurements are available for the organoid cohorts, so no reference could be constructed there.

Rank-normalized error density and the error-density ratio

The exact attribution returns two error masses, E_{add} and E_{int}, but the two subspaces are of very different size. Comparing the masses alone would confound where error concentrates with how much room each subspace has, so each mass was normalized by the dimension of its subspace.

We formed the bipartite graph whose vertices are the n_d compounds and n_s samples of an evaluation support, and whose edges are its n observed pairs. The dimension of the additive subspace is $d_H = n_d + n_s - c$, where c is the number of connected components of that graph. The interaction subspace then occupies the complementary $d_M = n - d_H$ dimensions. Computing d_H from the graph avoids forming H, which is never materialized: the additive design spans 55,166 levels over the 3,141,680 pairs of the full benchmark dataset. Every support on which the ratio was computed had a single connected component, so $d_H = n_d + n_s - 1$ there.

The share of squared prediction error absorbed by the additive subspace is $f_H = E_{\text{add}}/(E_{\text{add}} + E_{\text{int}})$, which by the exact attribution is E_{add} as a fraction of total prediction error. The rank-normalized error densities are $D_H = n\,E_{\text{add}}/d_H$ and $D_M = n\,E_{\text{int}}/d_M$, that is squared error per degree of freedom in each subspace. Their ratio is the additive-to-interaction error-density ratio $C_{\text{error}} = D_H/D_M$. It is dimensionless and unchanged by a common rescaling of the response. It equals one when error is spread in proportion to subspace size. A value above one means the additive subspace carries more error per degree of freedom than the interaction subspace. All four quantities are properties of one predictor on one support. They were computed on each held-out cell-line evaluation support, separately for every seed and both training arms, with the same operator that defines the attribution.

Distribution-shift regimes and data splitting

We evaluated under two distribution-shift regimes, each defined by the axis it withholds. Leave-cell-line-out (LCLO) evaluation partitions the benchmark dataset by cell line. Every response measured on a held-out cell line is absent from training, so the sample marginal of that cell line is never observed. Leave-small-molecule-out (LSO) evaluation partitions by compound, so the drug marginal of a held-out compound is never observed. Both partitions act on entities and not on pairs, so no test pair shares its held-out entity with any training pair.

Entities were split 70:15:15 into training, validation and test sets, by holding out 15% of entities for test and then 15% of the total for validation. The entity list was deduplicated and sorted before splitting, so a partition is a function of the entity set and the seed alone, not of table row order. The manifests cover five seeds, 3407 to 3411, in each regime. The component-supervised comparison used all five seeds; the representation benchmark used the first three. Every seed yields 690, 148 and 148 cell lines under LCLO, and 37,925, 8,128 and 8,127 compounds under LSO. Pair counts vary across seeds because entities carry unequal numbers of observations: training rows ranged from 1,959,171 to 2,222,051 under LCLO and from 2,194,319 to 2,206,545 under LSO.

The ten train, validation and test partitions were fixed before model training and used for all representations and training arms. Two molecular representations do not cover the whole compound set. Their uncovered pairs were dropped from all three splits, and those configurations are recorded as evaluated on fewer pairs. A seed selects a manifest together with the weight initialisation and the minibatch order, so a paired difference across seeds averages over splits and initialisations jointly.

Predictor architecture, training and checkpoint selection

The predictor is a nonlinear dual-tower network. The compound tower maps a 2,048-dimensional ECFP4 vector to 512 dimensions in one block. The transcriptome tower maps the 15,961-dimensional gene-level expression profile to 1,024 and then to 512 dimensions. A block is a linear layer, a ReLU, batch normalisation and dropout with $p = 0.4$. The two tower outputs are concatenated and passed to a joint head of width 1,024 and then 512, built from the same blocks. A linear layer then emits one scalar. With these input dimensions the network has 19,501,057 trainable parameters. The component-supervised predictor leaves all of these modules unchanged and adds two 512-wide scalar effect heads and a learnable intercept, 528,387 further parameters.

