Supplementary Information for Learning reproducible cellular perturbation responses from limited independent measurements

Weijing Zhao[1†], Bo Li[1†], Bob Zhang[1*], Yuqi Wang[1], Shuping Zhao[2], Chengliang Liu[1]

[1]Department of Artificial Intelligence, University of Macau, Taipa, Macau SAR, China.

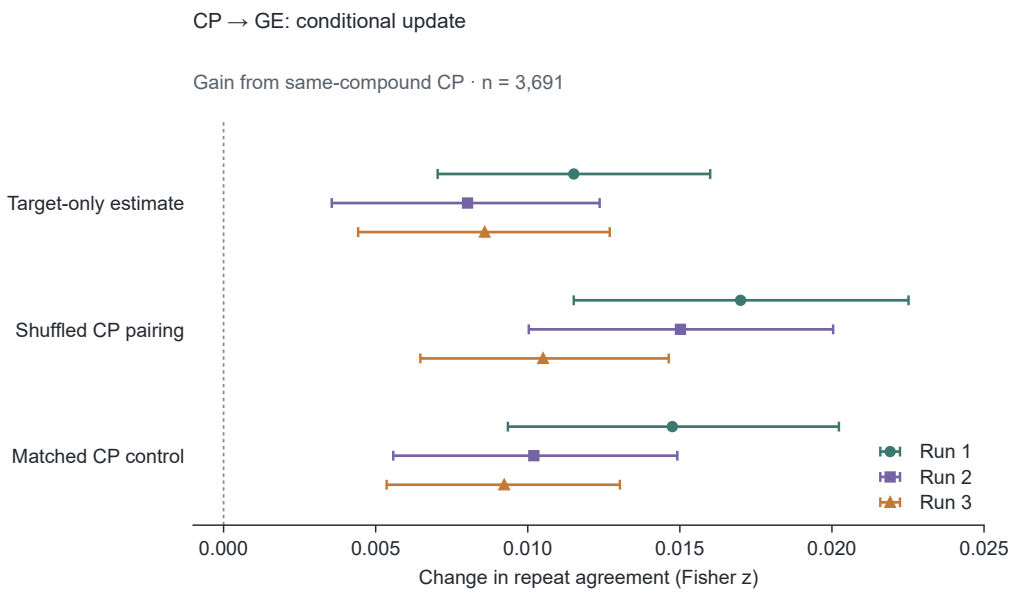
[2]School of Computer Science, Guangdong University of Technology, China.

*Corresponding author(s). E-mail(s): bobzhang@um.edu.mo;

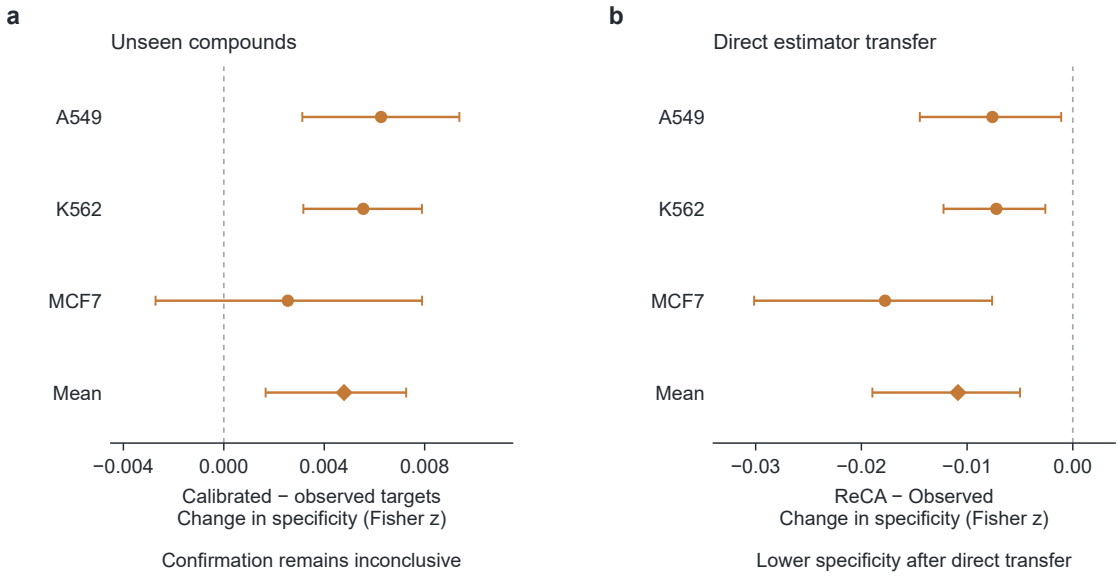
[†]These authors contributed equally to this work.

This file contains three supplementary figures, six supplementary tables and detailed methods supporting the main manuscript.

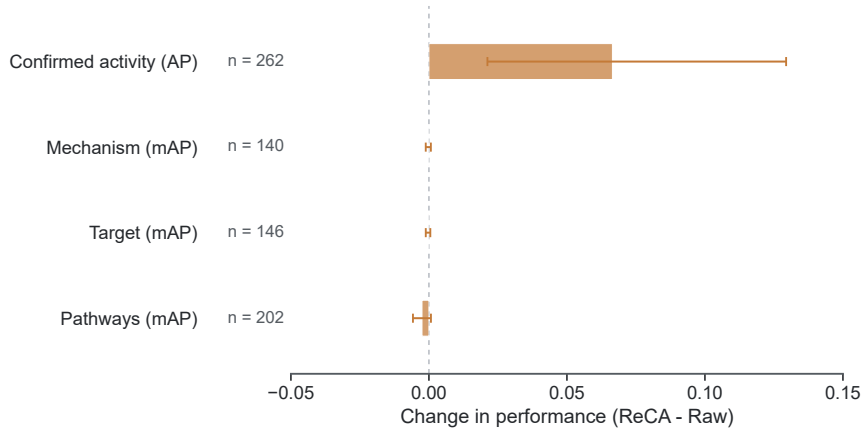
Supplementary figures



Supplementary Figure S1 | Training-run results for conditional CP-to-GE updating. Paired changes in held-out expression agreement for same-compound CP evidence relative to the target-only GE estimate, shuffled CP pairing and condition-matched CP. Each run uses 3,691 compounds. The mapper conditions on the current GE estimate and observed CP, and its update weight is selected on validation compounds. Runs 1, 2 and 3 use random seeds 3407, 42 and 2025, respectively. Points show run-specific mean paired differences. Effects are changes in repeat agreement (Fisher z). Compound-pooled estimates and intervals are shown in Fig. 3a (right).



Supplementary Figure S2 | Cross-environment transfer in sci-Plex3. a, Prediction for unseen compounds in held-out cell-line/plate blocks, using 36 compounds and 115 compound–dose conditions per target. The contrast compares matched predictors trained on calibrated versus observed responses. **b,** Direct transfer of the source-fitted ReCA estimator to an observed target support profile. Specificity is the same-minus-other-compound Fisher-z difference. Points show paired mean effects and bars show 95% percentile bootstrap intervals from 10,000 resamples. Target-specific intervals resample compounds; mean intervals resample targets and then compounds. Both experiments combine cell-line and plate-block shifts.



