DACL: PREDICTING RESPONSES TO PAIRS OF UNSEEN GENES BY REUSING SINGLE-PERTURBATION SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Combinatorial CRISPR screens measure a small fraction of possible gene pairs, so models must predict responses to gene pairs never perturbed in training. In this strict both-unseen setting, the only information about the test genes is a response-free descriptor such as Gene Ontology annotations. Learned structure helps only if these descriptors can retrieve it. We introduce Deployment-Addressable Constituent Learning (DACL), which learns one map from a gene's descriptor to its perturbation effect. Single perturbations supervise the map directly, while pairs supervise the sum of two predicted effects. An unseen pair is predicted by that sum. On the three official Norman Perturb-seq splits, DACL reaches a Pearson-Δ of 0.401 on pairs of unseen genes, against 0.337 for biolord and 0.191 for GEARS. Over 70 further splits it outperforms biolord and direct pair models given the same Gene Ontology descriptors, and the advantage stays positive, though unresolved at the gene level, with language-model gene embeddings. On a genetic-interaction screen, matched controls show that single perturbations help through the additive path, not a larger function class. Direct pair models remain preferable on some screens with strong pair-specific structure. In two prospective evaluations, training data alone always selected the direct model, which had lower normalized pair error in all 45 panels.

1 INTRODUCTION

Combinatorial perturbation screens map genetic interactions and help prioritize interventions that reshape cellular state, but the number of possible gene pairs far exceeds what can be measured. Perturb-seq and related CRISPR platforms read out high-dimensional responses at single-cell resolution (Dixit et al., 2016; Replogle et al., 2022), including responses to gene pairs (Norman et al., 2019; Wessels et al., 2023). Yet even the largest screens cover a small fraction of the pair space. Computational models are therefore asked to extrapolate. Latent response models such as scGen and CPA shift or compose perturbation embeddings, and GEARS propagates effects over a gene-interaction and Gene Ontology graph (Lotfollahi et al., 2019; 2023; Roohani et al., 2024). The biolord model adds constituent effects predicted from biological attributes (Piran et al., 2024), and scGPT and GeneCompass supply pretrained gene representations (Cui et al., 2024; Yang et al., 2024). PerturbNet, Scouter and TxPert draw on functional annotations, gene-description embeddings or knowledge graphs (Yu et al., 2025; Zhu & Li, 2026; Wenkel et al., 2026). For experimental design, the most useful predictions concern pairs of genes that have never been perturbed at all.

However, whatever these models store in learned perturbation embeddings or pair-specific parameters cannot be read out for a gene that was never perturbed. In the strict both-unseen setting, neither gene of a test pair occurs in any training perturbation, so only a response-free descriptor is available for the test genes. Benchmarks already show that complex predictors and foundation-model representations often fail to beat simple linear baselines (Ahlmann-Eltze et al., 2025), and that common metrics can reward systematic variation shared by all perturbations rather than perturbation-specific effects (Viñas Torné et al., 2026). Target-disjoint pilot tests show the same pattern for four natural extensions. Learned target-specific response shapes, graph propagation, predicted susceptibility and flexible correction of an additive prediction each failed to improve held-out prediction. The gain that susceptibility offered an oracle vanished on held-out targets (Appendix A.7). Response structure can thus be learnable

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

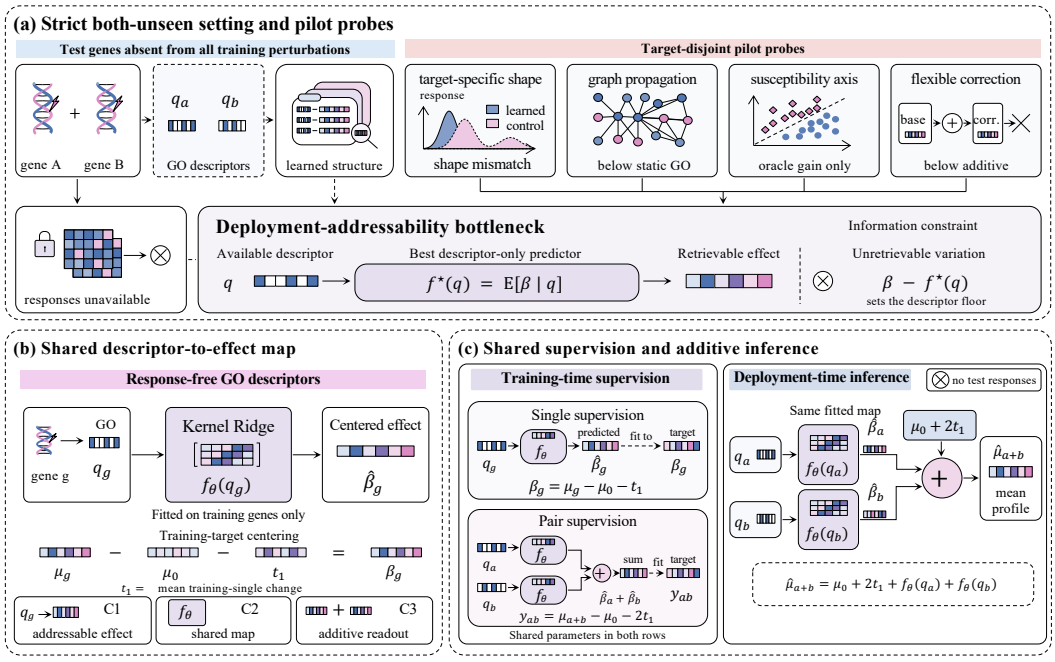


Figure 1: Strict both-unseen prediction and the DACL model. (a) A test pair shares no gene with any training perturbation, so only response-free descriptors are available. Target-disjoint pilot tests ask whether learned structure can be retrieved from them. (b) Kernel ridge regression maps a gene's Gene Ontology descriptor to its centered effect. (c) Single and pair losses fit the same map, and at test time both centered effects are added to $\mu_0 + 2t_1$.

without being retrievable from the descriptor available for a new gene, which makes the storage of learned structure, rather than its amount, the central design question.

To address this limitation, we propose Deployment-Addressable Constituent Learning (DACL), which stores what it learns about individual genes in a function of the descriptor. Kernel ridge regression maps a gene's Gene Ontology membership to its centered effect. Single perturbations supervise this map directly, and paired perturbations supervise it through the sum of the two predicted effects. An unseen pair is predicted by adding both predicted effects and twice the mean single effect to the control. Constituent prediction and additive composition have close precedents in biolord and CPA (Piran et al., 2024; Lotfollahi et al., 2023). Our addition is pair-sum supervision of that same ridge map, so everything learned from pairs remains available for unseen genes. Three elementary identities make this precise. No amount of training data lowers the error floor set by the descriptor. Pair rows move the map only through the residual left by single supervision. A direct pair model is preferable exactly when the approximation error it removes exceeds the estimation error it adds.

We test these predictions on seven screens. On the three official Norman splits, DACL reaches a Pearson-Δ of 0.401 on pairs of unseen genes, compared with 0.337 for biolord and 0.191 for GEARS (Table 1). Most of the gain is present with single supervision alone. Over 70 further splits it reaches 0.294, against 0.206 for biolord and 0.142 for GEARS, and it outperforms both and the direct pair models given the same Gene Ontology descriptors, with gene-level intervals that exclude zero (Appendix A.6). Permuting the assignment of Gene Ontology rows to genes removes most of the gain, and with GenePT embeddings (Chen & Zou, 2025) the advantage stays positive but is not resolved. On the GSE116198 genetic-interaction screen, adding single perturbations improves a factorized pair model by 0.034 but a richer direct one by less than 0.001. A sum-only reparameterization recovers the same gain, so the benefit comes from the additive path rather than a new function class. Direct pair models are preferable on some screens with strong pair-specific structure. On two scalar growth screens, the training-only error gap always favored DirectJoint, which beat DACL-C on pair error in all 45 prospective panels (Appendix F).