The reference predictor was trained to minimise squared error on the total response. Optimisation used Adam, a learning rate of 10^{-4}, weight decay 10^{-5}, a batch size of 2,048 and 30 epochs. The learning

rate was multiplied by 0.2 after ten epochs without improvement in the validation correlation $\mathrm{corr}(y - b, \hat{y} - b)$. For this scheduler, b was the training grand mean plus drug and sample mean deviations: the deviation for the held-out entity type was estimated from validation labels, and that for the other type from training labels. The same b was subtracted from both vectors. This scheduler statistic differs from the orthogonally decomposed interaction correlation used for evaluation. Optimiser, schedule and epoch budget were identical for every arm, representation and regime.

The reported checkpoint is the epoch with the lowest total prediction error on the validation split. This rule was prespecified and applied to every arm, representation and regime. It is the rule that reads the whole response: the decomposition is orthogonal, so total prediction error is exactly $E_{\mathrm{add}} + E_{\mathrm{int}}$, and both components enter the criterion. A criterion computed on one component alone is blind to failure in the other.

Seeds were set for PyTorch, CUDA, NumPy and Python, with cuDNN in deterministic mode. The training data loader was additionally given an explicit random generator seeded from the run seed. Every arm at a given seed therefore received the same sequence of minibatch permutations.

Representation benchmark and decoder dependence

The representation benchmark used an anchor design rather than a full Cartesian product. Twelve molecular representations were each paired with gene-level expression, and twelve transcriptomic representations were each paired with ECFP4. The pair (ECFP4, gene-level expression) is the anchor of both arms, so the benchmark comprises 23 distinct configurations. Every configuration was fitted under both distribution-shift regimes and at three seeds, 3407, 3408 and 3409, using the predictor, training protocol and checkpoint rule described above. The anchor was fitted once and mirrored into both arms, so the benchmark table holds 144 rows for 138 fits.

The molecular representations were ECFP4, ChemBERTa2, Chemprop, GeminiMol, InfoAlign, KPGT, MolCLR, MolT5, Mole_BERT, Ouroboros, UniMol and UniMolV2. Their dimensions were 2,048, 384, 300, 2,048, 300, 2,304, 512, 768, 300, 2,048, 512 and 1,024, respectively. Ten of them cover all 54,180 compounds of the benchmark dataset. GeminiMol covers 49,170 compounds (90.8%) and UniMolV2 covers 54,052 (99.8%). Both were evaluated on their covered subset ; the prespecified data partitions were unchanged. The transcriptomic representations were gene-level expression itself, across 15,961 genes, and eleven pretrained transcriptomic embeddings. Their dimensions were SCimilarity 128, Geneformer 256, CellPLM 512, scGPT 512, Tahoe_x1 512, BulkFormer 640, Cell2Sentence 1,024, tGPT 1,024, UCE 1,280, STATE 2,058 and scFoundation 3,072.

Aligning each embedding file to the transcriptomic input matrix determines what the comparison measures. The transcriptomic input matrix carries the CCLE profile of every sample, so the embedding entering the comparison has to be the one computed from that same profile. Otherwise a comparison between gene-level expression and an embedding would confound the representation with the source profile. For each embedding, we selected the CCLE-derived profiles, matched them by sample identifier and restricted them to the 1,406 samples of the CCLE transcriptomic matrix. All eleven embeddings covered these samples without missing values.

Decoder dependence was assessed on a bounded subset of the benchmark under the leave-cell-line-out regime, holding the training protocol fixed and varying only the decoder class. Four molecular representations (ECFP4, MolCLR, KPGT, UniMol) were compared against gene-level expression, and four transcriptomic representations (gene-level expression, scGPT, CellPLM, BulkFormer) against ECFP4, giving seven distinct pairs. Three decoder classes were fitted for each pair at seed 3407, for 21 fits. The additive decoder predicts a linear function of the molecular representation plus a linear function of the transcriptomic representation, 18,010 parameters at the anchor. The bilinear decoder adds a rank-32 bilinear term to that model, 594,298 parameters at the anchor, with the rank fixed a priori and never varied. The third is the nonlinear dual-tower decoder of the main benchmark, 19,501,057 parameters at the anchor.

The additive decoder was included precisely because it cannot express interaction. It is a structural control rather than a competitor: the decomposition should assign it an interaction amplitude of zero, and this was verified rather than assumed. The standard deviation of its predicted interaction component ran from 3.0×10^{-9} to 3.7×10^{-9} across its seven configurations. The same quantity ran from

1.3×10^{-2} to 2.8×10^{-2} across the fourteen configurations of the other two decoders. Any representation ranking on r_{int} obtained from an additive decoder therefore carries no information, and the decomposition exposes that limitation directly.