Supplementary Figure S3 | Task-dependent effects of ReCA on activity ranking and annotation retrieval. Bars show paired ReCA–Raw changes in confirmed-activity average precision (AP) and mechanism, target and curated Reactome target-neighbour mean average precision (mAP). The four endpoint-specific populations contain 262, 140, 146 and 202 compounds, respectively. Error bars are pointwise 95% percentile intervals from 10,000 paired compound-bootstrap resamples, using the seed-specific bootstrap procedure described in Supplementary Methods. All four endpoints use the one-support ReCA predictions. Raw and ReCA values and full paired intervals are given in Supplementary Table S3.

Supplementary Table S1 | Measurement aggregation, estimation and independent evaluation

Setting	Comparison	Endpoint	Estimate [95% CI]	n	Note
BBBC047 CP	One measurement per group	Median excess agreement (PCC)	+0.13919 [+0.13554, +0.14260]	18082	
BBBC047 CP	Two measurements per group	Median excess agreement (PCC)	+0.24312 [+0.23786, +0.24839]	18082	
BBBC047 CP	Two per group - one per group	Median paired change in PCC	+0.08612 [+0.08453, +0.08768]	18082	
BBBC047 GE	One measurement per group	Median excess agreement (PCC)	+0.02949 [+0.00034, +0.10547]	38	
BBBC047 GE	Two measurements per group	Median excess agreement (PCC)	+0.06311 [+0.01679, +0.19180]	38	
BBBC047 GE	Two per group - one per group	Median paired change in PCC	+0.03559 [+0.01288, +0.07607]	38	
BBBC047 CP; 1 support	IMR - measurement mean	Repeat agreement (Fisher z)	+0.08046 [+0.07138, +0.08906]	4023	
cpg0004 CP; 1 support	IMR - measurement mean	Repeat agreement (Fisher z)	+0.13279 [+0.11411, +0.15087]	127	Pooled over three runs
BBBC047 CP; 2 support	IMR - measurement mean	Repeat agreement (Fisher z)	+0.04757 [+0.03975, +0.05538]	3982	
cpg0004 CP; 2 support	IMR - measurement mean	Repeat agreement (Fisher z)	+0.06959 [+0.05706, +0.08217]	120	Pooled over three runs
BBBC047 CP; 3 support	IMR - measurement mean	Repeat agreement (Fisher z)	+0.02883 [+0.02089, +0.03657]	3588	
cpg0004 CP; 3 support	IMR - measurement mean	Repeat agreement (Fisher z)	+0.03161 [+0.02018, +0.04268]	120	Pooled over three runs
BBBC047 GE; 1 support	IMR - measurement mean	Repeat agreement (Fisher z)	+0.06976 [+0.06102, +0.07868]	3691	
BBBC047 GE; 2 support	IMR - measurement mean	Repeat agreement (Fisher z)	+0.08762 [+0.07710, +0.09792]	3121	
BBBC047 CP; 1 support	IMR - measurement mean	Specificity (Fisher z)	+0.06088 [+0.05157, +0.07052]	4023	
BBBC047 CP; 1 support	Aggregated-target IMR - measurement mean	Specificity (Fisher z)	+0.07828 [+0.06883, +0.08786]	4023	

Continued on next page

Supplementary Table S1 continued

Setting	Comparison	Endpoint	Estimate [95% CI]	n	Note
BBBC047 CP; 1 support	Empirical-Bayes shrinkage - measurement mean	Specificity (Fisher z)	+0.08588 [+0.07762, +0.09406]	4023	
BBBC047 CP; 2 support	IMR - measurement mean	Specificity (Fisher z)	+0.02397 [+0.01555, +0.03236]	3982	
BBBC047 CP; 2 support	Aggregated-target IMR - measurement mean	Specificity (Fisher z)	+0.03516 [+0.02688, +0.04355]	3982	
BBBC047 CP; 2 support	Empirical-Bayes shrinkage - measurement mean	Specificity (Fisher z)	+0.03965 [+0.03289, +0.04643]	3982	
BBBC047 CP; 3 support	IMR - measurement mean	Specificity (Fisher z)	+0.00646 [-0.00246, +0.01526]	3588	
BBBC047 CP; 3 support	Aggregated-target IMR - measurement mean	Specificity (Fisher z)	+0.00983 [+0.00113, +0.01854]	3588	
BBBC047 CP; 3 support	Empirical-Bayes shrinkage - measurement mean	Specificity (Fisher z)	+0.01117 [+0.00459, +0.01762]	3588	
BBBC047; 1 support	ReCA - best single estimator	Specificity (Fisher z)	+0.00059 [-0.00138, +0.00256]	2423	
BBBC047; 3 support	ReCA - best single estimator	Specificity (Fisher z)	+0.00182 [+0.00100, +0.00264]	2166	
cpg0004-LINCS; 1 support	ReCA - best single estimator	Specificity (Fisher z)	+0.01580 [+0.00964, +0.02184]	189	
cpg0004-LINCS; 3 support	ReCA - best single estimator	Specificity (Fisher z)	+0.01052 [+0.00625, +0.01485]	185	
Across four conditions	ReCA - best single estimator	Specificity (Fisher z)	+0.00718 [+0.00529, +0.00913]	4	Equal-weight mean of four conditions; n counts conditions
BBBC047; 1 support	ReCA - support mean	Specificity (Fisher z)	+0.02915	2423	Point estimate
BBBC047; 3 support	ReCA - support mean	Specificity (Fisher z)	+0.01680	2166	Point estimate
cpg0004-LINCS; 1 support	ReCA - support mean	Specificity (Fisher z)	+0.06934	189	Point estimate
cpg0004-LINCS; 3 support	ReCA - support mean	Specificity (Fisher z)	+0.05120	185	Point estimate
BBBC047; 1 support	ReCA - PCA	Specificity (Fisher z)	+0.02347	2423	Point estimate
BBBC047; 3 support	ReCA - PCA	Specificity (Fisher z)	+0.00898	2166	Point estimate

Continued on next page

Supplementary Table S1 continued

Setting	Comparison	Endpoint	Estimate [95% CI]	n	Note
cpg0004-LINCS; 1 support	ReCA - PCA	Specificity (Fisher z)	+0.04244	189	Point estimate
cpg0004-LINCS; 3 support	ReCA - PCA	Specificity (Fisher z)	+0.02693	185	Point estimate

Support budgets count measurements. Each budget has its own eligible population. The paired median change in the aggregation analysis is computed within compounds, so it differs from the difference between the marginal medians.

Specificity contrasts use prediction-specific foreign scores. BBBC047 component specificity contrasts use foreign profiles from support plates; Supplementary Table S4 uses foreign profiles from the held-out plate. IMR, independent-measurement regression; ReCA, reproducibility-calibrated aggregation; PCA, principal-component reconstruction. Aggregated-target IMR uses an aggregate of non-support measurements as its target. Empirical-Bayes shrinkage denotes the covariance-based empirical-Bayes estimator.