We make the following contributions.

- We formulate strict both-unseen pair prediction as a problem of storing learned response structure where a response-free descriptor can retrieve it. We make this precise with a descriptor error floor, the exact effect of pair supervision on a shared ridge map and a risk decomposition that states when a direct pair model should replace constituent composition.
- We introduce DACL, which trains one descriptor-to-effect map with single and paired perturbations and composes predicted effects additively. On the three official Norman splits it reaches a Pearson-Δ of 0.401 on pairs of unseen genes, against 0.337 for the strongest external model. Over 70 further splits it outperforms that model and direct pair models given the same Gene Ontology descriptors.
- We trace the gain to correctly assigned descriptors and additive reuse of single perturbations with matched interventions on target assignment, supervision and model class. We also show when pair-specific capacity helps, including two prospective evaluations run blind to test responses.

2 Related Work

Perturbation-response and compositional models. Single-cell perturbation prediction has moved from latent response models to structured predictors of unseen interventions. scGen shifts perturbation effects in a latent space (Lotfollahi et al., 2019). CPA learns compositional embeddings that extrapolate to unmeasured combinations of seen perturbations (Lotfollahi et al., 2023). CellOT maps control to perturbed cells with optimal transport, and CINEMA-OT separates perturbation effects from confounding variation (Bunne et al., 2023; Dong et al., 2023). GEARS combines co-expression and Gene Ontology graphs to predict multigene perturbations, including combinations with experimentally unseen genes (Roohani et al., 2024), and TxPert integrates several knowledge graphs for unseen perturbations and cell lines (Wenkel et al., 2026). We study which learned structure remains usable when both genes of a pair are absent from every training perturbation.

Unseen targets, descriptors and additive composition. Several methods predict unseen perturbations from information available before any experiment. biolord predicts genetic perturbation effects from gene attributes and forms double-perturbation predictions by adding predicted constituent changes (Piran et al., 2024). PerturbNet maps functional annotations to response distributions, including settings in which both genes are unseen (Yu et al., 2025), and Scouter replaces graph features with language-model embeddings of gene descriptions and evaluates pairs of unseen genes (Zhu & Li, 2026). Response decomposition aligns prior gene profiles with response components and fits direct pair models across availability strata (Molina & Zhang, 2026). scGPT, scFoundation and GeneCompass provide pretrained gene and cell representations (Cui et al., 2024; Hao et al., 2024; Yang et al., 2024). Under additive latent shifts and sufficiently diverse training perturbations, von Kügelgen et al. (2025) identify effects up to an orthogonal transformation and derive guarantees for unseen linear combinations of seen perturbations. DACL combines a response-free descriptor, constituent effects and additive composition. Its single-supervision variant combines two baselines of Ahlmann-Eltze et al. (2025), a linear model on perturbation embeddings and the sum of measured single-perturbation effects. DACL differs in where supervision enters. Singles and pairs train one descriptor-to-effect map, so the quantity refined by pair data is the quantity retrieved for an unseen gene.

Generalization and evaluation. Ahlmann-Eltze et al. (2025) report that deep and foundation-model predictors do not consistently outperform simple linear baselines on genetic perturbation tasks. Viñas Torné et al. (2026) show that common metrics can reward systematic variation shared across perturbations. We therefore evaluate against the mean of training perturbations, include descriptor-free and assignment-permuted controls, and give every internal direct control the same descriptors as DACL. Lippl & Stachenfeld (2025) show that structured representations do not guarantee compositional generalization, which depends on the statistics of the training data. Our risk decomposition makes this explicit for constituent and direct pair models.

3 METHOD

3.1 STRICT BOTH-UNSEEN PREDICTION

Let $\mu_0 \in \mathbb{R}^d$ be the mean response of unperturbed control cells, an expression profile or scalar phenotypes, and μ_g the mean response after perturbing gene g. For a pair (a, b) the target is μ_{a+b}. In the strict both-unseen setting, neither a nor b occurs in any training perturbation, and the only information about a test gene is a response-free descriptor q_g. Our descriptor is the sparse Gene Ontology membership vector of g (Ashburner et al., 2000; Gene Ontology Consortium et al., 2023), which for Norman uses the direct human annotations of the 2026-08-05 release (Carbon & Mungall, 2026). Held-out responses and pairs are not used for preprocessing, selection or fitting.

Any predictor of a gene-level response factor β must then be a function of q. If $\mathbb{E}\|\beta\|_2^2 < \infty$ and $f(q)$ is square integrable, the squared-loss optimum is $f^\star(q) = \mathbb{E}[\beta \mid q]$, and

$$\mathbb{E}\|\beta - f(q)\|_2^2 = \mathbb{E}\|\beta - f^\star(q)\|_2^2 + \mathbb{E}\|f^\star(q) - f(q)\|_2^2. \tag{1}$$

The first term is a floor set by the descriptor. More training responses can reduce the second term but cannot lower the floor. Structure that varies across genes with equal descriptors is therefore not retrievable at test time, however well a model has learned it (proof in Appendix H). In this sense, a stored quantity must be addressable by the descriptor.

3.2 DEPLOYMENT-ADDRESSABLE CONSTITUENT LEARNING

Let $\mathcal{S}$ be the set of training singles, $\delta_g = \mu_g - \mu_0$, and

$$t_1 = |\mathcal{S}|^{-1} \sum_{g \in \mathcal{S}} \delta_g, \qquad \beta_g = \delta_g - t_1. \tag{2}$$

DACL learns a map $f_\theta(q_g)$ that predicts the centered effect β_g. Let $\mathcal{S}_q \subseteq \mathcal{S}$ contain the n_s training singles with a descriptor, $\mathcal{P}_q$ the n_p training pairs whose two genes are in $\mathcal{S}_q$, and $\bar{\beta} = n_s^{-1} \sum_{g \in \mathcal{S}_q} \beta_g$. We use kernel ridge regression on a standardized, training-centered linear kernel of the descriptors with feature map $\phi(q)$, centered training Gram matrix K_c and trace scale $s_K = \mathrm{tr}(K_c)/n_s$, all fitted on $\mathcal{S}_q$ only. Single supervision solves

$$\widehat{W}_{\text{single}} = \arg\min_W \sum_{g \in \mathcal{S}_q} \|\phi(q_g)^\top W - (\beta_g - \bar{\beta})\|_2^2 + \alpha s_K \|W\|_F^2, \qquad f_{\text{single}}(q) = \bar{\beta} + \phi(q)^\top \widehat{W}_{\text{single}}. \tag{3}$$

For a training pair let $y_{ab} = \mu_{a+b} - \mu_0 - 2t_1$. Pair supervision adds pair rows to the same objective, constraining the sum of the two constituent predictions,

$$\begin{aligned} \widehat{W}_{\text{comp}} = \arg\min_W & \sum_{g \in \mathcal{S}_q} \|\phi(q_g)^\top W - (\beta_g - \bar{\beta})\|_2^2 \\ & + \lambda \frac{n_s}{n_p} \sum_{(a,b) \in \mathcal{P}_q} \|(\phi(q_a) + \phi(q_b))^\top W - (y_{ab} - 2\bar{\beta})\|_2^2 + \alpha s_K \|W\|_F^2, \end{aligned} \tag{4}$$

with $\lambda = 1$ fixed, so that the single and pair blocks carry equal total row weight. Setting $\lambda = 0$ recovers Eq. (3) exactly. The ridge parameter α is chosen by target-disjoint inner cross-validation on training singles. An unseen pair is predicted by adding two constituent effects,

$$\widehat{\mu}_{a+b} = \mu_0 + 2t_1 + f(q_a) + f(q_b). \tag{5}$$

With $f_{\text{comp}}(q) = \bar{\beta} + \phi(q)^\top \widehat{W}_{\text{comp}}$, DACL-S uses f_{single} and DACL-C uses f_{comp} in Eq. (5). No parameter is indexed by a gene or a pair, so every learned quantity is available for an unseen gene.