Two protocol adjustments were forced by this comparison and were applied identically to every configuration in it. First, features were standardized per dimension using training entities only. The additive and bilinear decoders carry no normalization layer, and these representations differ in scale by two orders of magnitude. Without standardization a representation would be ranked by the scale of its embedding rather than its content. Absolute values from this comparison are consequently not comparable to those of the main benchmark. Second, the default weight decay of 10^{-5} leaves the additive decoder effectively unregularized, because that decoder fits 15,961 sample-side coefficients from 690 training cell lines. The penalty was therefore selected on validation total prediction error from the fixed grid $\{10^{-5}, 10^{-3}, 10^{-1}, 10\}$, identically for all fourteen normalization-free configurations and never on test data. The default value is nested in that grid. The dual tower kept its original hyperparameters; validation-based regularization selection was used only for the additive and bilinear decoders.

Decoder dependence was quantified as the Spearman rank correlation between the four-representation orderings produced by two decoder classes. It was computed separately for r_{total}, r_{add} and r_{int}, together with agreement of the top-ranked representation and of the top-two set. On r_{int} the additive decoder was excluded from these comparisons, so only the bilinear and dual-tower orderings were compared. Each ordering holds four items, so the coefficient takes only eleven possible values, and the explicit orderings were retained alongside it. The transcriptomic arm was additionally refitted at seeds 3408 and 3409 under the bilinear and dual-tower decoders. The bilinear penalty was re-selected on each seed's own validation split, and stability was summarized by the mean pairwise seed-to-seed Spearman correlation.

Descriptive score separation relative to seed variation

This descriptive diagnostic compares two spreads that the benchmark already reports. Fix a distribution-shift regime, a side of the benchmark and one of the three correlations. The across-representation spread is then the difference between the largest and the smallest seed-mean value over the representations on that side. The seed-to-seed spread is the median, over those same representations, of the standard deviation of that correlation across the three seeds. The score- separation ratio is the ratio of the first quantity to the second. It is a ratio of two spreads, not a new performance metric. It describes the distance between seed-mean scores relative to the distance a representation moves when its seed changes; it does not estimate paired score covariance, pairwise uncertainty or ranking stability.

Only representations covering the entire benchmark dataset entered the statistic. GeminiMol and UniMolV2 were therefore excluded, so the molecular side rests on ten representations and the transcriptomic side on twelve. The standard deviations are sample standard deviations over three seeds, so the denominator is itself estimated from three values. No categorical threshold or inference label is applied to the ratio, and no P value attaches to it.

The statistic was computed for both distribution-shift regimes, both sides and all three correlations. It was also recomputed on the transcriptomic side with UCE removed, leaving eleven representations. UCE is a single outlier there, and it is the one representation whose r_{int} differs sharply between the two regimes, 0.013 under leave-cell-line-out against 0.276 under leave-small-molecule-out. Without UCE, the leave-cell-line-out transcriptomic ratios fall to 1.6 on r_{total}, 1.3 on r_{add} and 2.1 on r_{int}, and the leave-small-molecule-out ratios to 1.7, 1.5 and 2.7. UCE was retained in the reported statistic, and the recomputation is reported as a sensitivity check.

Component-wise supervision

Component-wise supervision is a targeted model intervention on the predictor described above. Its two encoders and its joint head are unchanged, and two scalar effect heads plus a learnable intercept are added. The additive head returns $\hat{b} = \mu + f_d(e_d) + f_s(e_s)$, where μ is the intercept. Here f_d and f_s are the two effect heads, each with one 512-unit hidden layer and a scalar output. They read the 512-dimensional drug embedding e_d and sample embedding e_s respectively. The interaction head is the baseline joint head, which reads both embeddings concatenated and returns a scalar $\hat{z}$. The prediction is $\hat{y} = \hat{b} + \hat{z}$. Because $\hat{b}$ is a drug function plus a sample function plus a constant, it lies in the additive

subspace by construction. Its measured interaction content, $\|M\hat{b}\|^2/\|\hat{b}\|^2$, stayed below 10^{-15} in every reported run. The encoders carry batch normalization, so the factorization is exact only at evaluation, where those layers are fixed. Applying the additive projection to the interaction head output gives the additive-subspace leakage $\|H\hat{z}\|^2/\|\hat{z}\|^2$, which we record for every run.