Intervals are paired compound intervals; for the cpg0004 budget comparisons, scores were averaged within compound across three training runs before resampling. Aggregation intervals use 2,000 bootstrap resamples and other intervals 10,000. ReCA weights and run-specific estimates are in Supplementary Data 1.

Supplementary Table S2 | Cross-modal controls and conditional evidence

Setting	Comparison	Endpoint	Estimate [95% CI]	n	Note
GE -> CP	GE residual update - target-only CP estimate	Repeat agreement (Fisher z)	+0.02150 [+0.01937, +0.02366]	4023	
GE -> CP	Same-compound evidence - shuffled pairing	Repeat agreement (Fisher z)	+0.03742 [+0.03488, +0.03995]	4023	
GE -> CP	Same-compound evidence - matched control	Repeat agreement (Fisher z)	+0.01081 [+0.00903, +0.01261]	3538	
CP -> GE	Conditional update - target-only GE estimate	Repeat agreement (Fisher z)	+0.00937 [+0.00572, +0.01296]	3691	Pooled over three runs
CP -> GE	Same-compound evidence - shuffled pairing	Repeat agreement (Fisher z)	+0.01417 [+0.01002, +0.01832]	3691	Pooled over three runs
CP -> GE	Same-compound evidence - matched control	Repeat agreement (Fisher z)	+0.01139 [+0.00732, +0.01543]	3691	Pooled over three runs

Continued on next page

Supplementary Table S2 continued

Setting	Comparison	Endpoint	Estimate [95% CI]	n	Note
CP -> GE	Joint prediction - target-only GE estimate	Repeat agreement (Fisher z)	+0.00492 [+0.00035, +0.00942]	3691	Pooled over three runs
CP -> GE	Conditional update - joint prediction	Repeat agreement (Fisher z)	+0.00446 [+0.00089, +0.00798]	3691	Pooled over three runs
Sequential CP/GE	GE|CP1	Repeat agreement (Fisher z)	+0.01942 [+0.01728, +0.02161]	3983	
Sequential CP/GE	GE|CP1+CP2	Repeat agreement (Fisher z)	+0.01487 [+0.01286, +0.01683]	3983	
Sequential CP/GE	CP2|CP1	Repeat agreement (Fisher z)	+0.02520 [+0.02175, +0.02862]	3983	
Sequential CP/GE	CP2|CP1+GE	Repeat agreement (Fisher z)	+0.02065 [+0.01742, +0.02379]	3983	
GE -> CP	Same-compound - exact-dose control	Repeat agreement (Fisher z)	+0.01854 [+0.01421, +0.02293]	715	Exact plate and dose matching
BBBC047; 1 support	Prior-augmented IMR - IMR	Repeat agreement (Fisher z)	+0.01714 [+0.01391, +0.02038]	4023	Run 1; prior weight 0.300
BBBC047; 1 support	Prior-augmented IMR - IMR	Repeat agreement (Fisher z)	+0.01517 [+0.01174, +0.01847]	4023	Run 2; prior weight 0.275
BBBC047; 1 support	Prior-augmented IMR - IMR	Repeat agreement (Fisher z)	+0.02782 [+0.02382, +0.03192]	4023	Run 3; prior weight 0.300
cpg0004-LINCS; 1 support	Prior-augmented IMR - IMR	Repeat agreement (Fisher z)	-0.00123 [-0.00301, +0.00049]	121	Run 1; prior weight 0.100
cpg0004-LINCS; 1 support	Prior-augmented IMR - IMR	Repeat agreement (Fisher z)	-0.00233 [-0.00420, -0.00045]	121	Run 2; prior weight 0.100
cpg0004-LINCS; 1 support	Prior-augmented IMR - IMR	Repeat agreement (Fisher z)	-0.00220 [-0.00402, -0.00037]	121	Run 3; prior weight 0.100

Unless stated otherwise, rows use BBBC047. GE, gene expression; CP, Cell Painting. Effects are changes in repeat agreement (Fisher z). Reassigned and matched-control evidence assess compound identity.

CP-to-GE updates condition on the current GE estimate. Pooled intervals average compound-level contrasts across three runs before 10,000 bootstrap resamples.

GE-to-CP controls reassign the residual-evidence vectors. CP1 and CP2 are successive support measurements; X|Y denotes adding X after Y. Sequential comparisons share 3,983 compounds and a held-out CP target. Run-specific estimates are in Supplementary Data 1.

Prior comparisons are reported per run (runs 1, 2 and 3 use seeds 3407, 42 and 2025) with validation-selected prior weights. BBBC047 uses one support and one held-out plate; cpg0004 uses six doses. Intervals use 10,000 and 2,000 compound resamples, respectively.

Supplementary Table S3 | ReCA morphological recovery, downstream performance and transfer boundaries

Endpoint	Raw	ReCA	Paired change [95% CI]	n
Feature-group agreement (Fisher-z PCC ↑; Fig. 4a)				
DNA	0.59804	0.70057	$+0.10253\ [0.08806,\ 0.11681]$	291
RNA	0.73426	0.89338	$+0.15912\ [0.14112,\ 0.17734]$	291
Endoplasmic reticulum	0.64549	0.78598	$+0.14049\ [0.12645,\ 0.15435]$	291
Mitochondria	0.53781	0.63939	$+0.10158\ [0.08384,\ 0.11982]$	291
Actin/Golgi/plasma membrane	0.62415	0.76107	$+0.13692\ [0.12232,\ 0.15145]$	291
Shape	0.37559	0.57317	$+0.19758\ [0.18397,\ 0.21170]$	291
Cross-channel	0.63367	0.78360	$+0.14994\ [0.13495,\ 0.16468]$	291
Seven-group mean	0.59272	0.73388	$+0.14117\ [0.13121,\ 0.15109]$	291
Dominant-feature recovery and dose concordance (↑; Fig. 4b,c)				
Top-25 overlap	0.26121	0.31962	$+0.05841\ [0.05333,\ 0.06352]$	291
Top-25 direction agreement	0.79016	0.86319	$+0.07302\ [0.06787,\ 0.07836]$	291
Dose concordance (Spearman ρ)	0.32673	0.37169	$+0.04496\ [0.03558,\ 0.05419]$	289
Downstream performance (↑; Supplementary Fig. S3)				
Confirmed activity (AP)	0.11260	0.17909	$+0.06649\ [0.02120,\ 0.12942]$	262
Mechanism retrieval (mAP)	0.04997	0.04977	$-0.00020\ [-0.00115,\ 0.00071]$	140
Target retrieval (mAP)	0.05018	0.04984	$-0.00034\ [-0.00114,\ 0.00051]$	146
Reactome target-neighbour retrieval (mAP)	0.11952	0.11701	$-0.00251\ [-0.00578,\ 0.00073]$	202

All rows use one-support ReCA predictions on cpg0004-LINCS. n, compounds; ↑, higher is better; AP, average precision; mAP, mean average precision. Changes are ReCA minus Raw; intervals are pointwise 95% paired compound-bootstrap intervals (10,000 resamples). Retrieval arm means average queries, whereas paired contrasts first average within compound; their subtraction can therefore differ slightly. Aggregation and seed-specific procedures are detailed in Supplementary Methods.