Pair rows refine the same hypothesis class instead of adding one. Let X and Z be the single and pair designs, B and P their centered targets, W_0 the single-only solution, $\omega_\lambda = \lambda n_s / n_p$ and $\rho = \alpha s_K$. Then

$$W_\lambda - W_0 = \omega_\lambda (X^\top X + \omega_\lambda Z^\top Z + \rho I)^{-1} Z^\top (P - Z W_0), \tag{6}$$

so pairs move the map only through the pair residual left by single supervision (Appendix H).

Calibration on training data. Where absolute error matters, an optional channel-wise affine map fitted on out-of-fold training predictions corrects amplitude and offset. With a positive slope, it leaves each scalar channel's Pearson correlation across held-out pairs unchanged (Appendix E.1).

3.3 Conditions favoring pair-specific capacity

Write the centered pair change as $Y_{ab} = \beta_a + \beta_b + \eta_{ab}$, where η_{ab} is the non-additive interaction, and let $e_g = f(q_g) - \beta_g$. The additive prediction $\widehat{Y}_{ab} = f(q_a) + f(q_b)$ has error

$$\widehat{Y}_{ab} - Y_{ab} = e_a + e_b - \eta_{ab}. \tag{7}$$

Two maps with equal single-gene risk can therefore have different pair risk, through the cross terms between e_a, e_b and η_{ab}. A direct pair model can exploit only the part of η_{ab} that is predictable from (q_a, q_b). To compare the two choices, let $Q = (q_a, q_b)$, $m(Q) = \mathbb{E}[Y_{ab} \mid Q]$, $R(h) = \mathbb{E}\|Y - h(Q)\|_2^2$ and $R_{\text{info}} = \mathbb{E}\|Y - m(Q)\|_2^2$. For a model class $\mathcal{H}_M$ with population minimizer h_M^* and fitted estimator $\widehat{h}_M$,

$$R(\widehat{h}_M) = R_{\text{info}} + A_M + X_M, \tag{8}$$

where $A_M = R(h_M^*) - R_{\text{info}}$ is the approximation gap of the class and $X_M = R(\widehat{h}_M) - R(h_M^*)$ its excess estimation risk. For the additive constituent class C and a richer direct pair class D,

$$R(\widehat{h}_C) - R(\widehat{h}_D) = (A_C - A_D) + (X_C - X_D), \qquad R(\widehat{h}_D) < R(\widehat{h}_C) \iff A_C - A_D > X_D - X_C. \tag{9}$$

The direct model has lower risk exactly when its approximation advantage $A_C - A_D$ exceeds its excess-risk difference $X_D - X_C$. This motivates a training-only diagnostic of single-effect predictability, pair support and residual predictability (Appendix D). The prospective tests estimate the risk difference of the two classes from target-disjoint inner-fold errors.

3.4 Controls with the same information

Every internal control uses the same descriptors, training responses, target-disjoint model selection, missing-descriptor fallback (a zero centered effect in constituent models, a zero descriptor in direct models) and evaluator as DACL. DirectJoint maps the symmetric pair feature $x_{ab} = [q_a + q_b,\ |q_a - q_b|,\ q_a \odot q_b]$ to the centered pair response with ridge regression or a one-hidden-layer MLP with tuned width, weight decay and epochs. On GSE116198, a neural factorial uses one descriptor encoder with a factorized head, its sum-only reparameterization or a richer direct head with interaction features. A flexible composer replaces the addition in Eq. (5) by a learned symmetric function of the two constituent predictions. Architectures, grids and parameter counts are in Appendix A.2.

4 Experimental Setup

Screens and splits. Norman is a CRISPR activation Perturb-seq screen in K562 cells with single and paired perturbations (Norman et al., 2019). We use the simulation splits of the released GEARS code for seeds 1 to 3 (Roohani et al., 2024), called official splits below, and score two strata of held-out pairs. Following GEARS, the strata count a pair's genes that occur in training or validation perturbations (none in Seen-0, two in Seen-2), and all models fit training conditions only (Appendix A.4). Seen-0 contains 25 test instances (23 unique pairs over 24 genes), so we also evaluate on 70 further splits generated by the same procedure (584 Seen-0 instances) and compute intervals by resampling held-out genes. Within each of six further screens, models are refitted under the same strict both-unseen rule, with held-out genes chosen by an outcome-free rule and every pair involving a held-out gene removed from training. GSE116198 (Horlbeck et al., 2018) is the primary mechanism test, GSE290153 (Nebenfuehr et al., 2026) separates reuse from factorization, GSE200201 (Otto et al., 2023) tests when pair-specific capacity pays off, GSE284018 (Simpson et al., 2023) provides a held-out endpoint, and PRJNA691742 (Zhao et al., 2021) and Shen2017 (Shen et al., 2017) are prospective tests. Wessels (Wessels et al., 2023), MCF10A (GSE164996, Zhao et al., 2021) and ComboSciPlex (Srivatsan et al., 2020; Lotfollahi et al., 2023) delimit the scope (Appendix E).

Table 1: Norman Pearson-Δ for pairs of unseen genes (SEEN-0) and unseen pairs of seen genes (SEEN-2). Scores are averaged over held-out conditions within a split and reported as mean $\pm$ s.d. over the three official splits and 70 further splits (584 SEEN-0 and 1263 SEEN-2 test instances). Bold marks the best value in each column, and dashes mark models that were not run or do not apply. External methods are our runs of their released code, not published scores. The last four rows use the Gene Ontology descriptor. RMSE, top-20 DE Pearson-Δ and centroid accuracy are in Table A.4.

	Official splits ($n = 3$)		Further splits ($n = 70$)	
Method	SEEN-0	SEEN-2	SEEN-0	SEEN-2
Perturbed Mean	-0.143 ± 0.246	-0.154 ± 0.065	-0.076 ± 0.207	-0.088 ± 0.114
Matching Mean	–	0.707 ± 0.053	–	0.660 ± 0.066
Descriptor-free additive	0.209 ± 0.182	0.164 ± 0.034	0.125 ± 0.207	0.129 ± 0.111
GEARS 0.1.2	0.191 ± 0.164	0.496 ± 0.079	0.142 ± 0.127	0.442 ± 0.099
biolord	0.337 ± 0.149	$\mathbf{0.787\pm0.064}$	0.206 ± 0.162	$\mathbf{0.762\pm0.056}$
scGPT	0.115 ± 0.033	0.545 ± 0.107	–	–
Scouter	0.217 ± 0.187	0.639 ± 0.066	0.208 ± 0.122	0.606 ± 0.062
Response decomp., Ridge	0.242 ± 0.100	0.262 ± 0.089	–	–
Response decomp., MLP	0.198 ± 0.105	0.230 ± 0.122	–	–
DepMap Ridge	0.173 ± 0.160	0.268 ± 0.048	0.055 ± 0.167	0.196 ± 0.121
DepMap MLP	0.227 ± 0.153	0.527 ± 0.045	0.068 ± 0.161	0.480 ± 0.100
DirectJoint Ridge	0.380 ± 0.221	0.776 ± 0.092	0.269 ± 0.189	0.751 ± 0.072
DirectJoint MLP	0.386 ± 0.241	0.707 ± 0.098	0.271 ± 0.196	0.703 ± 0.080
DACL-S (singles)	0.375 ± 0.210	0.773 ± 0.069	0.285 ± 0.186	0.738 ± 0.070
DACL-C (singles + pairs)	$\mathbf{0.401\pm0.211}$	0.770 ± 0.075	$\mathbf{0.294\pm0.186}$	0.760 ± 0.065

Metrics. The primary endpoint, Pearson-Δ, averages over held-out conditions the Pearson correlation of true and predicted profiles after subtracting from each the unweighted mean of the training perturbation profiles. Centering on this mean removes the component shared by all training perturbations (Viñas Torné et al., 2026). We also report the root mean squared error (RMSE), Pearson-Δ on the 20 genes that GEARS lists as most differentially expressed (DE) for each pair, and centroid accuracy, the fraction of other test conditions whose true profile is farther from the prediction than the correct one. Scalar readouts use Pearson correlation across held-out pairs, averaged over channels.