Each head was given its own target. The additive component Hy and the interaction component My were formed on the training support and used only as training targets. Evaluation used a projection rebuilt on the evaluation panel, never the training projection. Training minimized

$$L \;=\; \mathrm{MSE}(\hat{b}+\hat{z},\, y) \;+\; \frac{\mathrm{Var}(y)}{\mathrm{Var}(Hy)}\,\mathrm{MSE}(\hat{b},\, Hy) \;+\; \frac{\mathrm{Var}(y)}{\mathrm{Var}(My)}\,\mathrm{MSE}(\hat{z},\, My),$$

with all three variances computed once on the training support and then held fixed. Both targets are functions of training labels alone. The two multipliers put every term in the units of the total prediction error. The expression is $\mathrm{Var}(y)$ times the fully variance-normalized loss, so it has the same minimizer. It also leaves the first term numerically identical to the objective used without component supervision. Normalization matters because the additive component carries about three times the variance of the interaction component here. Across the ten training runs the two multipliers were 1.31 to 1.33 and 4.07 to 4.26.

The two effect heads and the intercept add 528,387 trainable parameters to the baseline's 19,501,057, giving 20,029,444. M3 is the matched supervision control: it has exactly M4's two-head architecture and 20,029,444 trainable parameters, but is trained only on the retained total-response term. M3 and M4 were trained on the current benchmark under both distribution shifts and all five seeds. Their same transcriptomic inputs, data partitions, test support and checkpoint-selection rule were identical (Supplementary Table 16).

Every other setting was inherited: representation, panel, prespecified split manifests, seeds, optimizer, scheduler, learning rate, weight decay, batch size, dropout and epoch budget. The checkpoint rule was the same for every arm, namely the epoch with the lowest validation total prediction error. Minibatch order came from a generator seeded by the run seed, so paired arms drew the same permutation. The objective was applied unchanged in both training regimes, leave-cell-line-out and leave-small-molecule-out. The zero-shot organoid evaluation reused the leave-cell-line-out checkpoints and involved no further training. No loss weight, layer width, head or hyperparameter was tuned per regime, and the two multipliers follow from the training labels rather than from a search. Nothing was selected on a test panel. Per-seed M0 and M4 results, including component errors and selected epochs, are given in Supplementary Table 20.

Zero-shot organoid evaluation and input-compatibility criteria

The organoid evaluation was zero-shot. The predictors trained on the cell-line panel were applied to the organoid cohorts with no organoid training and no organoid split. No organoid-specific loss term and no organoid-specific hyperparameter were used, and the whole organoid panel is a test set. The primary M0/M4 evaluation used the leave-cell-line-out checkpoints already selected by the prespecified rule, one M0 and one M4 checkpoint for each of seeds 3407 to 3411, hence ten checkpoints. The three-arm M0/M3/M4 comparison additionally evaluated the matched M3 checkpoint for each seed, hence fifteen checkpoints total. None was fine-tuned, reselected or recalibrated on organoid data. The decomposition was rebuilt on each reported panel, where r_{total}, r_{add} and r_{int} were computed.

Transfer is interpretable only where the transcriptomic input is the same kind of quantity as the training input. A prespecified input-compatibility check of three conditions was therefore applied to every cohort before any model was run. The reference is the CCLE-derived transcriptomic input matrix, whose mean per-sample level is 2.702, whose fraction of exact zeros is 0.1533 and whose minimum is 0. First, a cohort's mean per-sample level must lie within 1.0 of the reference level. Second, its fraction of exact zeros must lie between half and twice the reference fraction, that is within 0.0767 to 0.3067. Third, its minimum must not fall more than 0.5 below the reference minimum. UMPDO1 and UMPDO2 pass all three, at level gaps of 0.876 and 0.694. LICOB and HKUPDO fail all three, at level gaps of 2.303 and 2.694, zero fractions of 0.000 and 0.951, and minima of -1.80 and -9.97. These two are input-incompatible cohorts, reported separately and excluded from the primary estimate.