Supplementary Table S3 continued | Cross-environment transfer

Comparison	Target	Specificity change [95% CI]	n
Predictor-target replacement: seen compounds			
Calibrated − observed training targets	A549	$+0.00492\ [0.00352,\ 0.00631]$	183
	K562	$+0.00907\ [0.00702,\ 0.01137]$	183
	MCF7	$+0.00334\ [0.00067,\ 0.00597]$	183
	Mean	$+0.00578\ [0.00304,\ 0.00887]$	3 lines
Predictor-target replacement: unseen compounds (Supplementary Fig. S2a)			
Calibrated − observed training targets	A549	$+0.00626\ [0.00313,\ 0.00938]$	36
	K562	$+0.00556\ [0.00317,\ 0.00789]$	36
	MCF7	$+0.00255\ [-0.00272,\ 0.00790]$	36
	Mean	$+0.00479\ [0.00167,\ 0.00726]$	3 lines
Direct estimator transfer: unseen compounds (Supplementary Fig. S2b)			
Transferred ReCA − observed support	A549	$-0.00760\ [-0.01447,\ -0.00110]$	36
	K562	$-0.00723\ [-0.01223,\ -0.00260]$	36
	MCF7	$-0.01776\ [-0.03015,\ -0.00763]$	36
	Mean	$-0.01087\ [-0.01896,\ -0.00499]$	3 lines

These sci-Plex3 analyses use ridge-based ReCA targets (Supplementary Methods). Specificity is the same-minus-other-compound Fisher-z agreement. The same 183 seen or 36 unseen compounds recur across all three target cell lines; n is not additive across lines. Target-specific intervals use 10,000 paired compound-bootstrap resamples; mean intervals resample target cell lines and then compounds. Predictor-target replacement and direct estimator transfer are separate interventions.

Supplementary Table S4 | Classical and repeat-aware BBBC047 response-estimation benchmark

Method	1 support: E [95% CI]	2 supports: E [95% CI]	3 supports: E [95% CI]
Mean	0.18947 [0.17697, 0.20255]	0.22936 [0.21548, 0.24308]	0.26921 [0.25447, 0.28443]
Median	0.18947 [0.17648, 0.20236]	0.22936 [0.21562, 0.24320]	0.25990 [0.24568, 0.27406]
MODZ	0.18947 [0.17656, 0.20250]	0.22936 [0.21572, 0.24307]	0.26419 [0.24903, 0.27969]
PCA	0.20049 [0.18618, 0.21476]	0.23796 [0.22370, 0.25282]	0.27776 [0.26214, 0.29336]
Ridge	0.20632 [0.19371, 0.21868]	0.24029 [0.22701, 0.25326]	0.27280 [0.25876, 0.28691]
Noise2Self	0.20364 [0.19032, 0.21709]	0.23356 [0.21970, 0.24732]	0.27328 [0.25865, 0.28804]
IMR	0.18968 [0.17764, 0.20163]	0.22170 [0.20911, 0.23471]	0.25000 [0.23678, 0.26360]
Aggregated-target IMR	0.21019 [0.19781, 0.22262]	0.23541 [0.22231, 0.24847]	0.24741 [0.23395, 0.26107]
Empirical-Bayes shrinkage	0.21616 [0.20333, 0.22883]	0.25070 [0.23744, 0.26426]	0.28046 [0.26598, 0.29478]
ReCA	0.22117 [0.20820, 0.23406]	0.25416 [0.24069, 0.26803]	0.28208 [0.26762, 0.29663]

All methods received the same R physical 775-dimensional support profiles and were evaluated against the same held-out profile and 32 condition-matched foreign profiles. Perturbation specificity is $E = z_{\mathrm{same}} - z_{\mathrm{foreign}}$. Eligible compound counts were 3,149, 3,124 and 2,832 at one, two and three supports, respectively. Values are compound-level means with 95% percentile intervals from 10,000 compound-bootstrap resamples after averaging eligible dose conditions within compound.

Mean, Median, MODZ, PCA, Ridge and Noise2Self are external comparators. IMR, aggregated-target IMR and empirical-Bayes shrinkage are the component estimators of ReCA.

Supplementary Table S5 | Generic neural backbones and model-size sensitivity

Displayed family	Configuration	Parameters	1 support: E	2 supports: E	3 supports: E
MLP	Capacity-matched control; 3 blocks, width 768	2,373,127	0.20667	0.22973	0.25429
ResNet	Compact; 3 blocks, width 384	2,372,359	0.19929	0.22895	0.25850
ResNet	Larger; 3 blocks, width 512	3,949,319	0.20121	0.23304	0.26481
FT-Transformer	Compact; four blocks	2,704,303	0.19427	0.21876	0.22649
FT-Transformer	Larger; five blocks	4,856,555	0.19706	0.21030	0.23857
TabM	Compact; $k = 4$	2,530,872	0.20169	0.22813	0.25812
TabM	Larger; $k = 32$	13,799,360	0.20877	0.23763	0.26826
ReCA	Repeat-aware estimator	–	0.22117	0.25416	0.28208

Paired comparison	1 support: ΔE [95% CI]	2 supports: ΔE [95% CI]	3 supports: ΔE [95% CI]
Larger ResNet − compact ResNet	+0.00192 [+0.00001, +0.00386]	+0.00410 [+0.00210, +0.00612]	+0.00631 [+0.00215, +0.01042]
Larger FT-Transformer − compact FT-Transformer	+0.00279 [+0.00039, +0.00521]	−0.00846 [−0.01130, −0.00557]	+0.01208 [+0.00796, +0.01620]
Larger TabM − compact TabM	+0.00708 [+0.00408, +0.01005]	+0.00950 [+0.00689, +0.01224]	+0.01014 [+0.00655, +0.01357]
ReCA − larger ResNet	+0.01996 [+0.01547, +0.02443]	+0.02112 [+0.01638, +0.02592]	+0.01728 [+0.01122, +0.02318]
ReCA − larger FT-Transformer	+0.02411 [+0.01769, +0.03056]	+0.04387 [+0.03757, +0.05029]	+0.04351 [+0.03634, +0.05057]
ReCA − larger TabM	+0.01240 [+0.00766, +0.01718]	+0.01653 [+0.01188, +0.02117]	+0.01382 [+0.00896, +0.01855]

Every arm mapped the arithmetic mean of the same R physical 775-dimensional support profiles to the same disjoint 775-dimensional held-out profile. The compact configurations are capacity matched; the larger configurations test sensitivity to model size and differ in parameter count across architectures. Figure 2d shows the MLP control and the larger ResNet, FT-Transformer and TabM configurations.

All neural values are the means of predictions from seeds 3407, 42 and 2025. Epochs were selected using compound-disjoint splits of the training compounds, after which each model was refitted on all training compounds for the selected number of epochs. Conditions were averaged within compound before 10,000-draw paired compound bootstrap. The common populations contained 3,149, 3,124 and 2,832 compounds at one, two and three supports.