Baselines. Perturbed Mean predicts the cell-weighted average training perturbation, so its Pearson-Δ reflects only cell versus profile weighting. Matching Mean averages the two genes' measured single profiles when both exist. The descriptor-free additive baseline predicts $\mu_0 + 2t_1$. GEARS 0.1.2 is run with its released code on the same splits. The biolord model (Piran et al., 2024) uses its released Norman recipe to predict single-gene effects from Gene Ontology attributes and add the effects of the two genes. Scouter (Zhu & Li, 2026) uses its released Norman script and language-model embeddings of gene descriptions. scGPT (Cui et al., 2024) is fine-tuned with its perturbation tutorial on the official splits. Response decomposition (Molina & Zhang, 2026) is run with the released Ridge and MLP pair models, which are trained to map the summed DepMap profiles of the two genes to the pair response (Table C.1). DepMap Ridge and MLP predict constituent effects from DepMap gene-effect profiles (Tsherniak et al., 2017; Broad DepMap, 2024) (knockout effects on growth across cancer cell lines) and add these effects. All controls in Section 3.4 use the Gene Ontology descriptor, and Table G.1 lists how each comparator was run. Models are fitted on training responses only. Predictions are recorded before test responses are read, and no hyperparameter is tuned on a test split.

5 RESULTS

5.1 DACL PREDICTS PAIRS OF UNSEEN GENES ON NORMAN

On the three official Norman splits, DACL-C reaches a Pearson-Δ of 0.401 on pairs of unseen genes, compared with 0.337 for biolord, 0.242 for response decomposition, 0.227 for the strongest DepMap model, 0.217 for Scouter, 0.191 for GEARS and 0.115 for scGPT (Table 1). It also has the lowest RMSE, whereas response decomposition and DirectJoint Ridge score higher on the top-20 DE genes (Table A.4). GEARS does not improve on the descriptor-free additive prediction (0.209) in this

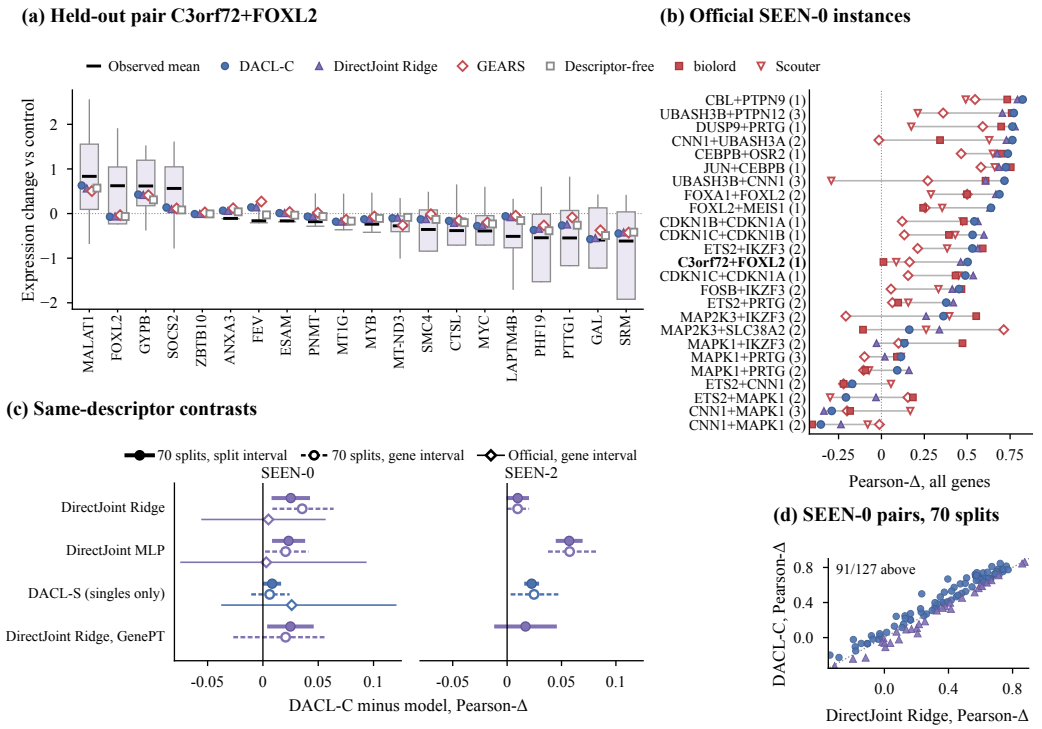


Figure 2: DACL-C predicts held-out pairs of unseen genes better than GEARS, and better than direct pair models with the same descriptor over 70 further splits. (a) Top-20 differentially expressed genes of C3orf72+FOXL2 (official split 1), the SEEN-0 instance with the median DACL-C score. Boxes and whiskers give the 25–75% and 5–95% ranges of the change from control across cells. (b) Pearson-Δ over all genes for the 25 official SEEN-0 test instances (split in parentheses, pair of (a) in bold). (c) DACL-C minus models with the same descriptor, with 95% t intervals over the 70 further splits and 95% intervals over resampled genes. (d) DACL-C against DirectJoint Ridge on the 127 unique SEEN-0 pairs of the 70 splits, with the pairs where DACL-C is higher in blue.

stratum. Figure 2(a,b) shows one pair gene by gene and every test instance, and Figures A.1 and A.2 show all splits and more pairs. Resampling held-out genes over the 23 unique pairs gives intervals that exclude zero for the gains over GEARS (0.216, $[0.012, 0.397]$), Scouter, scGPT, DepMap MLP and the descriptor-free baseline, but not over biolord or response decomposition (Table A.7).

The official splits score higher than the further splits. Over 70 further splits, DACL-C reaches 0.294 on pairs of unseen genes, against 0.208 for Scouter, 0.206 for biolord, 0.142 for GEARS, 0.125 for the descriptor-free baseline and 0.068 for DepMap MLP. DACL-C exceeds biolord by 0.088 over splits ($[0.059, 0.116]$) and by 0.079 over unique pairs with resampled genes ($[0.016, 0.143]$, Table A.7). Most of the gain is present with single supervision alone (DACL-S, 0.285). For unseen pairs of seen genes, biolord adds measured or predicted single-gene effects, is best on the official splits (0.787) and matches DACL-C over the 70 further splits (0.762 and 0.760). DACL fits and scores a split in 2 to 4 s on four CPU threads, whereas scGPT fine-tuning takes 3 to 5.5 h of GPU time per split (Table A.6).

Direct pair models given the same Gene Ontology descriptors provide the stronger comparison (Figure 2(c)). Over the 70 further splits, DACL-C exceeds DirectJoint Ridge by 0.025 ($[0.008, 0.043]$) and DirectJoint MLP by 0.023 ($[0.008, 0.038]$) on SEEN-0. Resampling held-out genes yields 0.035 ($[0.009, 0.064]$) and 0.020 ($[0.002, 0.041]$), and DACL-C scores higher than DirectJoint Ridge on 91 of the 127 unique pairs (Figure 2(d)). On SEEN-2 DirectJoint Ridge is on par with DACL-C (difference 0.010, $[-0.001, 0.020]$) and slightly better on the official splits, whereas the MLP trails by 0.057 (Table A.9). Over 70 further splits, pair supervision adds 0.008 ($[0.000, 0.016]$) on SEEN-0 and 0.023 ($[0.016, 0.029]$) on SEEN-2 (Table A.9, official splits in Figure 3(e)). Over resampled genes the SEEN-0 gain, 0.006 ($[-0.011, 0.024]$), is not resolved. A learned symmetric composer in place of the sum in Eq. (5) trails DACL-C by 0.064 on SEEN-0 and 0.494 on SEEN-2 (Table A.9).