UMPDO3 passes the second and third conditions, with a zero fraction of 0.094 and a minimum of 0. It misses the first by 0.005, at a level gap of 1.005. It was nonetheless retained in the primary cohort, which had been fixed before this evaluation as UMPDO1, UMPDO2 and UMPDO3 pooled, 4,886 pairs on 100 organoids and 78 compounds. We moved neither the threshold nor the cohort definition on discovering this near-miss. The three continuous quantities are shown per cohort, so a reader can see where the line falls, and every cohort estimate is reported beside the pooled one (Supplementary Table 21). The matched three-arm results are reported in Supplementary Table 22.

Organoid expression was mapped onto the 15,961 training genes. A gene not measured in a cohort was set to zero, the convention the training pipeline applies to missing expression. Missing entries among the measured genes were set to zero as well. Coverage of the training genes was 15,413, 15,218 and 14,651 for UMPDO1, UMPDO2 and UMPDO3, against 13,224 for LICOB and 8,866 for HKUPDO.

Statistical analysis

The unit of replication for every paired comparison in this study is the training seed, and n is five. Five seeds, 3407 to 3411, were run for M0, M3 and M4 in the comparisons. Each seed carries its own trained baseline, matched supervision control and component-supervised predictor. Under the two held-out-axis regimes each seed also carries its own prespecified split manifests, so the test panel changes with the seed. Comparisons between arms were paired within seed, so paired arms are always scored on the same panel. Where the panel changes with the seed, the seed-to-seed spread of the paired difference carries panel variation as well as training variation. The zero-shot organoid evaluation instead uses one fixed panel of 4,886 pairs at every seed.

For each paired quantity we report the mean difference $\bar{d}$ over the five seeds. The reported statistic is $t = \bar{d}/(s_d/\sqrt{5})$, where s_d is the sample standard deviation of the five paired differences. It is a one-sample t on the paired differences, with four degrees of freedom. Where a two-sided 95% interval is given, it is $\bar{d} \pm t_{0.975,4}\, s_d/\sqrt{5}$. That statistic assumes the five paired differences are approximately normal, and five values cannot test the assumption. Every t is therefore reported beside the number of seeds carrying a consistent sign, which assumes no distribution. No P values are reported and no multiplicity adjustment was applied; each comparison is read on the size of its effect and on those two quantities.

No statistical procedure was used to determine the number of drug–sample pairs, samples, compounds or response resources included. Every pair meeting the prespecified eligibility rule was retained, and the number of seeds was fixed at five before the comparisons were run. Blinding does not apply: each comparison is between two trained predictors scored by a fixed procedure on prespecified splits, and no investigator judgement enters between a prediction and its score. No scored observation was discarded. Two prespecified rules restrict what a given analysis is computed on rather than removing a scored value, the matched-support filter for the reproducibility reference and the input-compatibility check for the organoid cohorts, and both were fixed before the analyses were run. The primary organoid set was fixed before evaluation as UMPDO1, UMPDO2 and UMPDO3; the compatibility check was applied to every available cohort and is reported as a validity diagnostic, not as the rule that selected the primary set. The model-arm contrasts, perturbations and independent-assay model-selection test were specified before their results were inspected; every other comparison reported here is descriptive.

The representation benchmark was run at three seeds per configuration, 3407 to 3409. It is reported as a seed mean with the seed-to-seed standard deviation, and no paired test is applied to it. The descriptive score-separation ratio is the range of those seed means across representations, divided by the median of their seed-to-seed standard deviations. It is formed separately for each input axis, each distribution-shift regime and each of the three correlations, over the representations evaluated on the full panel; it does not establish ranking stability or pairwise separability.

Intervals on cross-assay reproducibility are cluster bootstraps of 600 replicates, and the resampling units are whole compounds and whole samples of the matched-pair support. One scheme resamples compounds, one resamples samples, and a two-way scheme resamples both. Each scheme uses its own fixed generator seed, 11, 12 and 13. The additive projection and the reproducibility estimate were recomputed inside every replicate. Intervals are the 2.5th and 97.5th percentiles of the replicate distribution. Replicates retaining fewer than 300 rows were discarded, and all 600 replicates were usable in each scheme. Repeated cluster draws are retained with multiplicity and relabelled; under two-way resam-

pling, a pair's multiplicity is the product of its compound and sample draw counts.