Supplementary Table S6 | sci-Plex3 prediction with ReCA-derived training targets

Training target	PCC ↑	△PCC ↑	E ↑
Within cell line, unseen drugs (36 drugs per target cell line)			
Raw (observed responses)	0.5816	0.6278	0.3603
ReCA	0.4691 −0.1125 [−0.1155, −0.1095]	0.4410 −0.1867 [−0.1902, −0.1833]	0.2507 −0.1096 [−0.1119, −0.1072]
ReCA+Res	0.5838 +0.0022 [+0.0013, +0.0030]	0.6312 +0.0034 [+0.0026, +0.0042]	0.3590 −0.0013 [−0.0019, −0.0006]
Across cell lines, unseen drugs (36 drugs per target cell line)			
Raw (observed responses)	0.8394	0.0068	0.0027
ReCA	0.8456 +0.0062 [+0.0061, +0.0064]	0.0096 +0.0028 [+0.0010, +0.0045]	0.0007 −0.0020 [−0.0032, −0.0007]
ReCA+Res	0.8360 −0.0034 [−0.0035, −0.0033]	0.0036 −0.0033 [−0.0043, −0.0022]	0.0022 −0.0005 [−0.0012, +0.0003]
Across cell lines, seen drugs (183 drugs per target cell line)			
Raw (observed responses)	0.8338	0.0089	0.0049
ReCA	0.8409 +0.0071 [+0.0070, +0.0072]	0.0110 +0.0021 [+0.0012, +0.0030]	0.0023 −0.0026 [−0.0033, −0.0019]
ReCA+Res	0.8310 −0.0028 [−0.0029, −0.0027]	0.0056 −0.0033 [−0.0038, −0.0028]	0.0042 −0.0007 [−0.0011, −0.0004]

Values are means over the three held-out target cell lines (A549, K562 and MCF7), each weighted equally. For ReCA and ReCA+Res, the second line gives the paired difference from Raw with its pointwise 95% interval (10,000 paired drug resamples within each target cell line). Raw trains on observed responses; ReCA trains on direct ReCA estimates; ReCA+Res predicts the ReCA estimate and the observed-minus-ReCA residual in separate branches and sums them, as for BBBC047 in Fig. 5. All arms were capacity-matched and evaluated against the same observed target profiles. PCC, full-profile correlation after adding the matched Vehicle baseline; △PCC, correlation of control-subtracted responses; E, perturbation specificity (same-drug minus other-drug Fisher-z agreement). With two measurements per condition, the sci-Plex3 ReCA adaptation reduces aggregated-target IMR to IMR and uses diagonal shrinkage (Supplementary Methods). Seen-drug transfer uses drugs that were also in the source training set.

Supplementary Methods

The cpg0004 feature representation

The cpg0004 preparation started from the normalized, variable-selected replicate-level CP table supplied with the LINCS Pilot 1 data [1]. It retained treatment profiles at the six standard dose levels, excluded two plate-wide reference compounds, and collapsed technical wells by the feature-wise median within compound–dose–plate. A same-plate control median defined the response baseline. The resulting table contained 45,621 treatment profiles on 136 plates and 1,568 ordinary compounds.

Features with at most 20% missing values across treatment and control profiles were retained. Imputation used global control medians; features absent from all controls used the pooled treatment/control median, and entirely missing features were discarded. These preprocessing quantities were computed once on the full table, before compound splitting; the model standardizers used in the experiments were fitted on training data only. The resulting 242-coordinate matrix retained one batch identifier, which was excluded from the 241-feature biological analysis. Image segmentation and feature normalization were those of the data providers.

Measurement noise and independent supervision

Let an observed response be $\boldsymbol{x}_{ir} = \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_{ir}$, where $\boldsymbol{\theta}_i$ is the condition-specific $\boldsymbol{\mu}_{c,m}$ of the main Methods with indices abbreviated, and r indexes independent measurements. Suppose that a held-out measurement error has conditional mean zero and is conditionally independent of the support observations given $\boldsymbol{\theta}_i$. For any estimator $f(\boldsymbol{x}_{i,S})$,

$$\mathbb{E}\big[\|f(\boldsymbol{x}_{i,S}) - \boldsymbol{x}_{ih}\|^2\big] = \mathbb{E}\big[\|f(\boldsymbol{x}_{i,S}) - \boldsymbol{\theta}_i\|^2\big] + \mathbb{E}\big[\|\boldsymbol{\epsilon}_{ih}\|^2\big]\,. \tag{S1}$$

The cross term vanishes under these assumptions, so the minimizers of the two risks coincide, following the independent-noisy-target argument of Noise2Noise [2]. Averaging independent held-out labels reduces their variance without changing their conditional mean. This argument motivates IMR and aggregated-target IMR. Shared plate effects, dose mixing or systematic errors violate its assumptions, so the empirical design separates measurements physically and uses matched controls. The identity holds for squared error, the fitting objective; agreement (PCC) is an evaluation measure.

Covariance-based empirical-Bayes shrinkage

Training conditions with at least two plate measurements were included. For condition i, let n_i denote its number of plates, $\boldsymbol{m}_i$ their mean, and N_i their sample covariance with denominator $n_i - 1$. Conditions received equal weight. We estimated

$$\widehat{\boldsymbol{\mu}} = \frac{1}{M}\sum_i \boldsymbol{m}_i, \tag{S2}$$

$$\widehat{\Sigma}_n = \frac{1}{M}\sum_i N_i, \tag{S3}$$

$$\widehat{\Sigma}_s = \frac{1}{M}\sum_i \big[(\boldsymbol{m}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{m}_i - \widehat{\boldsymbol{\mu}})^\top - N_i/n_i\big]\,. \tag{S4}$$

A diagonal term $10^{-6}\,\mathrm{tr}(\widehat{\Sigma}_n)/d$ stabilized the noise covariance. Let $LL^\top$ be its Cholesky factorization. We symmetrized $L^{-1}\widehat{\Sigma}_s L^{-\top}$, decomposed it as $V\Lambda V^\top$, and replaced negative eigenvalues by zero. For a support mean $\overline{\boldsymbol{x}}_S$ based on R plates, the estimate was

$$\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\mu}} + LV\,\mathrm{diag}\left(\frac{\lambda_j^+}{\lambda_j^+ + 1/R}\right) V^\top L^{-1}(\overline{\boldsymbol{x}}_S - \widehat{\boldsymbol{\mu}}). \tag{S5}$$

Under a Gaussian working model with whitened signal variance λ_j^+ and independent noise variance $1/R$, the factor is the conditional-mean coefficient. Learning a shared covariance structure follows the broader multivariate empirical Bayes principle [3]; here the estimator uses the repeated-measurement covariance definitions above. The rule is used as a point estimator; no posterior intervals are computed.