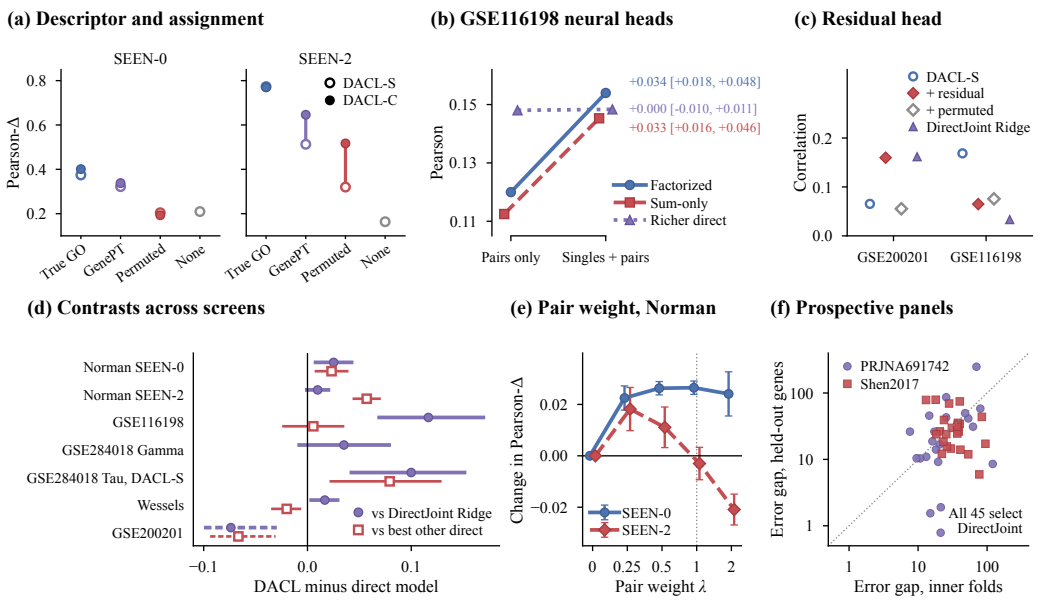


Figure 3: The Norman gain depends on correctly assigned descriptors, and the GSE116198 gain comes from additive reuse of single perturbations. On several other screens direct pair models are better. (a) Mean Pearson-Δ of DACL-S (open) and DACL-C (filled) over the official Norman splits with the true Gene Ontology descriptor, GenePT embeddings, permuted Gene Ontology rows and no descriptor. (b) GSE116198 Pearson of three neural heads sharing one encoder, trained on pairs only or on singles and pairs, with the gain from singles and its 95% interval. (c) DACL-S with a pair-residual head fitted to true or permuted residuals (32 permutations) and DirectJoint Ridge (Pearson-Δ on GSE200201, Pearson on GSE116198). (d) DACL minus DirectJoint Ridge (filled) and the best other direct model (open), with 95% intervals over Norman splits or held-out genes and leave-one-gene-out ranges (dashed). DACL is the prespecified DACL-S on GSE284018 Tau and DACL-C elsewhere, and the open GSE116198 marker compares the heads of (b). (e) Change in Pearson-Δ from $\lambda = 0$ as the pair weight grows on the official Norman splits, mean $\pm$ s.d., with $\lambda = 1$ in DACL-C. (f) DACL-C minus DirectJoint normalized pair mean squared error, the negative of the gap in Appendix F, in the 45 prospective panels on inner folds and held-out genes. The rule selected DirectJoint in every panel.

5.2 Correct gene-to-descriptor assignment carries the signal

If DACL works because the descriptor retrieves learned structure, breaking the gene-to-descriptor link should remove the gain. A fixed-point-free permutation of Gene Ontology rows across genes preserves the row multiset, sparsities and column marginals. It lowers DACL-S from 0.375 to 0.204 on SEEN-0, to the level of the descriptor-free baseline, and from 0.773 to 0.320 on SEEN-2 (Figure 3(a), Table B.4). On SEEN-0, at fixed pair support and weight, observed pair targets add 0.027, targets permuted across pairs subtract 0.085, and synthetic additive targets built from training singles add 0.009 (Table B.6). The pair gain thus depends on correctly assigned pair responses.

With GenePT gene-summary embeddings (Chen & Zou, 2025) in place of Gene Ontology, DACL-C still exceeds DirectJoint Ridge on SEEN-0 over the 70 further splits. The advantage is 0.025 over splits ($[0.004, 0.046]$) and 0.021 over unique pairs with resampled genes ($[-0.027, 0.057]$), unresolved at the gene level (Table B.3). GenePT is weaker on Norman (DACL-S 0.212 against 0.285), and pair supervision adds more (0.039 on SEEN-0 and 0.132 on SEEN-2), consistent with Eq. (6).

5.3 Single-perturbation supervision is reused through the additive path

GSE116198 is a CRISPR interference genetic-interaction screen with 467 genes and scalar phenotypes in two cell lines, evaluated on five splits that hold out 20% of genes (Appendix A.4). DACL-C reaches a Pearson correlation of 0.150, compared with 0.033 for DirectJoint Ridge and 0.028 for a DirectJoint

MLP on the raw pair features (difference from Ridge 0.116, $[0.069, 0.169]$, 9 of 10 panels, Tables C.3 and E.4). To locate the gain, a factorial comparison of neural models fixes the descriptor encoder, optimizer and model selection and varies only the prediction head and the supervision (Figure 3(b), Table D.2). With singles and pairs, the factorized head, which adds predicted gene effects, scores 0.154, and the richer direct head scores 0.148 with a lower RMSE (difference 0.006, $[-0.023, 0.034]$). Adding the same single rows improves the factorized head by 0.034 ($[0.018, 0.048]$) and the direct head by less than 0.001 ($[-0.010, 0.011]$), an interaction of 0.034 ($[0.016, 0.050]$). A sum-only head with the same parameter count, which algebraically reparameterizes the factorized head, gains 0.033. In a separate same-encoder comparison with pair-only training, the factorized and direct heads are on par at 0.127 and 0.129 (Table D.4). The gain therefore comes from the additive path by which single responses reach pair predictions, not from a new function class (Table D.3). On the independent GSE290153 screen, factorization helps with pairs only ($+0.018$) and much less once singles are added ($+0.004$), with all intervals including zero (Appendix D.1).

5.4 Other screens and the limits of constituent composition

Figure 3(d) and Table E.1 compare DACL with direct models across screens. On GSE284018, we fixed DACL-S and its calibration on Gamma, fitted them to training Tau responses and recorded predictions before reading held-out ones. On Tau it reaches 0.314, above DirectJoint Ridge by 0.100 ($[0.042, 0.151]$) and the sum-only model by 0.079 ($[0.023, 0.128]$) in every panel (Table E.3).

Section 3.3 predicts that direct capacity can win when the non-additive part of the pair response is predictable from the descriptors. GSE200201, a Perturb-seq screen of SWI/SNF complex subunits in MOLM13 cells, has on average 13 training pairs per panel whose two genes have descriptors, against about 52,000 in GSE116198. DirectJoint Ridge reaches 0.162 and DACL-C 0.088, and the direct model stays ahead after removing any single gene of this 27-pair screen (Table E.7). A residual head for the pair interaction raises DACL from 0.065 to 0.160 while slightly worsening RMSE, and permuting its targets across pairs removes the gain (Figure 3(c), Table D.6). The same head lowers Pearson on GSE116198 from 0.169 to 0.065 while improving RMSE, consistent with the association signal being carried by the additive part there. A training-only diagnostic distinguishes the screens by pair support but not by residual predictability, which is negative in both (Appendix D).