The 3,141,680 drug–sample pairs of the benchmark dataset are not the unit of replication for any paired test reported here. Pairs share compounds and samples, so they are not independent. A pair count states the size of an evaluation dataset. The n behind every t value in this paper is five seeds.

Software

The analyses were run in Python 3.11.10 with NumPy 1.26.4, pandas 1.5.3, SciPy 1.15.1, scikit-learn 1.5.2 and PyArrow 18.0.0. Predictors were trained with PyTorch 2.4.1 built against CUDA 11.8.

Acknowledgements

This work was supported in part by the Ministry of Education, Singapore (MOE-T1251RES2309, MOE-T2EP20125-0014, and MOE-T2EP20225-0028), the Agency for Science, Technology and Research (A*STAR), Singapore (IAF-PP H25J6a0034), and the National Research Foundation, Singapore (NRF-CRP33-2025-0003 and AI4SCT_2025-0014). B.Z. was supported by the University of Macau (MYRG-GRG2024-00205-FST) and the Science and Technology Development Fund, Macau SAR (0028/2023/RIA1).

Data availability

All drug-response and transcriptomic data analysed in this study are derived from public resources. Response measurements were taken from the DROMA collection of harmonized preclinical drug-response and omics data[40], which aggregates the eleven resources used here: NCI60[41], PRISM[5], CTRP1[42], CTRP2[43], GDSC1[3], GDSC2[44], CCLE[4], GRAY[45], gCSI[46], FIMM[47] and UHNBreast[48]. The processed DROMA collection is archived at https://doi.org/10.5281/zenodo.18503188; its input-data provenance record is https://doi.org/10.5281/zenodo.17498421. The transcriptomic input matrix is derived from the Cancer Cell Line Encyclopedia[4]. Supplementary Data 1 records the DROMA project aliases used in this study and, where verifiable from the available metadata, the corresponding upstream publications and accessions. For several organoid cohorts, the DROMA metadata does not expose a recoverable upstream publication or accession; the cited DROMA processed record is therefore the direct reproducible source used here.

Code availability

The code for data construction and preprocessing, data splitting, M0/M3/M4 training and evaluation, and the analyses reported here is available at https://github.com/Boom5426/DrugDis.

References

[1] Bunne, C. *et al.* How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell* **187**, 7045–7063 (2024).

[2] Roohani, Y. H. *et al.* Virtual Cell Challenge: Toward a Turing test for the virtual cell. *Cell* **188**, 3370–3374 (2025).

[3] Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).

[4] Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

[5] Corsello, S. M. *et al.* Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nat. Cancer* **1**, 235–248 (2020).

[6] Carli, F. *et al.* Learning and actioning general principles of cancer cell drug sensitivity. *Nat. Commun.* **16**, 1654 (2025).

[7] Bernett, J. *et al.* Critical evaluation of drug response prediction models with DrEval. *Nat. Commun.* **17**, 4238 (2026).

[8] Xia, X., Zhu, C., Zhong, F. & Liu, L. TransCDR: a deep learning model for enhancing the generalizability of drug activity prediction through transfer learning and multimodal data fusion. *BMC Biol.* **22**, 227 (2024).

[9] Chawla, S. *et al.* Gene expression based inference of cancer drug sensitivity. *Nat. Commun.* **13**, 5680 (2022).

[10] Codicè, F. *et al.* The specification game: rethinking the evaluation of drug response prediction for precision oncology. *J. Cheminformatics* **17**, 33 (2025).

[11] Viñas Torné, R. *et al.* Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nat. Biotechnol.* **44**, 1050–1059 (2026).

[12] Sharifi-Noghabi, H. *et al.* Drug sensitivity prediction from cell line-based pharmacogenomics data: guidelines for developing machine learning models. *Brief. Bioinform.* **22**, bbab294 (2021).

[13] Xia, F. *et al.* A cross-study analysis of drug response prediction in cancer cell lines. *Brief. Bioinform.* **23**, bbab356 (2022).

[14] Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

[15] Heid, E. *et al.* Chemprop: A machine learning package for chemical property prediction. *J. Chem. Inf. Model.* **64**, 9–17 (2024).

[16] Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **4**, 279–287 (2022).

[17] Liu, G. *et al.* Learning molecular representation in a cell. In *The Thirteenth International Conference on Learning Representations* (2025).