Fitting partitions and numerical settings

The CP budget benchmark used the same support and reference assignments, made from identifiers alone, across seeds 42, 3407 and 2025. BBBC047 neural estimates used five compound-level fits for out-of-fold predictions where required. Its checkpoints were selected within the training compounds. The GE and cpg0004 budget regressors used their validation partitions for checkpoint selection. Standard deviations smaller than 10^{-6} were floored before standardization. Model outputs were returned to response coordinates before scoring.

For the independent ReCA experiment, a compound partition drawn with seed 3407 allocated 60% of original training compounds to fitting, 20% to calibration and 20% to evaluation. The IMR and aggregated-target IMR fitting subset was internally split by compound for epoch selection, then both models were refitted on the whole fitting subset. Input statistics and direct-reference target statistics were fitted on this subset. Aggregated-target IMR labels included all eligible non-support plate means, including the selected direct reference, but excluded every support plate.

ReCA weights minimized unstandardized response-coordinate MSE over the three-dimensional simplex. The minimum was found by comparing the vertex, edge and feasible interior solutions, followed by the fixed shrinkage towards equal weights described in Methods, $\boldsymbol{w} = (1-\gamma)\mathbf{1}/3 + \gamma\hat{\boldsymbol{w}}$ with $\gamma = n/(n+3)$ for n calibration compounds. PCA candidate ranks were 16, 32, 64, 128, 256 and 512, clipped to feasible dimensions; rank selection used calibration specificity. Four criteria on point estimates were specified before evaluation: no ReCA-minus-best-single effect below -0.01, at least three non-negative effects, a positive equal-condition macro-average, and positive effects against support means and PCA in all four conditions.

The four independent-ReCA evaluation conditions contained 2,423 and 2,166 BBBC047 compounds and 189 and 185 cpg0004-LINCS compounds at one and three supports, respectively. The strongest individual estimator in each condition was identified from the candidate predictions without refitting.

Foreign-reference construction and score aggregation

Support eligibility and descriptive aggregation.

For one, two and three CP support measurements, the matched-budget IMR comparisons included 4,023, 3,982 and 3,588 BBBC047 compounds and 127, 120 and 120 cpg0004-LINCS compounds, respectively. The cpg0004 comparisons used matched-control subsets shared across three runs. BBBC047 GE comparisons included 3,691 and 3,121 compounds at one and two supports. The dataset-wide aggregation analysis compared single-plate agreement with agreement between two disjoint two-plate means. It required four physical plates per compound and sampled up to 20 plate quartets. The analogous GE analysis included the 38 compounds with four eligible measurements, some of which combined doses.

In BBBC047 aggregation, each comparison used four distinct plate slots. Foreign comparisons preserved these slots, used one foreign compound for each two-plate side, and required different compounds between sides. Sixteen null draws and up to 20 candidate quartets per compound were used. Plate- or quartet-level statistics were reduced to one compound-level value before calculating population medians and their uncertainty.

The fixed-null repeat-budget analyses constructed foreign measurement pairs rather than comparing each prediction with foreign measurements. BBBC047 matching preserved the selected support and reference plate slots. cpg0004 matching preserved compound dose, support/reference plates and well row. Technical wells within a plate were not counted as independent repeats. Thirty-two matched-null draws were used in the standard budget experiments. A scalar PCC was calculated for each eligible pair and transformed before averaging within compound. The fixed null was reused for observed means and model estimates. Thus the difference of their corrected scores is exactly the difference of their same-reference transformed agreements.

The direct estimator decomposition and independent ReCA experiments instead compared each prediction with its selected same-compound and foreign-compound reference profiles. In the BBBC047 direct decomposition, foreign profiles came from the selected support plates and were averaged across these plates for multiple supports. In the BBBC047 same-space benchmark, foreign profiles came from the held-out plate. Their foreign correction therefore depends on the method prediction. In sci-Plex3, each foreign correlation is likewise transformed before averaging. Source-data tables give the scoring definition and the eligible sample count for each endpoint.

Same-space benchmarks and representation retrieval

The BBBC047 same-space benchmark preserved the compound-disjoint train, validation and test partitions and physical $(\mathrm{compound}, \mathrm{dose}, \mathrm{plate})$ identity. Support plates, one disjoint held-out plate and 32 same-dose, different-compound foreign profiles were assigned with seed 3407. Technical wells were aggregated only within a physical profile. Mean, Median, MODZ, PCA reconstruction, multi-output Ridge, Noise2Self-style denoising, IMR, aggregated-target IMR, empirical-Bayes shrinkage and ReCA all mapped R support profiles in the same 775-dimensional space to one estimate. Hyperparameters and ReCA weights were selected using training or validation compounds only; the benchmark ReCA weights minimized validation mean squared error over the simplex, without shrinkage towards

equal weights. PCA was selected on validation data as the external comparator for the prespecified macro contrast, the equal-weight mean of ReCA − PCA over the three budgets: $+0.0137$ ($+0.0207$, $+0.0162$ and $+0.0043$ at one, two and three supports; point estimates from Supplementary Table S4). The primary endpoint was prediction-specific perturbation specificity E, defined in the main Methods. Scores were averaged over eligible dose conditions within compound, and intervals used 10,000 paired compound-bootstrap samples.

The neural architecture stress test used the same support, held-out and foreign assignments, support means, held-out targets and 775-dimensional output. A capacity-matched track compared an MLP, ResNet, FT-Transformer and TabM; a separate size-sensitivity track used larger reference configurations for ResNet, FT-Transformer and TabM [4, 5]. The main figure displays the MLP control and these larger configurations. Their parameter counts were 2,373,127, 3,949,319, 4,856,555 and 13,799,360, respectively. Budgets were trained independently, and each reported prediction was the mean across seeds 3407, 42 and 2025. Within each seed, a compound-disjoint split of the training compounds selected the epoch by standardized held-response mean squared error. Feature and target standardizers used inner-training compounds only, after which the model was refitted on all training compounds for the selected epoch count.

cpDistiller-C was evaluated by representation retrieval because it maps 1,783-dimensional source-well CellProfiler features to a 50-dimensional latent representation. The published cpDistiller-C architecture and exponential-moving-average checkpoint were fitted on 51,992 treatment and 24,996 mock-control wells from training compounds only. Test wells were embedded independently and then aggregated to physical $(\mathrm{compound}, \mathrm{dose}, \mathrm{plate})$ profiles. For budget R, the query was the mean of R support embeddings; the gallery contained one held-out same-compound, same-dose, different-plate profile and the same 32 foreign identities used for the corresponding Mean and ReCA arms. With one positive, per-condition average precision equalled reciprocal held-profile rank. mAP@33 was averaged across dose conditions within compound and then across compounds. All arms and paired contrasts used 10,000 compound-bootstrap samples.

Cross-modal controls and joint prediction

The GE-to-CP residual experiment reused fixed CP estimates and virtual priors. Its shuffled control reassigned residual vectors between compounds; its foreign control matched available GE plate-set, dose and response-magnitude metadata. An analysis with exact plate and dose matching used 715 compounds. These analyses measure the increment that GE adds to fixed CP estimates.