In two prospective evaluations, predictions were recorded before test responses were read. The training-only error gap favored DirectJoint over DACL-C and matched the direction of the test gap in all 21 PRJNA691742 and 24 Shen2017 panels, which coincides with always choosing DirectJoint (Figure 3(f), Appendix F). Here DACL has 20 to 30 times the normalized pair error of DirectJoint.

6 Limitations

We refitted models within each screen, with Gene Ontology as the main descriptor and GenePT on Norman only. Transfer of a fitted model across studies remains untested, and adding DepMap profiles to Gene Ontology did not improve Norman (Table B.2). The official Norman SEEN-0 splits have 23 unique pairs, and the 70 further splits overlap and are not independent replicates. Pair supervision helps on Wessels, ComboSciPlex drug combinations (Table C.2) and Norman pairs of seen genes over the 70 further splits. It offers little benefit on Norman pairs of unseen genes and does not help on GSE284018. The direct MLP is better on Wessels, where every model has a negative Pearson-Δ (Table E.5), and direct pair models are better on GSE200201, in Pearson on Shen2017 and in absolute error on both prospective screens (Appendix E). The training-only error gap has not been validated as a switch between model classes. GSE284018 Tau is a new endpoint of the same screen, not an independent replication. Predictions are mean responses rather than cell-level distributions.

7 Conclusion

DACL learns a descriptor-indexed map of single-gene effects from single and paired perturbations and adds predicted effects for pairs of unseen genes. On such pairs it outperforms GEARS, biolord and same-descriptor direct models on Norman, while direct pair models remain better on some screens with strong pair-specific structure.

AI USE STATEMENT

We used large language models solely for code implementation assistance, manuscript copy-editing and polishing, and generating conceptual schematic icons. All experimental workflows, mathematical formulations, data analyses and scientific claims were designed, conducted and verified by the authors.

ETHICS STATEMENT

This work reanalyzes published CRISPR perturbation screens. It involves no human participants, no new experiments and no identifiable personal data, and each dataset is used under the terms of its public release. Predicted perturbation responses are intended to prioritize experiments rather than replace them, and any predicted interaction requires experimental validation. The authors have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 define the model, the strict both-unseen splits and the evaluation. Appendix A.2 gives the grids and architectures of all controls, Appendix A.4 the split sizes, and Appendix H the proofs. All screens are public. The anonymized supplementary archive contains DACL, DACL-S, DirectJoint Ridge, the descriptor-free additive baseline, the evaluator and the Norman splits, with a script that runs these four models on Norman.

REFERENCES

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

Michael Ashburner et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1): 25–29, 2000. doi: 10.1038/75556.

Broad DepMap. DepMap 24Q4 Public. Figshare+ dataset, 2024. URL `https://doi.org/10.25452/figshare.plus.27993248.v1`.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

Seth Carbon and Chris Mungall. Gene Ontology Data Archive, version 2026-08-05. Zenodo dataset, 2026. URL `https://doi.org/10.5281/zenodo.21844811`.

Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nature Biomedical Engineering*, 9(4):483–493, 2025. doi: 10.1038/s41551-024-01284-6.

Haotian Cui et al. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038.

Mingze Dong, Bao Wang, Jessica Wei, Antonio H. de O. Fonseca, Curtis J. Perry, Alexander Frey, Feriel Ouerghi, Ellen F. Foxman, Jeffrey J. Ishizuka, Rahul M. Dhodapkar, and David van Dijk. Causal identification of single-cell experimental perturbation effects with CINEMA-OT. *Nature Methods*, 20(11):1769–1779, 2023. doi: 10.1038/s41592-023-02040-5.

Gene Ontology Consortium, Suzi A. Aleksander, et al. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023. doi: 10.1093/genetics/iyad031.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, 2024. doi: 10.1038/s41592-024-02305-7.

Max A. Horlbeck et al. Mapping the genetic landscape of human cells. *Cell*, 174(4):953–967.e22, 2018. doi: 10.1016/j.cell.2018.06.010.

Samuel Lippl and Kimberly Stachenfeld. When does compositional structure yield compositional generalization? a kernel theory. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/572a6f16ec44f794fb3e0f8a310acbc6-Abstract-Conference.html.

Tianyu Liu, Tianqi Chen, Wangjie Zheng, Xiao Luo, Yiqun Chen, and Hongyu Zhao. Embeddings from language models are good learners for single-cell data analysis. *Patterns*, 7(2):101431, 2026. doi: 10.1016/j.patter.2025.101431.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, 2019. doi: 10.1038/s41592-019-0494-8.

Mohammad Lotfollahi et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. doi: 10.15252/msb.202211517.

Alexis Molina and Xinyi Zhang. Perturbation response decomposition enables biologically aligned generalization to unseen perturbations and cellular contexts. *bioRxiv preprint*, 2026. doi: 10.64898/2026.07.24.740459.

Benjamin Nebenfuehr et al. Uncovering genetic interactions in the DNA repair network in response to endogenous damage and ionizing radiation. *Cell Reports*, 45(1):116850, 2026. doi: 10.1016/j.celrep.2025.116850.

Thomas M. Norman et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438.

Jordan E. Otto et al. Structural and functional properties of mSWI/SNF chromatin remodeling complexes revealed through single-cell perturbation screens. *Molecular Cell*, 83(8):1350–1367.e7, 2023. doi: 10.1016/j.molcel.2023.03.013.

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42(11):1678–1683, 2024. doi: 10.1038/s41587-023-02079-x.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024. doi: 10.1038/s41587-023-01905-6.

John Paul Shen et al. Combinatorial CRISPR–Cas9 screens for de novo mapping of genetic interactions. *Nature Methods*, 14(6):573–576, 2017. doi: 10.1038/nmeth.4225.

Danny Simpson, Jia Ling, Yangwode Jing, and Britt Adamson. Mapping the genetic interaction network of PARP inhibitor response. *bioRxiv preprint*, 2023. doi: 10.1101/2023.08.19.553986. URL https://www.biorxiv.org/content/10.1101/2023.08.19.553986.

Sanjay R. Srivatsan et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020. doi: 10.1126/science.aax6234.

Aviad Tsherniak, Francisca Vazquez, Phil G. Montgomery, et al. Defining a cancer dependency map. *Cell*, 170(3):564–576.e16, 2017. doi: 10.1016/j.cell.2017.06.010.

Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbić. Systema: A framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 44(6):1050–1059, 2026. doi: 10.1038/s41587-025-02777-8.

Julius von Kügelgen, Jakob Ketterer, Michael Vollenweider, Michael Scholkemper, Xinwei Shen, Nicolai Meinshausen, and Jonas Peters. Extrapolation guarantees for perturbation modeling under the additive latent shift assumption. *arXiv preprint arXiv:2504.18522*, 2025. doi: 10.48550/arXiv.2504.18522. URL https://arxiv.org/abs/2504.18522. Version 3, 2026.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Liam Hodgson, Ihab Bendidi, et al. TxPert: Using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-026-03113-4.

Hans-Hermann Wessels et al. Efficient combinatorial targeting of RNA transcripts in single cells with Cas13 RNA Perturb-seq. *Nature Methods*, 20(1):86–94, 2023. doi: 10.1038/s41592-022-01705-x.

Xiaodong Yang, Guole Liu, Guihai Feng, Dechao Bu, Pengfei Wang, Jie Jiang, Shubai Chen, Qinmeng Yang, Hefan Miao, Yiyang Zhang, Zhenpeng Man, et al. GeneCompass: Deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Research*, 34(12):830–845, 2024. doi: 10.1038/s41422-024-01034-y.

Hengshi Yu, Weizhou Qian, Yuxuan Song, and Joshua D. Welch. PerturbNet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular Systems Biology*, 21(8):960–982, 2025. doi: 10.1038/s44320-025-00131-3.

