Supplementary Information

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

This Supplementary Information contains three Supplementary Notes on the response decomposition, zero-amplitude predictions and cross-assay reproducibility; 22 Supplementary Tables with detailed benchmark results; and one Supplementary Data item describing the data sources.

Contents

Supplementary Notes **2**

Note 1: the decomposition on an unbalanced observed support 2

Note 2: the degenerate case of the interaction-component correlation 2

Note 3: interpretation of the cross-assay reproducibility reference 3

Supplementary Tables **5**

Supplementary Data **14**

Supplementary Notes

Note 1: the decomposition on an unbalanced observed support

The projection H is fixed by the observed support alone, and the support of a pharmacogenomic dataset is neither complete nor balanced. Three consequences follow, and each one changes how a component value may be read.

The drug and sample marginals do not have separate variance shares. On a complete balanced grid the two sets of indicator columns are orthogonal, and the variance carried by the additive component splits uniquely into a drug part and a sample part. On an unbalanced support they are correlated, so no unique split exists. We therefore report three quantities rather than two. On the benchmark dataset the drug marginal alone accounts for 72.65% of total response variance and the sample marginal alone for 4.17%, but 0.97% of the variance is carried by both. The parts unique to each are 71.68% and 3.20%, and the three sum to the 75.85% the additive component carries. The shared part is small here and it is not constant: across the ten test sets it ranges from 0.46% to 1.40% (Supplementary Table 2). A statement of the form "the drug marginal explains x% of the variance" is well defined only once the sharing is quoted with it, which is why the main text quotes the unique shares.

The interaction component is indexed by the support that defines it. Two evaluation sets with different supports define different projections, so their interaction components are different quantities and not two measurements of one. The dimension of the additive subspace is $\mathrm{rank}(H) = n_{\mathrm{drug}} + n_{\mathrm{sample}} - c$ with c the number of connected components of the bipartite support graph, and $c = 1$ for the benchmark dataset and for every test set. The fraction of degrees of freedom the projection removes varies by an order of magnitude between evaluation sets: 1.76% on the benchmark dataset, 9.52% to 11.45% on the held-out cell-line test sets, and 1.92% to 1.98% on the held-out compound test sets. This is why the error density in Methods is normalized by the rank of each subspace rather than by the number of observations, and why component values are compared within a distribution-shift regime rather than across regimes.

A level observed once is a limiting case of the same fact. Its own response is its marginal, so the projection assigns all of it to the additive component and none to the interaction component. Sparsely observed compounds and samples therefore contribute to the additive component by construction, whatever their pharmacology.

The marginals are properties of the screening design as well as of the biology. The observed support covers 3,141,680 of the 54,180 $\times$ 986 possible drug–sample pairs, that is 5.88% of the grid, and which pairs were screened was decided by the resources that ran them rather than at random. The drug marginal is the mean response of a compound over the samples on which that compound happens to have been screened, and the sample marginal is the mean response of a sample over the compounds in the libraries applied to it. Neither estimates an intrinsic potency or an intrinsic sensitivity. The additive component that the decomposition removes is accordingly a property of the observed design as much as of pharmacology, and the interaction component is the deviation from the best additive explanation of that design. Every component value in this work is stated on a named support, the benchmark dataset, a test set or a single response resource (Supplementary Table 3), and carries that qualification.

Note 2: the degenerate case of the interaction-component correlation

The interaction-component correlation is $\mathrm{corr}(My, M\hat{y})$. When the prediction lies in the additive subspace, $M\hat{y} = 0$, both the numerator and the denominator of the Pearson formula vanish, and the quantity is undefined. This is not an edge case: it is the exact situation of every predictor that carries no drug–sample interaction information at all.

Which predictors fall in it. A constant predictor does, and so does any predictor assembled from drug-level and sample-level terms, including one built from training-set marginals. A structurally additive decoder does as well, for the same reason. Across the 28 runs of the additive ridge decoder in

the decoder comparison, the standard deviation of the emitted interaction lies between 1.6×10^{-9} and 1.2×10^{-8}, against an observed interaction standard deviation of order 7×10^{-2}, and the emitted amplitude A_{int} is zero to six decimal places in every one (Supplementary Table 11). What a Pearson formula returns for those runs spans -0.019 to $+0.017$, and it is the correlation of floating-point noise. Ranking additive decoders on the interaction component therefore orders noise, which is why the main text reads that comparison from the emitted amplitude and not from the correlation (Fig. 3g).

How the two cases are separated. We record the emitted amplitude $A_{\text{int}} = \text{sd}(M\hat{y})/\text{sd}(My)$ beside every correlation. For generic prediction records and the null calibration, a record is flagged as emitting no interaction when $A_{\text{int}} \leq 10^{-8}$. A decoder that is structurally additive by construction is separately declared non-emitting, regardless of its finite-precision residual; this is the status of the additive ridge in Fig. 3g. In the null-calibration records the numeric threshold falls in an empty region: predictors that emit nothing reach 9.4×10^{-18} to 1.8×10^{-9}, whereas emitting records range from 0.2775 to 1.9275. For a flagged record the interaction-component correlation is reported as undefined. Supplementary Table 5 reports it that way, and reports the total response and additive-component correlations of the constant predictor as undefined for the same reason: a constant vector carries no variance in any subspace, so none of the three correlations exists for it.

The informative null is the other one. A predictor that emits interaction but carries none of its structure gives a measured zero on a defined quantity. Shuffled predictions emit 1.3480 to 1.9275 times the amplitude the observed interaction contains and score -0.001 under held-out cell lines, $+0.000$ under held-out compounds and $+0.000$ in organoids. That is what calibrates the scale, and it is a different statement from the undefined case.

Note 3: interpretation of the cross-assay reproducibility reference

Cross-assay reproducibility q_c is the correlation between GDSC1 and GDSC2 measurements of the same drug–sample pair, computed component by component with one projection built on the matched support and applied identically to both arms. On the 32,724 raw matched keys that pass benchmark key eligibility it is 0.650 for total response, 0.754 for the additive component and 0.498 for the interaction component. The empirical reproducibility reference is $\sqrt{q_c}$, that is 0.807, 0.869 and 0.706, and recovery relative to it is $\eta_c = r_c/\sqrt{q_c}$ (Supplementary Table 6).

The square root has a narrow, model-based interpretation. If two measurements obey $y_A = z + \epsilon_A$ and $y_B = z + \epsilon_B$, where z is a common latent quantity, the errors have equal variance, are mutually independent and are independent of z, then $q_c = \text{Var}(z)/\text{Var}(y_A)$ and a perfect predictor of z has correlation $\sqrt{q_c}$ with one measured arm. These are classical measurement-model assumptions, not properties established for GDSC1 and GDSC2. In particular, their assay-generation difference can violate the exchangeability and independent-error assumptions. We therefore use $\sqrt{q_c}$ only as an empirical comparison reference, never as an attainable performance limit or a biological truth.

Analysis supports. Different analyses use the following subsets of matched observations. At every stage the source field is DROMA `Sensitivity`, passed without rescaling. The source metadata document its numerical range but do not recover a uniform upstream assay metric or biological direction for each resource.

Stage	Key and transformation	Count	Use and interpretation
Benchmark response panel	Canonical (`SMILES`, `Sample_ID`); mean of retained `Sensitivity` rows	3,141,680	Training and evaluation target for the benchmark.
Raw GDSC match	Exactly one GDSC1 and one GDSC2 record for the same canonical key, giving y_A, y_B	43,622; 89 drugs, 799 samples	Starting set for cross-assay reproducibility and independent model-selection analyses.
Benchmark eligibility filter	Raw GDSC match whose canonical key is admitted by the benchmark eligibility rules	32,724; 89 drugs, 600 samples	Full-support q_c reference. Key membership alone does not establish that a prediction target equals y_A.
Exact recovery support	Prediction row on a raw GDSC match with stored target equal to y_A within 10^{-6}	LCLO 2,876–3,051; LSO 1,970–3,873 per seed	Support for η_c only. This target-equality filter excludes rows carrying another resource's measurement.
Independent model-selection support	Per-split, per-seed inner join of matched raw GDSC keys to both M0 and M4 prediction rows	LCLO 4,927–5,411 (mean 5,138); LSO 2,723–7,164 (mean 5,027)	GDSC1 selection metrics and GDSC2 independent endpoint. It does not apply the exact-recovery target-equality filter.