The CP-to-GE residual input contained the 977-dimensional GE estimate and 775-dimensional observed CP aggregate. Both input statistics and residual-target statistics were fitted on training data. Five-fold compound-out-of-fold GE estimates constructed its training residual labels. Mapper checkpoint selection used an inner training-compound split. Validation selected β from $\{0, 0.025, 0.05, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, 1\}$, with ties favouring the smaller value. Validation selected a positive weight in all three seeds. Shuffled CP reassigned profiles across compounds without self-pairs. Foreign CP donors were selected using available acquisition-count, dose and profile-magnitude metadata. All control arms reused the same target-only estimate and trained mapper.

Direct fusion concatenated measured GE support, rather than its denoised estimate, with observed CP and predicted the independent GE reference. Its architecture was 1752–256–32–256–977, with GELU activations, training-only input/target standardization, AdamW at learning rate 3×10^{-4} and weight decay 10^{-5}, batch size 256 and up to 80 epochs. Validation-reference MSE selected its checkpoint. The paired comparison with the conditional mapper used 3,691 common compounds with shared target-only and null scores.

Virtual priors and GE-informed CP predictors

The virtual CP prior in the BBBC047 support-estimation experiments used a 775-dimensional baseline CP profile and a 2048-bit radius-two Morgan fingerprint. Its architecture was 2823–512–128–775 with GELU hidden activations. It predicted the measured aggregate CP response, with no treated CP or GE input. Baseline and target standardizers were fitted using training compounds, and AdamW used learning rate 3×10^{-4}, weight decay 10^{-5}, batch size 256 and a maximum of 40 epochs. The checkpoint subset was a fixed subset of about 10% of the training compounds; the best checkpoint was retained without a subsequent refit including those checkpoint compounds. Where training-compound prior predictions were required, the prior was cross-fitted over five compound folds, with every checkpoint subset inside the remaining folds. The current estimate was $P_0 = (1-\lambda)\widehat{\boldsymbol{y}}_{\mathrm{IMR}} + \lambda V$, with λ selected by validation agreement over 0 to 0.30 in increments of 0.025.

The BBBC047 GE-informed CP predictor was a deterministic baseline-conditioned multimodal network. Separate CP-baseline, GE-baseline and molecular encoders each produced embeddings with 512 coordinates. Their concatenation formed a 1536-dimensional representation; CP and GE decoders used 768-unit hidden layers. An additional encoder mapped observed 977-dimensional GE responses to 512 units. A dedicated CP residual branch took its concatenation with the 1536-dimensional base representation and used a 768-unit hidden layer before the 775-dimensional CP output. This branch did not feed the GE decoder. Encoder and decoder hidden layers used batch normalization, LeakyReLU with slope 0.1 and dropout 0.3.

Training used the standard compound split, seeds 3407, 42 and 2025 and 40 epochs. The baseline training configuration used Adam, learning rate 10^{-3}, weight decay 10^{-4} and batch size 1,000. The loss summed CP and GE smooth-L1 response-difference losses with transition parameter 1 and auxiliary target-profile MSE weighted by 0.05. An additional CP direction term, weighted by 0.1, compared the centered unit prediction difference with the training plate-consensus direction. Validation selected the checkpoint using the reconstruction objective without the direction term. Predictions were converted back from model coordinates before evaluating raw CP response agreement. The residual used in the update was the full GE-informed CP response minus the separately fitted virtual prior.

For cpg0004, the prior additionally encoded six dose categories: ECFP4, the 242-dimensional held-plate control baseline and six dose indicators formed a 2296–512–128–242 GELU network. The prior was trained for 60 epochs with batches of 256 and AdamW with learning rate 3×10^{-4} and weight decay 10^{-5}. Validation MSE selected its checkpoint and the fixed IMR–prior combination weight. The GE-informed predictor concatenated 993 observed GE coordinates with the same structure, baseline and dose inputs, giving a 3289–512–128–242 network with the same optimizer family. The update was $P_0 + \beta(G - V)$; validation MSE selected β on $[0, 0.5]$ in steps of 0.05. Its GE-shuffling and foreign-donor

controls retained the query compound, dose and baseline. These models are used in the prior and cross-modal-update analyses; Figure 4 and Supplementary Fig. S3 use the one-support ReCA predictions described below.

Morphological features, dose organization and retrieval

The one-support cpg0004-LINCS ReCA model combined IMR, aggregated-target IMR and empirical-Bayes shrinkage with weights $(0.1128535, 0.2115124, 0.6756341)$. Conditions with exactly five physical profiles were eligible for the two support rotations. This yielded 1,743 conditions and 3,486 physical support views per seed; 130 conditions without a four-profile reference were excluded before prediction. Candidate fitting, checkpoint selection and weight calibration used train or validation compounds only.

The biological feature groups contained 30 DNA, 21 RNA, 33 endoplasmic-reticulum, 17 mitochondrial, 31 actin/Golgi/plasma-membrane, 66 shape and 43 cross-channel features. Robust standardization used the median and 1.4826 times the median absolute deviation of training responses, with training standard-deviation and unit fallbacks for degenerate features. Top features were selected by absolute standardized magnitude, with feature index breaking ties. Direction agreement compared the signs of predicted and reference standardized values at the reference-defined indices.

Figure 4 uses one-support ReCA predictions on two physical support rotations. Each support is compared with the mean of the other four physical measurements of the same compound–dose condition. Feature-group PCCs are Fisher-transformed before aggregation; these are direct reference-agreement scores, not foreign-null-subtracted specificity scores. Top-25 overlap and direction are likewise scored per physical support view. The resulting scores are averaged within compound across doses, rotations and seeds, yielding 291 compounds for both feature-group and top-25 analyses. For dose concordance, all 15 pairs among six doses define correlation distances within each rotation. Spearman correlations between estimated and reference distances are averaged equally across seven groups, then across rotations and seeds within compound; 289 compounds are eligible. Neither ReCA profiles nor Raw profiles are averaged across the two support rotations before scoring.

Confirmed activity used one support measurement and two disjoint pairs of confirmation plates. The confirmation-pair reproducibility score and robustly standardized four-repeat effect norm both had to exceed their validation matched-null 95th percentiles. Prediction ranking used the Raw or ReCA effect norm. The AP cohort contained 262 compounds, seven of which had confirmed activity. Mechanism/target galleries used training and validation compounds with at least three repeats per dose; test queries required five repeats and the two available support rotations. Similarity was PCC, positives shared curated labels, and eligible gallery classes had at least five distinct compounds. The fixed mechanism, target and Reactome query populations contained 140, 146 and 202 compounds, respectively. Reactome retrieval selected 25 neighbours, formed the union of their curated target genes and used hypergeometric enrichment with a gallery-derived gene background. The Reactome 2022 gene-set collection contained 1,818 terms [6]. Benjamini–Hochberg correction was applied within each query [7].

sci-Plex response preparation and source-only calibration

The sci-Plex3 analysis used 24-hour single-compound pseudobulks with at least 50 cells and two replicate identifiers. Counts were converted to $\log(1+10^6 x/\sum x)$ before subtracting the matched Vehicle profile. The feature mapping retained 110,983 annotated features. For each held-out target cell line, 2,000 variable features were selected from unique source Vehicle groups. Variance estimates from the two source cell lines received equal weight, independently of their numbers of treated conditions.