Xiaoyu Zhao, Jinyu Li, Zhimin Liu, and Scott Powers. Combinatorial CRISPR/Cas9 screening reveals epistatic networks of interacting tumor suppressor genes and therapeutic targets in human breast cancer. *Cancer Research*, 81(24):6090–6105, 2021. doi: 10.1158/0008-5472.CAN-21-2555.

Ouyang Zhu and Jun Li. Scouter predicts transcriptional responses to genetic perturbations with large language model embeddings. *Nature Computational Science*, 6(1):21–28, 2026. doi: 10.1038/s43588-025-00912-8.

APPENDIX

The appendix follows the order of evidence in the main text. Appendix A gives the evaluation protocol, the architectures of all controls, dataset and split sizes, the complete Norman metrics and gene-level intervals behind Section 5.1, and the pilot tests cited in the introduction. Appendix B contains the descriptor, assignment and pair-information controls of Section 5.2. Appendices C and D cover comparators, transfer, the GSE116198 and GSE290153 factorials and the pair-residual diagnostic of Sections 5.3 and 5.4. Appendix E reports the other screens and the calibration procedure. Appendix F presents the prospective capacity tests, Appendix H gives the proofs and Appendix G describes how each comparator was run.

CONTENTS

A Evaluation protocol and target-disjoint pilot tests

B Descriptor assignment and alternative explanations

C Comparators, transfer and the primary mechanism screen

D Supervision reuse and pair-specific capacity

E External screens and scope

F Prospective capacity tests

G Comparators and the questions they test

H Theory and metric geometry

A EVALUATION PROTOCOL AND TARGET-DISJOINT PILOT TESTS

Throughout the appendix we abbreviate Gene Ontology as GO.

A.1 MODEL FITTING AND SELECTION

DACL-S and DACL-C use a standardized, training-centered linear kernel of the descriptors, a fixed intercept and target-disjoint inner validation. The ridge grid is $\{0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10\}$, and pair supervision uses a fixed pair-block weight of $\lambda = 1$. Models are fitted in float64 and predictions are stored in float32. We verified that $\lambda = 0$ reproduces DACL-S before reporting DACL-C.

The DirectJoint MLP receives the symmetric Gene Ontology pair feature $[q_a + q_b, |q_a - q_b|, q_a \odot q_b]$ and has one hidden layer. Its grid covers $\{32, 64\}$ hidden units, weight decay $\{10^{-4}, 10^{-3}\}$ and $\{100, 300\}$ epochs, with learning rate 0.003, gradient clipping at 5 and three fixed training seeds. It predicts coordinates in a weighted singular value decomposition basis of the training responses, with ranks 67, 65 and 64 on the three official splits. Hyperparameters are selected by four target-disjoint inner folds on training responses only. The MLP receives the same information as DACL and its capacity is tuned independently, so parameter count is not used to match the two models.

A.2 ARCHITECTURES OF THE MECHANISM CONTROLS

The GSE116198 factorial crosses model class (factorized or direct) with supervision (pairs only, or singles and pairs). The factorial uses a separate response target so that the gain from adding singles is not confounded with the centering used by DACL. Every configuration receives the same 2263-dimensional sparse Gene Ontology vector q_g and the same one-hidden-layer ReLU encoder $h_g = \mathrm{ReLU}(q_g W_e + b_e)$ with 48 hidden units and 108,672 encoder parameters. In this factorial only, every training row targets the change from control, $r = \mu - \mu_0$. Models trained on pairs only therefore use only pair rows and compute no t_1, response basis or other statistic of single responses. The main DACL fit uses the centered quantities of Eqs. (2)–(5).

For the two direct heads, let $m_{ab} = 1$ for a pair row and $m_{ab} = 0$ for a single row. The second encoded gene is the fixed zero vector when $m_{ab} = 0$, and the count block is $1 + m_{ab}$. A single perturbation therefore enters as $G(h_g, 0, 0)$ and a pair as $G(h_a, h_b, 1)$, so the fourth feature block is 1 for a single and 2 for a pair and is not derived from responses. In the factorial the target is always $r = \mu - \mu_0$, and the evaluator restores the control as $\widehat{\mu} = \mu_0 + \widehat{r}$ for every model.

Table A.1: Mechanism controls on GSE116198. All configurations use the same GO input, encoder, target-disjoint folds, optimizer and paired encoder initializations, and select hyperparameters by pooled mean squared error (MSE) on pairs only. The gain from adding singles is compared within a model, whereas the final score is compared between models.

Model	Single input and pair input	Target and offset	Supervision and selection	Parameters
Factorized	$b_g = h_g W_c + c_c$; single b_g; pair $b_a + b_b$	change from control r in the factorial; the main DACL fit adds $\mu_0 + 2t_1$ after centered composition	pair-only uses pair rows; joint uses single and pair rows with n_s/n_p pair weight; four target-disjoint folds; $\{100, 300\}$ epochs and $\{10^{-4}, 10^{-2}\}$ decay	108,770
Richer DirectJoint	$G(h_a, h_b, m) = [h_a + h_b, \lvert h_a - h_b \rvert, h_a \odot h_b, 1 + m]W_p + c_p$; single $G(h_g, 0, 0)$; pair $G(h_a, h_b, 1)$; shared head	change-from-control target; the control is added back at evaluation	same two supervision blocks, folds, optimizer, paired initialization, and pooled pair-only MSE selection	108,964
Parameter-matched sum-only Direct	$G_{\mathrm{sum}}(h_a, h_b, m) = [h_a + h_b, 1 + m]W_{\mathrm{sum}}$; single $G_{\mathrm{sum}}(h_g, 0, 0)$; pair $G_{\mathrm{sum}}(h_a, h_b, 1)$; no free head bias	change-from-control target; same response and offset in evaluation	same two supervision blocks, folds, optimizer, paired initialization, and pooled pair-only MSE selection	108,770

In the models trained on singles and pairs, pair rows receive weight n_s/n_p and single rows unit weight. At the fixed setting $\lambda = 1$ the two blocks therefore carry equal total weight in the objective, although individual rows are weighted differently. The richer DirectJoint model has a free head bias and interaction blocks. The sum-only direct model has no free head bias and sees only the sum of the two hidden states plus the fixed count block. All settings are selected on training responses only, and predictions are exported before any test response is read. The sum-only direct model therefore stays in the function class of the factorized model and differs only in parameterization, with the descriptor, parameter budget, supervision blocks, optimization and selection criterion fixed. The richer DirectJoint head is the comparison that changes the function class.

The raw-feature DirectJoint MLP is a separate capacity control on GSE116198. It receives the full 6,789-dimensional symmetric feature $[q_a + q_b, \lvert q_a - q_b \rvert, q_a \odot q_b]$ without hashing or projection. Its grid covers width $\{16, 64\}$, weight decay $\{10^{-6}, 10^{-4}, 10^{-2}, 10^{-1}\}$ and $\{100, 300, 600\}$ epochs, with learning rate 0.003, gradient clipping at 5, four target-disjoint inner folds and three fixed training seeds for the final fit. Selection uses pooled pair MSE and only training responses. The control receives the same information as DACL and its capacity is tuned. DACL's Pearson advantage over this control is interpreted separately from absolute error and does not imply that DACL is better on every absolute-error metric.

The nonlinear map in the descriptor-by-capacity factorial is a separate capacity test, not a DirectJoint comparator. It has one hidden layer with 32 units and an 8-dimensional response basis fitted on all training singles of the split. Adam settings are fixed (learning rate 0.01, weight decay 0.01, 800 epochs), without early stopping. The same input map serves singles and pairs, and pair supervision acts only through the sum of the two constituent outputs. The constant, true GO and DepMap-24Q4 descriptors and their fallbacks for missing features were specified before the experiment was run. The design also included a permutation of GO feature columns. This permutation only renames feature coordinates, so it does not break the assignment of terms to genes and leaves the standardized linear Gram matrix unchanged. We therefore do not use it as a negative control.