Four properties qualify what that reference supports.

It is not technical replication. The two arms are GDSC1 and GDSC2, two assay generations run by one consortium. The estimate therefore mixes measurement noise with an assay change, and it is the reproducibility of a measurement across that change, and not of one protocol run twice. Pairings across different resources were also estimated and none was adopted: each spans 389 to 3,174 pairs on fewer than 20 compounds, and each adds protocol divergence on top of the same assay difference.

It is a summary, not an attainable limit. A single scalar per component does not describe a quantity whose reproducibility varies with its magnitude. On the matched pairs, q_{int} rises from 0.08 in the smallest quartile of interaction magnitude to 0.62 in the largest. Recovery ratios above one are therefore possible, and one occurs: under held-out cell lines the additive η averages 1.00 across five seeds with a range of 0.97 to 1.03 (Supplementary Table 14). This illustrates why the reference should not be interpreted as an attainable limit.

Its support is not the evaluation support. The reference exists only where measurements repeat. Restricted to the pairs a regime actually holds out, the support falls to 2,876 to 3,051 pairs under held-out cell lines, over 79 to 86 compounds and 90 to 101 held-out cell lines. Under held-out compounds it holds 1,970 to 3,873 pairs but only 7 to 19 held-out compounds, and with so few compound clusters the ratio is unstable: it ranges from 0.03 to 0.54 across the five seeds for the interaction component and from 0.32 to 0.65 for the additive component, and 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 η_c under held-out cell lines and report its absence under held-out compounds. Supplementary Table 14 lists the per-seed values in both regimes so that the instability is visible rather than asserted. No repeated cross-assay measurements exist for the organoid cohorts, so no reference is defined there.

Its uncertainty is dominated by the compound axis. Observations sharing a compound or a sample are not independent, so intervals come from cluster bootstraps rather than from rows. For total response the sample bootstrap returns 0.640 to 0.660 across 600 clusters, while the compound bootstrap returns 0.529 to 0.755 across 89. The two-way scheme, which resamples both axes with draw multiplicity retained, gives 0.528 to 0.750. Intervals on η_c resample both axes once per replicate and recompute the model correlation and the reference together, so the numerator and the denominator carry their uncertainty jointly.

For the exact recovery calculation, the scored prediction target is required to equal the single arm-A measurement, not an average of the two GDSC arms. Had that target been an average of n measurements, the reference would need rescaling by the Spearman-Brown relation $q_n = nq/(1 + (n-1)q)$ before any recovery ratio was formed (Spearman, *Br. J. Psychol.* **3**, 271–295, 1910; Brown, *Br. J. Psychol.* **3**, 296–322, 1910). This statement is specific to the exact recovery support; it does not relabel the ordinary canonical-mean benchmark target described in the map above.

Supplementary Tables

Supplementary Table 1: Benchmark dataset composition and provenance. Composition of the cell-line benchmark. The transcriptomic input matrix is CCLE-derived and restricted to 15,961 genes. A sample enters the benchmark dataset only if a CCLE baseline profile exists for it, and the Tavor project is excluded on provenance grounds. Response rows are counted per resource before pair-level response aggregation across eleven response resources, and after the overlap rule of the processed response table (Methods). The single-resource scale rule in this table applies to the CCLE-derived transcriptomic input, not to the response data. Rows sharing a canonical drug–sample pair were averaged. The CCLE transcriptomic matrix contains one profile per sample.

Item	Value
Transcriptomic input	CCLE, gene whitelist 15,961
Profiled samples in the input matrix	1,406
Drug–sample pairs in the benchmark dataset	3,141,680
Cell-line samples	986
Compounds	54,180
Excluded project	Tavor
Transcriptome-profile duplicate rule	already unique
Transcriptomic-input scale rule	single resource
Checkpoint rule	Minimum validation mean squared error
NCI60 (response rows)	2,175,581
Prism (response rows)	504,493
CTRP2 (response rows)	269,725
GDSC1 (response rows)	125,914
GDSC2 (response rows)	44,797
CTRP1 (response rows)	39,784
CCLE (response rows)	10,730
GRAY (response rows)	1,888
gCSI (response rows)	1,646
FIMM (response rows)	890
UHNBreast (response rows)	54
All response resources	**3,175,502**

Supplementary Table 2: Orthogonal response decomposition on the benchmark dataset and on every test set. rank(H) is the dimension of the additive subspace and d_H/N the fraction of degrees of freedom it occupies, in per cent. Additive and Interaction are the percentages of total response variance carried by the two components. Drug and Sample are the percentages uniquely attributable to each marginal and Shared the percentage carried by both; Supplementary Note 1 explains why that third quantity is reported separately. |corr| is the absolute correlation between the two components, in units of 10^{-11}, and is the numerical check that the decomposition is orthogonal wherever it is applied.

Evaluation set	Pairs	Compounds	Samples	rank(H)	d_H/N	Additive	Interaction	Drug	Sample	Shared	|corr|
Benchmark dataset	3,141,680	54,180	986	55,165	1.76	75.85	24.15	71.68	3.20	0.97	2.72
LCLO, seed 3407	557,484	54,176	148	54,323	9.74	78.36	21.64	74.62	2.34	1.40	1.39
LCLO, seed 3408	474,279	54,176	148	54,323	11.45	76.22	23.78	71.31	4.46	0.46	1.68
LCLO, seed 3409	486,991	54,172	148	54,319	11.15	77.16	22.84	73.81	2.46	0.90	3.20
LCLO, seed 3410	570,863	54,175	148	54,322	9.52	76.76	23.24	72.24	3.86	0.67	2.58
LCLO, seed 3411	530,351	54,175	148	54,322	10.24	76.93	23.07	73.40	2.78	0.75	2.68
LSO, seed 3407	468,871	8,127	986	9,112	1.94	74.78	25.22	70.73	3.19	0.86	2.09
LSO, seed 3408	466,230	8,127	984	9,110	1.95	75.83	24.17	71.82	2.98	1.03	2.04
LSO, seed 3409	459,058	8,127	984	9,110	1.98	75.93	24.07	71.56	3.37	1.00	2.62
LSO, seed 3410	475,001	8,127	986	9,112	1.92	75.82	24.18	71.64	3.31	0.87	1.01
LSO, seed 3411	474,086	8,127	984	9,110	1.92	76.17	23.83	72.45	3.19	0.53	0.57

Supplementary Table 3: Orthogonal response decomposition within each response resource. Each resource is decomposed on its own eligible screen: its DROMA measurements that meet the inclusion criteria of the benchmark dataset, with rows sharing a canonical drug–sample pair averaged within the resource and no value from another resource. The benchmark dataset row repeats Supplementary Table 2. Without NCI60 is the benchmark dataset rebuilt from the other ten resources by the rules that built it. Resources are ordered by the number of pairs in their screen. Columns as in Supplementary Table 2, in per cent. [b]Pairs and additive share on the rows that the benchmark dataset attributes to the resource, which are fewer where the same compound name and sample were also measured by a resource earlier in the processing order (Methods). The absolute correlation between the two components is below 10^{-9} in every row.