For every compound–dose condition, one repeat was assigned as support by a fixed pseudorandom rule and the other as reference. This direction was shared between predictor arms. ECFP4 fingerprints used radius two and 2,048 bits. The query input included the Vehicle baseline corresponding to the held/query repeat, plus standardized log-dose. Input and target standardizers were identical between the two arms and derived from the initial source training set; target standardization used observed labels in both arms.

The source label estimator was a simplified ReCA adaptation. Within each drug-out-of-fold training split and source cell line, standardized repeat-to-repeat ridge regression used $\alpha = 10$ and both repeat directions. With only two repeats, the aggregated-target IMR prediction equalled IMR. The third candidate was diagonal shrinkage of each response coordinate towards zero, using the between-condition variance of two-repeat means divided by that variance plus mean within-condition variance and 10^{-8}. This diagonal rule replaces the full-covariance estimator above. Weights were selected on a simplex grid with increments of 0.05 by source calibration specificity. Every held drug fold was excluded from regression, normalization, shrinkage, calibration and foreign donors. During final all-source refitting, the fitting and calibration pools could overlap, and both excluded the held drug.

The predictor had GELU hidden activations, a maximum of 40 epochs, batch size 256 and the AdamW settings used above [8]. Each arm and seed selected its epoch by MSE against the same source-validation raw held-out responses, then refitted using all eligible source compounds for that fixed epoch count. Previously seen-compound transfer used a source-only 80/20 checkpoint split before refitting all 183 compounds. Unseen-compound transfer used 111 base and 36 calibration compounds and evaluated the remaining 36 compounds, comprising 115 compound–dose conditions. Target inference used 20 distinct same-dose, other-drug reference profiles, with identical donor identities between arms.

Predictor implementation and target construction

The BBBC047 predictor comparison (main-manuscript Fig. 5) and the sci-Plex3 comparison with MLP-based source-response estimates (Supplementary Table S6) used seeds 3407, 42, 2025, 1337 and 7331.

For BBBC047, one compound–dose condition and exactly $k = 2, 3, 4$ distinct physical plates were selected per compound. Unselected treatment repeats were excluded from response-estimator fitting, target construction, downstream predictor training, validation and test evaluation. The one-support response-estimation stage used cyclic directed pairs of selected single-repeat profiles, retaining a one-support reference at every budget; five-fold compound-out-of-fold fitting excluded the recipient compound when generating training targets. ReCA was applied to each selected view before averaging. The Raw predictor used the selected-repeat mean as its training target. The ReCA+Res predictor used the mean of

calibrated views and the raw-minus-calibrated residual as separate branch targets, then summed the two predicted components. Both predictors contained 1,612,558 parameters in two branches with hidden widths $(256, 80)$. Inputs comprised only molecular fingerprints and the control profile. Training used AdamW, learning rate 3×10^{-4}, weight decay 10^{-5}, batch size 256 and at most 40 epochs. The common unseen-compound test target was the selected-repeat raw mean.

For sci-Plex3, source-cell response estimation used a fixed 2000–512–128–512–2000 GELU MLP with independently fitted normalization and calibration. ReCA training targets were generated out of fold by drug, then averaged across physical repeats and equally across the source cell lines. Raw and ReCA predictor configurations each averaged two independently trained $(256, 80)$ branches, giving 2,397,264 parameters per configuration. Source-only selection used 111 fitting and 36 validation drugs, with branch-target validation MSE selecting epochs. Fixed-epoch refits used 147 source drugs for the 36-drug unseen evaluation and all 183 source drugs for seen-drug transfer. Seen-drug evaluations reused these drugs in source cell lines and were analysed separately from unseen-drug transfer.

Paired intervals

For Figure 4, scores are averaged across seeds within compound before 10,000 paired compound-bootstrap resamples. Supplementary Fig. S3 uses a seed-specific procedure: each seed has its own compound resamples, and the resulting bootstrap effects are averaged across seeds before taking the 2.5th and 97.5th percentiles. Activity resampling carries all eligible queries of each sampled compound and recalculates AP; resamples without positive labels are omitted. Retrieval contrasts average query-level differences within compound before resampling. Intervals are pointwise, without correction across endpoints.

Training seeds are model realizations. In pooled repeat-budget, direct estimator-decomposition and cross-modal comparisons, scalar scores were averaged across the three seeds within compound before the paired bootstrap. The independent ReCA experiment fitted one candidate set per dataset–budget condition, and prior comparisons are reported per seed; the cpg0004 prior comparisons used 2,000 resamples. The independent ReCA macro-bootstrap sampled compound-level differences with replacement within each of the four conditions and averaged the four means. Compounds shared between the one- and three-support budgets were resampled independently in each, so the interval treats the four conditions as independent. For sci-Plex3, prediction vectors were averaged across seeds, then scored, then averaged across doses within drug. Its macro-interval resampled the three target cell lines and then drugs within each sampled line.

Paired confidence intervals for three cpg0004 support-budget comparisons and five GE-target contrasts were computed from compound-level scores using 10,000 bootstrap resamples [9]. The GE-target comparisons shared bootstrap draws, including the contrast appearing in two figure panels. Compounds were excluded from an endpoint when they failed its acquisition, annotation or finite-score criteria.

Supplementary data files

Supplementary Data 1 contains figure source data, and Supplementary Data 2 contains table source data.

References

[1] Way, G. P. *et al.* Morphology and gene expression profiling provide complementary information for mapping cell state. *Cell Systems* **13**, 911–923.e9 (2022). URL https://doi.org/10.1016/j.cels.2022.10.001.

[2] Lehtinen, J. *et al.* Noise2Noise: Learning Image Restoration without Clean Data. In Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research 80, 2965–2974 (2018). URL https://proceedings.mlr.press/v80/lehtinen18a.html.

[3] Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics* **51**, 187–195 (2019). URL https://doi.org/10.1038/s41588-018-0268-8.

[4] Gorishniy, Y., Rubachev, I., Khrulkov, V. & Babenko, A. Revisiting Deep Learning Models for Tabular Data. *Advances in Neural Information Processing Systems* **34**, 18932–18943 (2021). URL https://proceedings.neurips.cc/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html.

[5] Gorishniy, Y., Kotelnikov, A. & Babenko, A. TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling. *The Thirteenth International Conference on Learning Representations* (2025). URL https://openreview.net/forum?id=Sd4wYYOhmY.

[6] Gillespie, M. *et al.* The Reactome pathway knowledgebase 2022. *Nucleic Acids Research* **50**, D687–D692 (2022). URL https://doi.org/10.1093/nar/gkab1028.

[7] Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995). URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[8] Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. International Conference on Learning Representations (2019). URL https://openreview.net/forum?id=Bkg6RiCqY7.

[9] Efron, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7**, 1–26 (1979). URL https://doi.org/10.1214/aos/1176344552.