[18] Ahmad, W., Simon, E., Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa-2: Towards chemical foundation models. Preprint at https://doi.org/10.48550/arXiv.2209.01712 (2022).

[19] Xia, J. *et al.* Mole-BERT: Rethinking pre-training graph neural networks for molecules. In *The Eleventh International Conference on Learning Representations* (2023).

[20] Edwards, C. *et al.* Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 375–413 (Association for Computational Linguistics, 2022).

[21] Li, H. *et al.* A knowledge-guided pre-training framework for improving molecular representation learning. *Nat. Commun.* **14**, 7568 (2023).

[22] Zhou, G. *et al.* Uni-Mol: A universal 3D molecular representation learning framework. In *The Eleventh International Conference on Learning Representations* (2023).

[23] Ji, X. *et al.* Exploring molecular pretraining model at scale. In *Advances in Neural Information Processing Systems*, vol. 37, 46956–46978 (2024).

[24] Wang, L. *et al.* Conformational space profiling enhances generic molecular representation for AI-powered ligand-based drug discovery. *Adv. Sci.* **11**, 2403998 (2024).

[25] Wang, L. *et al.* Learned conformational space and pharmacophore into molecular foundational model. *Adv. Sci.* **13**, e13556 (2026).

[26] Kang, B., Fan, R., Yi, M., Cui, C. & Cui, Q. BulkFormer: a large-scale foundation model for bulk transcriptomes. *Cell Syst.* **17**, 101657 (2026).

[27] Gandhi, S. *et al.* Tahoe-x1: Scaling perturbation-trained single-cell foundation models to 3 billion parameters. Preprint at https://doi.org/10.1101/2025.10.23.683759 (2025).

[28] Heimberg, G. *et al.* A cell atlas foundation model for scalable search of similar human cells. *Nature* **638**, 1085–1094 (2025).

[29] Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

[30] Shen, H. *et al.* Generative pretraining from large-scale transcriptomes for single-cell deciphering. *iScience* **26**, 106536 (2023).

[31] Levine, D. *et al.* Cell2Sentence: Teaching large language models the language of biology. In *Proceedings of the 41st International Conference on Machine Learning*, vol. 235, 27299–27325 (PMLR, 2024).

[32] Hao, M. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* **21**, 1481–1491 (2024).

[33] Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).

[34] Adduri, A. K. *et al.* Predicting cellular responses to perturbation across diverse contexts with State. *Cell* **189**, 5914–5931.e20 (2026).

[35] Rosen, Y. *et al.* Universal cell embedding provides a foundation model for cell biology. *Nature* **656**, 183–191 (2026).

[36] Wen, H. *et al.* CellPLM: Pre-training of cell language model beyond single cells. In *The Twelfth International Conference on Learning Representations* (2024).

[37] Guimarães, P. & Portugal, P. A simple feasible procedure to fit models with high-dimensional fixed effects. *Stata J.* **10**, 628–649 (2010).

[38] Cameron, A. C., Gelbach, J. B. & Miller, D. L. Bootstrap-based improvements for inference with clustered errors. *Rev. Econ. Stat.* **90**, 414–427 (2008).

[39] Owen, A. B. The pigeonhole bootstrap. *Ann. Appl. Stat.* **1**, 386–411 (2007).

[40] Peng, Y. DROMA datasets (DROMA_DB). Zenodo (2026). URL https://doi.org/10.5281/zenodo.18503188.

[41] Holbeck, S. L., Collins, J. M. & Doroshow, J. H. Analysis of Food and Drug Administration–approved anticancer agents in the NCI60 panel of human tumor cell lines. *Mol. Cancer Ther.* **9**, 1451–1460 (2010).

[42] Basu, A. *et al.* An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* **154**, 1151–1161 (2013).

[43] Seashore-Ludlow, B. *et al.* Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* **5**, 1210–1223 (2015).

[44] Iorio, F. *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).

[45] Daemen, A. *et al.* Modeling precision treatment of breast cancer. *Genome Biol.* **14**, R110 (2013).

[46] Haverty, P. M. *et al.* Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* **533**, 333–337 (2016).

[47] Pemovska, T. *et al.* Individualized systems medicine strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. *Cancer Discov.* **3**, 1416–1429 (2013).

[48] Marcotte, R. *et al.* Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance. *Cell* **164**, 293–309 (2016).