Response support	Pairs	Compounds	Samples	d_H/N	Additive	Interaction	Drug	Sample	Shared	Pairs[b]	Additive[b]
Benchmark dataset	3,141,680	54,180	986	1.76	75.85	24.15	71.68	3.20	0.97	—	—
Without NCI60	993,482	1,926	984	0.29	66.54	33.46	60.59	4.13	1.83	—	—
NCI60	2,170,974	52,908	68	2.44	82.54	17.46	79.48	2.69	0.37	2,157,687	82.59
Prism	612,596	1,375	473	0.30	81.27	18.73	79.37	1.90	-0.01	503,644	81.61
CTRP2	292,817	462	822	0.44	69.74	30.26	60.06	8.55	1.13	269,725	69.94
GDSC1	167,886	293	701	0.59	59.04	40.96	53.17	5.93	-0.06	125,914	58.17
GDSC2	88,507	174	605	0.88	59.03	40.97	52.37	6.66	0.00	44,797	56.01
CTRP1	40,085	307	234	1.35	73.17	26.83	57.82	14.04	1.31	39,784	73.32
gCSI	11,495	41	523	4.90	71.51	28.49	65.13	7.31	-0.93	1,646	83.50
CCLE	10,730	24	472	4.61	79.27	20.73	73.99	5.02	0.27	10,730	79.27
GRAY	3,087	95	50	4.66	72.75	27.25	65.04	7.84	-0.13	1,888	74.76
FIMM	2,345	52	46	4.14	56.59	43.41	46.73	9.80	0.06	890	58.13
UHNBreast	166	6	39	26.51	48.04	51.96	24.45	21.55	2.04	54	87.98

Supplementary Table 4: Behaviour of existing evaluation views and DrugDis under prespecified perturbations. Entries are mean changes from the original prediction over M0 and M4 at five seeds. Each metric is computed on identical predictions and pair support within a run. Drug-only and cell-only offsets have a standard deviation equal to half that of the observed response. The two interaction perturbations retain the projected additive prediction and multiply only the projected interaction prediction.

Regime	Metric	Drug-only offset	Cell-only offset	Interaction x 0.5	Interaction x 2
Unseen cell lines	Global	-0.1185	-0.1216	-0.0015	-0.0209
	Fixed drug	0.0000	-0.1203	0.0111	-0.0471
	Fixed cell	-0.1056	0.0000	-0.0014	-0.0324
	DrEval normalized	-0.1544	-0.1663	-0.0485	0.0230
	DrugDis interaction direction	-0.0000	0.0000	0.0000	0.0000
	DrugDis interaction amplitude	0.0000	-0.0000	-0.1807	0.3614
	DrugDis interaction R^2	-0.0000	0.0000	-0.0138	-0.1702
Unseen compounds	Global	-0.1230	-0.1207	-0.0030	-0.0216
	Fixed drug	0.0000	-0.2051	0.0021	-0.0950
	Fixed cell	-0.1111	-0.0000	0.0008	-0.0291
	DrEval normalized	-0.1254	-0.1232	-0.0035	-0.0213
	DrugDis interaction direction	0.0000	0.0000	0.0000	0.0000
	DrugDis interaction amplitude	0.0000	-0.0000	-0.1833	0.3667
	DrugDis interaction R^2	0.0000	0.0000	-0.0088	-0.1842

Supplementary Table 5: Null-predictor calibration. Mean and standard deviation across five seeds for four predictors under each evaluation regime. Total, Additive and Interaction are the three component-wise correlations. Entries read *undefined* where the predicted vector carries no variance in the corresponding subspace: for the constant predictor in all three, and for the constant and marginal predictors in the interaction subspace, which is the case Supplementary Note 2 treats. $\mathrm{sd}(M\hat{y})$ is the emitted interaction amplitude in absolute units and $\mathrm{sd}(My)$ the amplitude the observed interaction carries. The organoid rows are the prespecified primary pooled set (UMPDO1–3), not a set selected by the compatibility diagnostic.

Regime	Predictor	n	Total	Additive	Interaction	$\mathrm{sd}(M\hat{y})$	$\mathrm{sd}(My)$
Unseen cell lines	Constant	5	undefined	undefined	undefined	1.88×10^{-17}	0.0711
	Training marginals	5	0.849 (0.010)	0.967 (0.007)	undefined	1.29×10^{-18}	0.0711
	Shuffled predictions	5	-0.001 (0.001)	-0.001 (0.005)	-0.001 (0.002)	0.1203	0.0711
	Trained predictor (M0)	5	0.858 (0.009)	0.963 (0.007)	0.315 (0.040)	0.0247	0.0711
Unseen compounds	Constant	5	undefined	undefined	undefined	1.25×10^{-16}	0.0732
	Training marginals	5	0.196 (0.006)	0.226 (0.008)	undefined	1.22×10^{-10}	0.0732
	Shuffled predictions	5	0.000 (0.002)	-0.002 (0.011)	0.000 (0.001)	0.1037	0.0732
	Trained predictor (M0)	5	0.623 (0.008)	0.697 (0.006)	0.291 (0.017)	0.0274	0.0732
Organoids, primary pooled	Constant	5	undefined	undefined	undefined	8.72×10^{-18}	0.2407
	Training marginals	5	0.465 (0.001)	0.588 (0.001)	undefined	6.59×10^{-18}	0.2407
	Shuffled predictions	5	0.007 (0.015)	0.050 (0.065)	0.000 (0.013)	0.3878	0.2407

Supplementary Table 6: Cross-assay reproducibility of each response component, with cluster-bootstrap intervals. q_c is the correlation between the GDSC1 and GDSC2 measurements of the same drug–sample pair, estimated on the 32,724 raw matched keys that pass benchmark key eligibility, over 89 compounds and 600 samples. Each bootstrap column gives the replicate mean and the 95% interval under one resampling scheme, 600 replicates each; the decomposition is rebuilt inside every replicate. $\sqrt{q_c}$ is the empirical reproducibility reference against which recovery is expressed. Supplementary Note 3 explains the assumptions underlying this reference; it does not define an attainable limit.

Component	q_c	Two-way cluster	Compound cluster	Sample cluster	$\sqrt{q_c}$
Total response	0.650	0.636 [0.528, 0.750]	0.636 [0.529, 0.755]	0.650 [0.640, 0.660]	0.807
Additive component	0.754	0.728 [0.567, 0.866]	0.728 [0.562, 0.868]	0.754 [0.743, 0.764]	0.869
Interaction component	0.498	0.498 [0.444, 0.554]	0.499 [0.451, 0.547]	0.498 [0.481, 0.515]	0.706

Supplementary Table 7: Resource-composition sensitivity of the M4–M0 interaction effect. Differences are paired within seed. Positive gives the number of seeds whose interaction-correlation difference is positive. Supports carry at least 1,000 matched prediction rows. These are rescored subsets of the same trained predictions and not external-validation cohorts.

Regime	Support	Mean pairs	Seeds	Mean Δr_{int}	s.d.	Positive	Mean $\Delta R^2_{\mathrm{int}}$
Unseen cell lines	NCI60	369,208	5	-0.0220	0.0368	1/5	-0.0139
	NON_NCI60	155,147	5	0.0078	0.0227	4/5	0.0020
	Prism	80,984	5	0.0240	0.0243	5/5	0.0223
	GDSC1	19,096	5	-0.0004	0.0070	2/5	-0.0039
	GDSC2	6,971	5	-0.0083	0.0406	3/5	-0.0229
	CTRP2	42,081	5	-0.0302	0.0514	1/5	-0.0466
	CTRP1	5,766	5	0.0090	0.0387	3/5	-0.0069
	CCLE	1,700	5	-0.0024	0.0233	2/5	-0.0047
Unseen compounds	NCI60	323,805	5	0.0237	0.0170	5/5	0.0216
	NON_NCI60	145,156	5	0.0081	0.0039	5/5	0.0110
	Prism	72,812	5	0.0181	0.0041	5/5	0.0264
	GDSC1	16,767	5	0.0040	0.0228	2/5	0.0102
	GDSC2	7,093	5	-0.0127	0.0310	1/5	0.0250
	CTRP2	40,817	5	0.0247	0.0251	4/5	0.0187
	CTRP1	6,111	5	0.0105	0.0200	3/5	0.0110
	CCLE	2,539	4	0.0137	0.0624	3/4	0.0121

Supplementary Table 8: Representation benchmark under held-out cell lines. Mean and standard deviation across seeds for every configuration, under a fixed nonlinear dual-tower decoder. *Full panel* records whether the representation covers the entire benchmark dataset; the two that do not are scored on the subset they cover and are not directly comparable with the rest. A_{int} is the emitted interaction amplitude. The two component-supervised arms are listed first and are the five-seed M0 and M4 runs, so they carry five seeds where the representations carry three. In the configuration names, `raw` denotes the untransformed gene-level expression input.

Class	Representation	Full panel	Seeds	Total	Additive	Interaction	A_{int}
Component-supervised arms	M0 (ECFP4 x raw)	yes	5	0.858 (0.009)	0.963 (0.007)	0.315 (0.040)	0.346 (0.042)
	M4 (ECFP4 x raw)	yes	5	0.858 (0.006)	0.963 (0.004)	0.307 (0.042)	0.376 (0.028)
Molecular representations	ECFP4	yes	3	0.859 (0.007)	0.964 (0.003)	0.308 (0.053)	0.356 (0.010)
	GeminiMol	no	3	0.828 (0.010)	0.933 (0.007)	0.283 (0.047)	0.323 (0.030)
	Mole_BERT	yes	3	0.827 (0.008)	0.929 (0.004)	0.274 (0.039)	0.338 (0.053)
	ChemBERTa2	yes	3	0.821 (0.011)	0.925 (0.007)	0.255 (0.057)	0.328 (0.014)
	MolT5	yes	3	0.829 (0.012)	0.934 (0.007)	0.249 (0.078)	0.324 (0.025)
	MolCLR	yes	3	0.742 (0.013)	0.834 (0.008)	0.238 (0.073)	0.231 (0.017)
	KPGT	yes	3	0.769 (0.016)	0.865 (0.019)	0.236 (0.069)	0.286 (0.037)
	UniMolV2	no	3	0.757 (0.028)	0.853 (0.029)	0.225 (0.035)	0.276 (0.039)
	UniMol	yes	3	0.768 (0.020)	0.868 (0.015)	0.208 (0.024)	0.292 (0.049)
	InfoAlign	yes	3	0.757 (0.022)	0.862 (0.015)	0.183 (0.106)	0.385 (0.031)
	Ouroboros	yes	3	0.678 (0.018)	0.766 (0.016)	0.170 (0.030)	0.222 (0.016)
	Chemprop	yes	3	0.513 (0.021)	0.587 (0.031)	0.118 (0.002)	0.284 (0.047)
Transcriptomic representations	SCimilarity	yes	3	0.865 (0.004)	0.969 (0.005)	0.318 (0.054)	0.350 (0.035)
	raw	yes	3	0.859 (0.007)	0.964 (0.003)	0.308 (0.053)	0.356 (0.010)
	Tahoe_x1	yes	3	0.857 (0.007)	0.963 (0.004)	0.293 (0.050)	0.368 (0.029)
	scGPT	yes	3	0.860 (0.006)	0.967 (0.003)	0.290 (0.019)	0.374 (0.022)
	STATE	yes	3	0.849 (0.014)	0.959 (0.007)	0.276 (0.073)	0.433 (0.011)
	scFoundation	yes	3	0.858 (0.003)	0.967 (0.002)	0.272 (0.064)	0.376 (0.074)
	CellPLM	yes	3	0.857 (0.013)	0.964 (0.008)	0.271 (0.053)	0.306 (0.042)
	Cell2Sentence	yes	3	0.852 (0.015)	0.963 (0.012)	0.257 (0.034)	0.420 (0.035)
	BulkFormer	yes	3	0.849 (0.017)	0.960 (0.009)	0.252 (0.030)	0.356 (0.085)
	Geneformer	yes	3	0.853 (0.012)	0.965 (0.011)	0.243 (0.056)	0.392 (0.024)
	tGPT	yes	3	0.846 (0.015)	0.961 (0.009)	0.209 (0.020)	0.391 (0.017)
	UCE	yes	3	0.813 (0.006)	0.938 (0.015)	0.013 (0.049)	0.288 (0.069)

Supplementary Table 9: Representation benchmark under held-out compounds. Columns as in Supplementary Table 8.

Class	Representation	Full panel	Seeds	Total	Additive	Interaction	A_{int}
Component-supervised arms	M4 (ECFP4 x raw)	yes	5	0.609 (0.013)	0.678 (0.014)	0.308 (0.013)	0.359 (0.012)
	M0 (ECFP4 x raw)	yes	5	0.623 (0.008)	0.697 (0.006)	0.291 (0.017)	0.374 (0.007)
Molecular representations	ECFP4	yes	3	0.625 (0.010)	0.699 (0.008)	0.301 (0.014)	0.371 (0.002)
	GeminiMol	no	3	0.585 (0.016)	0.663 (0.016)	0.232 (0.012)	0.313 (0.033)
	Mole_BERT	yes	3	0.575 (0.017)	0.649 (0.016)	0.228 (0.011)	0.327 (0.025)
	MolCLR	yes	3	0.573 (0.009)	0.648 (0.011)	0.224 (0.017)	0.288 (0.007)
	MolT5	yes	3	0.531 (0.005)	0.599 (0.004)	0.214 (0.008)	0.296 (0.017)
	UniMolV2	no	3	0.565 (0.012)	0.640 (0.013)	0.207 (0.005)	0.280 (0.009)
	UniMol	yes	3	0.520 (0.019)	0.590 (0.020)	0.189 (0.003)	0.302 (0.013)
	Ouroboros	yes	3	0.537 (0.007)	0.611 (0.005)	0.186 (0.003)	0.302 (0.026)
	ChemBERTa2	yes	3	0.512 (0.003)	0.581 (0.006)	0.183 (0.005)	0.316 (0.038)
	KPGT	yes	3	0.360 (0.021)	0.411 (0.024)	0.097 (0.011)	0.186 (0.043)
	Chemprop	yes	3	0.173 (0.014)	0.201 (0.015)	0.006 (0.011)	0.058 (0.011)
	InfoAlign	yes	3	0.160 (0.007)	0.187 (0.007)	-0.001 (0.005)	0.067 (0.004)
Transcriptomic representations	STATE	yes	3	0.628 (0.009)	0.702 (0.007)	0.303 (0.019)	0.400 (0.019)
	scGPT	yes	3	0.625 (0.012)	0.698 (0.012)	0.303 (0.013)	0.378 (0.009)
	raw	yes	3	0.625 (0.010)	0.699 (0.008)	0.301 (0.014)	0.371 (0.002)
	Tahoe_x1	yes	3	0.623 (0.008)	0.697 (0.008)	0.300 (0.003)	0.384 (0.033)
	scFoundation	yes	3	0.620 (0.007)	0.693 (0.007)	0.296 (0.009)	0.376 (0.022)
	BulkFormer	yes	3	0.620 (0.017)	0.694 (0.016)	0.293 (0.012)	0.384 (0.015)
	SCimilarity	yes	3	0.613 (0.012)	0.686 (0.013)	0.290 (0.006)	0.353 (0.021)
	tGPT	yes	3	0.615 (0.016)	0.689 (0.017)	0.287 (0.016)	0.376 (0.023)
	CellPLM	yes	3	0.614 (0.016)	0.687 (0.015)	0.285 (0.017)	0.364 (0.024)
	Cell2Sentence	yes	3	0.618 (0.009)	0.693 (0.007)	0.280 (0.011)	0.342 (0.018)
	UCE	yes	3	0.618 (0.006)	0.693 (0.006)	0.276 (0.011)	0.355 (0.030)
	Geneformer	yes	3	0.610 (0.011)	0.685 (0.011)	0.268 (0.013)	0.339 (0.020)

Supplementary Table 10: Descriptive score separation in the representation benchmark. Spread is the range of the mean score across representations on one axis, median seed s.d. is the median across representations of the seed-to-seed standard deviation, and Ratio is their quotient. This ratio describes score separation relative to seed variation; it is not a categorical test of pairwise separation, statistical equivalence or ranking stability.

Regime	Axis	Component	Spread	Median seed s.d.	Ratio
Unseen cell lines	Molecular	Total response	0.3464	0.0147	23.6
		Additive component	0.3771	0.0114	33.0
		Interaction component	0.1897	0.0550	3.5
	Transcriptomic	Total response	0.0520	0.0097	5.4
		Additive component	0.0310	0.0079	3.9
		Interaction component	0.3055	0.0514	5.9
Unseen compounds	Molecular	Total response	0.4650	0.0097	47.9
		Additive component	0.5119	0.0095	54.0
		Interaction component	0.3020	0.0095	31.9
	Transcriptomic	Total response	0.0186	0.0106	1.8
		Additive component	0.0175	0.0098	1.8
		Interaction component	0.0352	0.0126	2.8

Supplementary Table 11: Component-wise recovery under three decoders. Mean and standard deviation across the runs of each configuration. The additive ridge decoder is additive by construction, so its prediction lies in the additive subspace and it emits an interaction of standard deviation 10^{-9} to 10^{-8}, against an observed interaction standard deviation of order 7×10^{-2}; its interaction-component correlations are the correlations of floating-point noise and are not a ranking, as Supplementary Note 2 sets out. Three of the sixty low-rank bilinear runs, all at the largest weight decay, collapse to the same condition. $\mathrm{sd}(M\hat{y})$ is the emitted interaction amplitude in absolute units. `raw` denotes the untransformed gene-level expression input.

Decoder	Molecular	Transcriptomic	Runs	Parameters	Total	Additive	Interaction	$\mathrm{sd}(M\hat{y})$
Additive ridge	ECFP4	BulkFormer	4	2,689	0.408 (0.103)	0.461 (0.117)	0.009 (0.007)	0.0000
	ECFP4	CellPLM	4	2,561	0.483 (0.111)	0.546 (0.125)	0.005 (0.004)	0.0000
	ECFP4	raw	4	18,010	0.254 (0.183)	0.287 (0.206)	0.004 (0.006)	0.0000
	ECFP4	scGPT	4	2,561	0.470 (0.094)	0.531 (0.106)	-0.007 (0.010)	0.0000
	KPGT	raw	4	18,266	0.176 (0.067)	0.199 (0.076)	-0.001 (0.002)	0.0000
	MolCLR	raw	4	16,474	0.099 (0.094)	0.112 (0.106)	-0.000 (0.001)	0.0000
	UniMol	raw	4	16,474	0.151 (0.124)	0.171 (0.140)	0.000 (0.001)	0.0000
Low-rank bilinear	ECFP4	BulkFormer	12	88,705	0.351 (0.114)	0.419 (0.118)	0.051 (0.040)	0.0559
	ECFP4	CellPLM	12	84,481	0.493 (0.106)	0.569 (0.126)	0.080 (0.053)	0.0185
	ECFP4	raw	12	594,298	0.266 (0.147)	0.304 (0.170)	0.047 (0.038)	0.0501
	ECFP4	scGPT	12	84,481	0.391 (0.104)	0.466 (0.108)	0.058 (0.050)	0.0431
	KPGT	raw	4	602,746	0.181 (0.087)	0.207 (0.100)	0.024 (0.021)	0.0581
	MolCLR	raw	4	543,610	0.074 (0.075)	0.084 (0.085)	0.009 (0.009)	0.0632
	UniMol	raw	4	543,610	0.125 (0.092)	0.141 (0.105)	0.018 (0.027)	0.0370
Nonlinear dual tower	ECFP4	BulkFormer	3	3,812,353	0.845 (0.011)	0.951 (0.009)	0.292 (0.010)	0.0282
	ECFP4	CellPLM	3	3,681,281	0.856 (0.020)	0.962 (0.015)	0.275 (0.049)	0.0224
	ECFP4	raw	3	19,501,057	0.856 (0.011)	0.957 (0.007)	0.332 (0.064)	0.0262
	ECFP4	scGPT	3	3,681,281	0.857 (0.008)	0.962 (0.007)	0.300 (0.028)	0.0257
	KPGT	raw	1	19,632,129	0.836	0.933	0.278	0.0202
	MolCLR	raw	1	18,714,625	0.817	0.911	0.288	0.0209
	UniMol	raw	1	18,714,625	0.837	0.933	0.294	0.0219

Supplementary Table 12: Agreement between representation rankings under different decoders. ρ_s is the Spearman correlation between the two rankings named in Comparison, over the representations that both decoders were run on. Top 1 and Top 2 record whether the leading representation and the leading pair are the same under both. Main-figure decoder-pair rows compare four representations at seed 3407; the `3seed` rows compare a decoder against itself across three seeds and give the mean ordering it produces. Comparisons on the interaction component are omitted where one of the two decoders emits no interaction. `raw` denotes the untransformed gene-level expression input.

Scope	Axis	Component	Comparison	ρ_s	Top 1	Top 2	Mean ordering
seed3407	drug side (transcriptome = raw)	Total	ridge vs bilinear	0.80	yes	no	
seed3407	drug side (transcriptome = raw)	Total	ridge vs tower	1.00	yes	yes	
seed3407	drug side (transcriptome = raw)	Total	bilinear vs tower	0.80	yes	no	
seed3407	drug side (transcriptome = raw)	Additive	ridge vs bilinear	0.80	yes	no	
seed3407	drug side (transcriptome = raw)	Additive	ridge vs tower	1.00	yes	yes	
seed3407	drug side (transcriptome = raw)	Additive	bilinear vs tower	0.80	yes	no	
seed3407	drug side (transcriptome = raw)	Interaction	bilinear vs tower	0.20	yes	no	
seed3407	transcriptome side (drug = ECFP4)	Total	ridge vs bilinear	0.80	no	yes	
seed3407	transcriptome side (drug = ECFP4)	Total	ridge vs tower	0.40	yes	no	
seed3407	transcriptome side (drug = ECFP4)	Total	bilinear vs tower	0.00	no	no	
seed3407	transcriptome side (drug = ECFP4)	Additive	ridge vs bilinear	1.00	yes	yes	
seed3407	transcriptome side (drug = ECFP4)	Additive	ridge vs tower	0.80	yes	yes	
seed3407	transcriptome side (drug = ECFP4)	Additive	bilinear vs tower	0.80	yes	yes	
seed3407	transcriptome side (drug = ECFP4)	Interaction	bilinear vs tower	-0.80	no	no	
3seed	transcriptome	Total	bilinear, seed stability	0.87	no	—	scGPT > CellPLM > BulkFormer > raw
3seed	transcriptome	Additive	bilinear, seed stability	1.00	yes	—	CellPLM > scGPT > BulkFormer > raw
3seed	transcriptome	Interaction	bilinear, seed stability	1.00	yes	—	scGPT > CellPLM > BulkFormer > raw
3seed	transcriptome	Total	tower, seed stability	-0.20	no	—	scGPT > raw > CellPLM > BulkFormer
3seed	transcriptome	Additive	tower, seed stability	0.40	no	—	CellPLM > scGPT > raw > BulkFormer
3seed	transcriptome	Interaction	tower, seed stability	-0.33	no	—	raw > scGPT > BulkFormer > CellPLM

Supplementary Table 13: Component-wise error geometry, per run. d_H is the dimension of the additive subspace and d_H/N its share of the degrees of freedom, in per cent. f_H is the share of total squared prediction error that falls in the additive subspace. D_H and D_M are the rank-normalized error densities of the two subspaces and $C_{\text{error}} = D_H/D_M$ their ratio. E_{add} and E_{int} are the two error terms of the exact attribution MSE $= E_{\text{add}} + E_{\text{int}}$. The three error quantities are given in units of 10^{-3}.

Regime	Arm	Seed	N	d_H	d_H/N	f_H	D_H	D_M	C_{error}	E_{add}	E_{int}
Unseen cell lines	M0	3407	557,484	54,323	9.74	0.257	15.82	4.93	3.21	1.54	4.45
	M0	3408	474,279	54,323	11.45	0.222	11.44	5.19	2.20	1.31	4.59
	M0	3409	486,991	54,319	11.15	0.207	11.29	5.42	2.08	1.26	4.82
	M0	3410	570,863	54,322	9.52	0.303	19.99	4.83	4.14	1.90	4.37
	M0	3411	530,351	54,322	10.24	0.182	9.92	5.09	1.95	1.02	4.57
	M4	3407	557,484	54,323	9.74	0.255	15.75	4.96	3.18	1.53	4.47
	M4	3408	474,279	54,323	11.45	0.239	12.48	5.13	2.43	1.43	4.54
	M4	3409	486,991	54,319	11.15	0.206	11.25	5.43	2.07	1.25	4.83
	M4	3410	570,863	54,322	9.52	0.247	15.17	4.86	3.12	1.44	4.40

Supplementary Table 13: Continued.

Regime	Arm	Seed	N	d_H	d_H/N	f_H	D_H	D_M	C_{error}	E_{add}	E_{int}
	M4	3411	530,351	54,322	10.24	0.195	11.19	5.28	2.12	1.15	4.74
Unseen compounds	M0	3407	468,871	9,112	1.94	0.632	431.36	4.97	86.81	8.38	4.87
	M0	3408	466,230	9,110	1.95	0.634	448.38	5.15	87.04	8.76	5.05
	M0	3409	459,058	9,110	1.98	0.642	449.94	5.08	88.52	8.93	4.98
	M0	3410	475,001	9,112	1.92	0.643	472.71	5.12	92.28	9.07	5.02
	M0	3411	474,086	9,110	1.92	0.641	444.60	4.88	91.19	8.54	4.78
	M4	3407	468,871	9,112	1.94	0.651	459.78	4.88	94.31	8.94	4.78
	M4	3408	466,230	9,110	1.95	0.655	484.06	5.09	95.14	9.46	4.99
	M4	3409	459,058	9,110	1.98	0.652	467.14	5.05	92.45	9.27	4.95
	M4	3410	475,001	9,112	1.92	0.640	455.80	5.01	91.04	8.74	4.91
	M4	3411	474,086	9,110	1.92	0.658	468.39	4.78	97.99	9.00	4.69

Supplementary Table 14: Recovery relative to the empirical reproducibility reference, per run. $\eta_c = r_c/\sqrt{q_c}$ on the matched support that each regime actually holds out, whose size is given in the fourth column. The held-out compound rows are listed for completeness and are not interpreted in the main text: only 7 to 19 held-out compounds contribute to them, and Supplementary Note 3 gives the per-seed instability that follows.

Regime	Arm	Seed	Matched pairs	η_{total}	η_{add}	η_{int}
Unseen cell lines	M0	3407	2,968	0.861	1.032	0.358
	M0	3408	3,015	0.829	1.009	0.365
	M0	3409	2,904	0.831	1.011	0.361
	M0	3410	3,051	0.815	0.973	0.391
	M0	3411	2,876	0.858	0.997	0.375
	M4	3407	2,968	0.833	1.002	0.334
	M4	3408	3,015	0.822	1.003	0.375
	M4	3409	2,904	0.824	1.008	0.350
	M4	3410	3,051	0.820	0.980	0.412
	M4	3411	2,876	0.855	0.986	0.406
Unseen compounds	M0	3407	2,721	0.297	0.322	0.120
	M0	3408	3,470	0.365	0.443	0.261
	M0	3409	1,970	0.552	0.564	0.538
	M0	3410	3,873	0.590	0.646	0.221
	M0	3411	2,998	0.301	0.338	0.032
	M4	3407	2,721	0.344	0.366	0.162
	M4	3408	3,470	0.380	0.476	0.239
	M4	3409	1,970	0.263	0.177	0.675
	M4	3410	3,873	0.574	0.630	0.234
	M4	3411	2,998	0.140	0.163	0.005

Supplementary Table 15: Prespecified input-compatibility check, per organoid cohort. Level and s.d. summarize the distribution of the transcriptomic input over each cohort, Gap vs CCLE is the absolute difference in level from the training reference in the first row, and Zeros is the fraction of entries equal to zero. Samples counts the expression profiles of each cohort, including organoids with no evaluated response pair; the 173 organoids of the zero-shot evaluation (Fig. 1b) are those with at least one, and Rows counts their drug–organoid pairs. *Primary* is the prespecified protocol decision, fixed before this evaluation was run. *Compatible* is the automated three-condition diagnostic. The two agree on four of the five cohorts. UMPDO3 sits 0.005 outside the level-mean threshold of 1.0 while matching UMPDO1 and UMPDO2 on zero fraction, minimum and range, so the binary boundary falls inside a scale family that the continuous evidence does not separate. The thresholds and primary cohort set were prespecified and retained.

Cohort	Samples	Genes	Level	s.d.	Gap vs CCLE	Zeros	Min	Primary	Compatible	Rows
CCLE (training reference)	—	—	2.702	0.221	0.000	0.153	0.00	no	yes	—
HKUPDO	52	8,866	0.008	0.075	2.694	0.951	-9.97	no	no	324
LICOB	65	13,224	0.399	0.015	2.303	0.000	-1.80	no	no	4,800
UMPDO1	90	15,413	3.578	0.128	0.876	0.114	0.00	yes	yes	2,744
UMPDO2	34	15,218	3.397	0.159	0.694	0.158	0.00	yes	yes	1,470
UMPDO3	14	14,651	3.707	0.198	1.005	0.094	0.00	yes	no	672

Supplementary Table 16: Matched three-arm comparison. M0, M3 and M4 were trained on the same dataset, manifests and five seeds and selected by minimum validation total-response error. M3 and M4 have identical architecture and parameter count; only M4 receives component-wise supervision. Entries are mean (s.d.) over seeds.

Regime	Arm	Seeds	Mean pairs	Total	Additive	Interaction	A_{int}	Interaction R^2	Parameters
Unseen cell lines	M0	5	523,994	0.858 (0.009)	0.963 (0.007)	0.315 (0.040)	0.346 (0.042)	0.097 (0.029)	19,501,057
	M3	5	523,994	0.855 (0.010)	0.960 (0.007)	0.306 (0.044)	0.369 (0.030)	0.090 (0.031)	20,029,444
	M4	5	523,994	0.858 (0.006)	0.963 (0.004)	0.307 (0.042)	0.376 (0.028)	0.090 (0.032)	20,029,444
Unseen compounds	M0	5	468,649	0.623 (0.008)	0.697 (0.006)	0.291 (0.017)	0.374 (0.007)	0.078 (0.014)	19,501,057
	M3	5	468,649	0.618 (0.013)	0.692 (0.012)	0.279 (0.018)	0.354 (0.023)	0.072 (0.012)	20,029,444
	M4	5	468,649	0.609 (0.013)	0.678 (0.014)	0.308 (0.013)	0.359 (0.012)	0.092 (0.009)	20,029,444

Supplementary Table 17: Model choices and their independent GDSC2 endpoints. Each criterion selects M0 or M4 on GDSC1 separately by split and seed; M4 of 5 is the number of M4 choices. Δ MAE is the paired independent GDSC2 mean absolute error relative to global-correlation selection, so positive is worse. Sign, PCC and Spearman are secondary endpoints. The prespecified primary success rule was not met.

Regime	Selection criterion	M4 of 5	GDSC2 MAE	Δ MAE	Sign	PCC	Spearman
Unseen cell lines	Global	1	0.13716	0.00000	0.5767	0.2815	0.2490
	Fixed drug	4	0.14079	0.00363	0.5778	0.2795	0.2489
	Fixed cell	1	0.13716	0.00000	0.5767	0.2815	0.2490
	DrEval normalized	1	0.13716	0.00000	0.5767	0.2815	0.2490
	DrugDis interaction direction	3	0.13961	0.00245	0.5792	0.2879	0.2558
	DrugDis interaction R^2	1	0.13716	0.00000	0.5767	0.2815	0.2490
Unseen compounds	Global	2	0.12980	0.00000	0.5636	0.2979	0.2269
	Fixed drug	2	0.13056	0.00076	0.5651	0.2900	0.2321
	Fixed cell	2	0.12980	0.00000	0.5636	0.2979	0.2269
	DrEval normalized	2	0.12980	0.00000	0.5636	0.2979	0.2269
	DrugDis interaction direction	4	0.13023	0.00043	0.5655	0.3094	0.2334
	DrugDis interaction R^2	3	0.13009	0.00029	0.5662	0.3018	0.2330

Supplementary Table 18: Amplitude-sensitive interaction variance explained. Interaction is the component correlation, A_{int} is the emitted amplitude relative to the observed interaction, and $R^2_{\text{int}} = 2A_{\text{int}}r_{\text{int}} - A^2_{\text{int}}$. Entries are mean (s.d.) over five runs selected by minimum validation mean squared error. Unlike a squared correlation, this quantity accounts for amplitude and can be negative.

Regime	Arm	Seeds	Interaction	A_{int}	R^2_{int}
Unseen cell lines	M0	5	0.315 (0.040)	0.346 (0.042)	0.097 (0.029)
	M4	5	0.307 (0.042)	0.376 (0.028)	0.090 (0.032)
Unseen compounds	M0	5	0.291 (0.017)	0.374 (0.007)	0.078 (0.014)
	M4	5	0.308 (0.013)	0.359 (0.012)	0.092 (0.009)
Organoids, primary pooled	M0	5	0.067 (0.010)	0.179 (0.049)	-0.010 (0.012)
	M4	5	0.087 (0.009)	0.209 (0.021)	-0.008 (0.007)

Supplementary Table 19: Independent assay-generation selectivity evaluation by model arm. GDSC1 is the model-selection assay generation and GDSC2 the independent test generation. Each split and seed uses 50,000 support-sampled double-difference rectangles. Values are means over five seeds.

Regime	Arm	Seeds	Mean pairs	GDSC1 MAE	GDSC1 sign	GDSC1 PCC	GDSC2 MAE	GDSC2 sign	GDSC2 PCC
Unseen cell lines	M0	5	5,138	0.1946	0.5696	0.2315	0.1371	0.5782	0.2817
	M4	5	5,138	0.1979	0.5694	0.2252	0.1429	0.5747	0.2629
Unseen compounds	M0	5	5,027	0.1804	0.5525	0.2133	0.1303	0.5624	0.2837
	M4	5	5,027	0.1794	0.5572	0.2361	0.1303	0.5651	0.3055

Supplementary Table 20: The component-supervised predictor against the baseline, per seed. M0 is the total-response-supervised baseline. M4 adds two effect heads, component-wise supervision and 2.7% trainable parameters. Arms are paired within seed, and the two regimes use the same five prespecified splits. MSE, E_{add} and E_{int} are given in units of 10^{-3} and satisfy MSE $= E_{\text{add}} + E_{\text{int}}$ exactly. Epoch is the checkpoint selected by the prespecified rule, the epoch of lowest validation total prediction error.

Regime	Arm	Seed	N	Total	Additive	Interaction	A_{int}	MSE	E_{add}	E_{int}	Epoch	Parameters
Unseen cell lines	M0	3407	557,484	0.8671	0.9661	0.3160	0.346	5.99	1.54	4.45	16	19,501,057
	M0	3408	474,279	0.8566	0.9608	0.3561	0.356	5.90	1.31	4.59	24	19,501,057
	M0	3409	486,991	0.8535	0.9642	0.2507	0.365	6.07	1.26	4.82	30	19,501,057
	M0	3410	570,863	0.8466	0.9514	0.3173	0.277	6.27	1.90	4.37	12	19,501,057
	M0	3411	530,351	0.8667	0.9714	0.3371	0.388	5.59	1.02	4.57	30	19,501,057
	M4	3407	557,484	0.8653	0.9654	0.3101	0.364	6.01	1.53	4.47	19	20,029,444
	M4	3408	474,279	0.8604	0.9640	0.3724	0.416	5.97	1.43	4.54	24	20,029,444
	M4	3409	486,991	0.8547	0.9648	0.2570	0.394	6.08	1.25	4.83	27	20,029,444
	M4	3410	570,863	0.8501	0.9556	0.3087	0.357	5.84	1.44	4.40	20	20,029,444
	M4	3411	530,351	0.8588	0.9668	0.2886	0.351	5.88	1.15	4.74	19	20,029,444
Unseen compounds	M0	3407	468,871	0.6157	0.6925	0.2869	0.373	13.26	8.38	4.87	22	19,501,057
	M0	3408	466,230	0.6358	0.7083	0.3154	0.369	13.81	8.76	5.05	25	19,501,057
	M0	3409	459,058	0.6244	0.6965	0.3012	0.370	13.91	8.93	4.98	22	19,501,057
	M0	3410	475,001	0.6201	0.6953	0.2750	0.372	14.09	9.07	5.02	14	19,501,057
	M0	3411	474,086	0.6195	0.6940	0.2762	0.386	13.33	8.54	4.78	26	19,501,057
	M4	3407	468,871	0.5900	0.6584	0.3086	0.362	13.72	8.94	4.78	22	20,029,444
	M4	3408	466,230	0.6144	0.6806	0.3290	0.351	14.45	9.46	4.99	21	20,029,444
	M4	3409	459,058	0.6126	0.6817	0.3051	0.346	14.22	9.27	4.95	16	20,029,444
	M4	3410	475,001	0.6233	0.6960	0.3044	0.376	13.65	8.74	4.91	28	20,029,444
	M4	3411	474,086	0.6034	0.6723	0.2947	0.361	13.69	9.00	4.69	30	20,029,444

Supplementary Table 21: Zero-shot organoid evaluation, pooled and per cohort. Mean and standard deviation across five seeds. No organoid data entered training, checkpoint selection or tuning in any row. The first block is the prespecified primary analysis over UMPDO1–3, of which UMPDO1–2 meet every compatibility condition and UMPDO3 is a prespecified near-miss; the second pools every cohort and assesses sensitivity to including cohorts with different input scales; the remaining rows give each cohort separately, with its own eligibility flags.

Cohort	Primary	Compatible	Arm	Seeds	N	Total	Additive	Interaction	A_{int}
Primary pooled			M0	5	4,886	0.460 (0.007)	0.607 (0.011)	0.067 (0.010)	0.179 (0.049)
			M4	5	4,886	0.466 (0.036)	0.614 (0.046)	0.087 (0.009)	0.209 (0.021)
All cohorts, pooled			M0	5	10,010	0.271 (0.097)	0.356 (0.128)	0.040 (0.030)	0.206 (0.041)
			M4	5	10,010	0.382 (0.048)	0.500 (0.058)	0.078 (0.026)	0.235 (0.015)
HKUPDO	no	no	M0	5	324	0.442 (0.088)	0.493 (0.106)	-0.005 (0.092)	0.563 (0.165)
	no	no	M4	5	324	0.358 (0.096)	0.402 (0.106)	-0.015 (0.096)	0.736 (0.051)
LICOB	no	no	M0	5	4,800	0.268 (0.045)	0.377 (0.063)	0.066 (0.019)	0.118 (0.096)
	no	no	M4	5	4,800	0.283 (0.035)	0.402 (0.036)	0.054 (0.066)	0.224 (0.075)
UMPDO1	yes	yes	M0	5	2,744	0.561 (0.014)	0.662 (0.015)	0.054 (0.017)	0.221 (0.049)
	yes	yes	M4	5	2,744	0.547 (0.034)	0.648 (0.036)	0.077 (0.013)	0.279 (0.023)
UMPDO2	yes	yes	M0	5	1,470	0.358 (0.011)	0.552 (0.017)	0.058 (0.019)	0.092 (0.017)
	yes	yes	M4	5	1,470	0.364 (0.023)	0.560 (0.037)	0.069 (0.014)	0.097 (0.024)
UMPDO3	yes	no	M0	5	672	0.383 (0.024)	0.461 (0.038)	0.140 (0.052)	0.147 (0.036)
	yes	no	M4	5	672	0.391 (0.048)	0.466 (0.049)	0.167 (0.107)	0.187 (0.009)

Supplementary Table 22: Zero-shot organoid evaluation of the benchmark M3 control. LCLO checkpoints were applied without retraining or model selection on organoid data. The primary pooled set is UMPDO1, UMPDO2 and UMPDO3. M0 and M4 values are included for the matched three-arm comparison. Entries are mean (s.d.) over five seeds.

Cohort	Arm	Seeds	N	Additive	Interaction	A_{int}	Interaction R^2
Primary pooled	M0	5	4,886	0.607 (0.011)	0.067 (0.010)	0.179 (0.049)	-0.010 (0.012)
	M3	5	4,886	0.621 (0.024)	0.061 (0.013)	0.165 (0.059)	-0.009 (0.011)
	M4	5	4,886	0.614 (0.046)	0.087 (0.009)	0.209 (0.021)	-0.008 (0.007)
UMPDO1	M0	5	2,744	0.662 (0.015)	0.054 (0.017)	0.221 (0.049)	-0.028 (0.020)
	M3	5	2,744	0.653 (0.028)	0.048 (0.026)	0.213 (0.099)	-0.031 (0.032)
	M4	5	2,744	0.648 (0.036)	0.077 (0.013)	0.279 (0.023)	-0.035 (0.010)
UMPDO2	M0	5	1,470	0.552 (0.017)	0.058 (0.019)	0.092 (0.017)	0.002 (0.003)

Supplementary Table 22: Continued.

Cohort	Arm	Seeds	N	Additive	Interaction	A_{int}	Interaction R^2
	M3	5	1,470	0.548 (0.025)	0.052 (0.018)	0.074 (0.020)	0.002 (0.004)
	M4	5	1,470	0.560 (0.037)	0.069 (0.014)	0.097 (0.024)	0.004 (0.002)
UMPDO3	M0	5	672	0.461 (0.038)	0.140 (0.052)	0.147 (0.036)	0.019 (0.013)
	M3	5	672	0.483 (0.051)	0.115 (0.089)	0.136 (0.059)	0.015 (0.017)
	M4	5	672	0.466 (0.049)	0.167 (0.107)	0.187 (0.009)	0.028 (0.040)
ALL	M0	5	10,010	0.356 (0.128)	0.040 (0.030)	0.206 (0.041)	-0.027 (0.016)
	M3	5	10,010	0.354 (0.150)	0.032 (0.031)	0.186 (0.038)	-0.023 (0.009)
	M4	5	10,010	0.500 (0.058)	0.078 (0.026)	0.235 (0.015)	-0.019 (0.011)
LICOB	M0	5	4,800	0.377 (0.063)	0.066 (0.019)	0.118 (0.096)	-0.005 (0.021)
	M3	5	4,800	0.410 (0.026)	0.072 (0.048)	0.106 (0.059)	0.002 (0.013)
	M4	5	4,800	0.402 (0.036)	0.054 (0.066)	0.224 (0.075)	-0.032 (0.047)
HKUPDO	M0	5	324	0.493 (0.106)	-0.005 (0.092)	0.563 (0.165)	-0.326 (0.154)
	M3	5	324	0.468 (0.096)	0.013 (0.054)	0.690 (0.166)	-0.482 (0.264)
	M4	5	324	0.402 (0.106)	-0.015 (0.096)	0.736 (0.051)	-0.565 (0.146)

Supplementary Data

Supplementary Data 1: Data sources for cell-line and organoid response measurements

The benchmark uses the processed DROMA collection (Zenodo record 10.5281/zenodo.18503188; input-data provenance record 10.5281/zenodo.17498421) and the CCLE-derived transcriptomic matrix. The table below gives the DROMA project aliases used by the analysis, the direct processed source and any upstream publication or accession that is verifiable from the available metadata. Upstream accession identifiers or source-release identifiers are absent from the DROMA metadata for these resources. The resources and their response-row counts are given in Supplementary Table 1, and the organoid cohorts in Supplementary Table 15.

Supplementary Data 1 contains the direct analysed source, direct DOI, upstream publication and accession fields where available, and a provenance note for every response resource and PDO cohort. Entries marked "not exposed" indicate that an upstream identifier is absent from the DROMA metadata.

Resource	DROMA project	Primary publication
NCI60	DROMA `NCI60_drug`	Holbeck et al., DOI 10.1158/1535-7163.MCT-10-0106
PRISM	DROMA `PRISM_drug`	Corsello et al., DOI 10.1038/s43018-019-0018-6
CTRP1	DROMA `CTRP1_drug`	Basu et al., DOI 10.1016/j.cell.2013.08.003
CTRP2	DROMA `CTRP2_drug`	Seashore-Ludlow et al., DOI 10.1158/2159-8290.CD-15-0235
GDSC1	DROMA `GDSC1_drug`	Garnett et al., DOI 10.1038/nature11005
GDSC2	DROMA `GDSC2_drug`	Iorio et al., DOI 10.1016/j.cell.2016.06.017
CCLE	DROMA `CCLE_drug` + CCLE expression	Barretina et al., DOI 10.1038/nature11003
GRAY	DROMA `GRAY_drug`	Daemen et al., DOI 10.1186/gb-2013-14-10-r110
gCSI	DROMA `gCSI_drug`	Haverty et al., DOI 10.1038/nature17987
FIMM	DROMA `FIMM_drug`	Pemovska et al., DOI 10.1158/2159-8290.CD-13-0350
UHNBreast	DROMA `UHNBreast_drug`	Marcotte et al., DOI 10.1016/j.cell.2015.11.062
UMPDO1	DROMA `UMPDO1_drug`	upstream publication not identifiable from available metadata
UMPDO2	DROMA `UMPDO2_drug`	upstream publication not identifiable from available metadata
UMPDO3	DROMA `UMPDO3_drug`	upstream publication not identifiable from available metadata
LICOB	DROMA `LICOB_drug`	Ji et al., DOI 10.1126/scitranslmed.adg3358
HKUPDO	DROMA `HKUPDO_drug`	upstream publication not identifiable from available metadata

Data sources. Sources for the eleven cell-line response resources and five patient-derived organoid cohorts. The harmonized DROMA record is the direct reproducible route used for the analysed rows; upstream accessions that are not present in the DROMA metadata are reported as a provenance limitation.