For each outer split, feature standardization and response centering are fitted on its training conditions. Regularization is then selected using deterministic target-disjoint inner folds, DACL-S and DACL-C are fitted, and a separate evaluator scores the exported predictions. Held-out responses are not used

for feature processing, model selection or fitting. Genes without a response-free descriptor receive the same fallback in every matched method.

A.3 DATASETS AND SPLITS

Norman. Both strata use the same Pearson-Δ relative to the perturbation mean, averaged over conditions. For a held-out condition c, let $z_c = \mu_c - \mu_{\mathrm{ref}}$ and $\hat{z}_c = \hat{\mu}_c - \mu_{\mathrm{ref}}$, where μ_{ref} is the unweighted mean of the training perturbation profiles. The condition score is $\mathrm{Pearson}(z_c, \hat{z}_c)$, and the reported value is its unweighted mean over held-out conditions within each split and stratum. In SEEN-0 neither gene of a test pair occurs in a training or validation perturbation, and in SEEN-2 both genes occur as training or validation singles. DirectJoint Ridge and MLP use the symmetric pair features $[q_a + q_b, |q_a - q_b|, q_a \odot q_b]$ and the same training responses and response basis as the factorized model. The flexible composer, which we denote by U_{free}, receives the same out-of-fold DACL-S predictions, rank-8 response basis, fixed offset, support, evaluator and fallback as explicit addition, and replaces only the addition. The contrast $f_{\mathrm{single}} - U_{\mathrm{free}}$ therefore tests the composition operator alone. The robustness comparison of f_{comp} with U_{free} also changes pair supervision, so it is a joint structural control rather than a test of the operator alone.

Auxiliary Norman panel. A separate strict both-unseen panel of Norman pairs is used for the matched contrast $f_{\mathrm{comp}} - f_{\mathrm{single}}$. We report this panel separately from the larger Norman panels because its test pairs differ.

Wessels. For Wessels we use condition centroids from the THP-1 CaRPool-seq screen. Training excludes every single and pair that involves a held-out gene, and each test condition contains two held-out genes. The GO and DepMap panels differ only in the response-free gene descriptor, and the objective, split rule, inference rule, metric and aggregation are fixed. The matched controls keep the three original splits, GO rows, training responses and evaluator, and add DirectJoint Ridge and DirectJoint MLP with the same symmetric pair features. Their predictions were exported before the test responses were read.

MCF10A GSE164996. For MCF10A we use the official perturbation graph and our own strict both-unseen split over 11 genes. For each of three fixed seeds, five genes are held out by a fixed pseudorandom rule that does not use outcomes. No training pair contains a held-out gene, and every test pair contains two. The same evaluator scores the test pairs, with scores averaged over conditions. This result delimits the scope of DACL under our own split and does not use an official benchmark partition.

GSE90063. GSE90063 (Dixit et al., 2016) is an independent genetic screen of double perturbations, analyzed with the same fixed method and evaluation protocol. Its aggregate uncertainty is reported with the other external screens.

GSE200201. GSE200201 is an independently generated CRISPR knockout Perturb-seq screen of mSWI/SNF subunits in MOLM13 cells. An outcome-free rule on the pair graph assigns each of the 27 eligible pairs among 20 genes to the strict both-unseen test set of exactly one of seven fixed splits. Models are refitted within the dataset. The primary statistic is the mean of $f_{\mathrm{comp}} - f_{\mathrm{single}}$ over unique pairs, with a 10,000-replicate two-way bootstrap over gene clusters. The signs of individual splits are descriptive, and every split is reported.

ComboSciPlex. ComboSciPlex (Lotfollahi et al., 2023), a screen of drug combinations, uses a fixed split of 24 training and seven held-out conditions. The same model formulations are refitted on the training responses of this dataset alone. Held-out responses are not used for fitting or selection and are read only by the evaluator after the predictions are exported. Pearson-Δ and RMSE are point estimates averaged over conditions, and the uncertainty of $f_{\mathrm{comp}} - f_{\mathrm{single}}$ is a 10,000-replicate bootstrap over the seven held-out conditions. This panel tests whether the formulation carries over to a new dataset after refitting. Zero-shot transfer with fixed parameters and strict both-unseen prediction on genetic screens are outside its scope.

Table A.2: Role of each dataset. Each dataset serves one role, and results from the transfer and scope analyses are not used as evidence for the main mechanism claim.

Dataset	Role	Evidence provided	Not evidence for
Norman	Primary benchmark	Strict both-unseen performance on the official benchmark splits, over several metrics	Causal attribution of the gain to what the model learns
GSE116198	Primary mechanism test	Strong DirectJoint controls, the model-class by supervision factorial and the pair-residual analysis	Zero-shot transfer across studies with fixed parameters
GSE290153	Independent separation of factorization and reuse	Factorization with pairs only versus reuse of single responses, on a separate screen	A general claim that factorization or reuse replicates
GSE200201	Setting where direct models win	Whether a pair residual predicted from the descriptors recovers direct-pair performance	Superiority of constituent models over direct models
GSE284018	Development and held-out endpoint	Gamma selects the predictor, and the held-out Tau endpoint tests the fixed constituent map with calibration on training data	Replication on an independent dataset (Tau comes from the same screen), proof that pair supervision helps, or general superiority
Wessels, MCF10A, GSE90063	External scope	Descriptor-assignment controls on Wessels and MCF10A, and refitted formulation checks on all three screens	Pooling with the GSE116198 mechanism estimate
ComboSciPlex	Transfer of the formulation after refitting	Pair-supervision gain after refitting on this dataset	Strict both-unseen prediction on genetic screens or transfer with fixed parameters

A.4 Data and split sizes

Table A.3 gives the sizes that define each evaluation set, without listing individual genes. The GEARS strata are defined over the training and validation perturbations, whereas our models are fitted on the training conditions only. Every SEEN-0 pair is therefore unseen in fitting, and in 15 of the 54 official SEEN-2 instances and 332 of the 1,263 further ones at least one single is a validation condition that our models do not fit on. Norman is fitted on condition centroids rather than individual cells. The official Norman data have 5,045 readout genes and 277 conditions after filtering (284 before), and the GO matrix has 1,250 term columns and 105 target rows. It holds the direct annotations of the human UniProt annotation file of the 2026-08-05 release, without propagation to ancestor terms and excluding annotations with the NOT qualifier, with gene symbols matched exactly. The 1,250 terms (862 biological process, 229 molecular function, 159 cellular component) are all terms annotated to at least one of the 100 target genes with annotations, and the matrix has 2,672 nonzero entries. The auxiliary Norman panel counts are the exact counts of that panel. The 70 further splits (Norman, seeds 16 to 85) and the 34 resplits of the auxiliary Norman panel (seeds 5 to 38) used in the robustness analyses share genes and pairs, so they are not independent biological replicates. For the three official Norman splits, the gene-cluster analysis below reports pair overlap directly and clusters the uncertainty by held-out gene.

A.5 Complete Norman metrics

Tables A.4 and A.5 complete Table 1 with the RMSE, Pearson-Δ on the top 20 differentially expressed genes (T20-Δ) and centroid accuracy (CA). The 70 further splits are GEARS simulation splits with seeds 16 to 85, generated and filtered in the same way as the official splits. They were fixed before any model was scored on them, and every split is reported. Matching Mean falls back to Perturbed Mean in SEEN-0 and is omitted there. The biolord model was run with its released Norman recipe (Appendix G). Over the 70 further splits, the paired differences DACL-C minus biolord are $+0.088$ ($[0.059, 0.116]$) on SEEN-0 and -0.002 ($[-0.010, 0.006]$) on SEEN-2, with 95% t intervals over splits.

Table A.3: Data and split sizes. Counts are conditions, target genes or pairs as indicated, and a dash marks a quantity that does not apply. The descriptor column gives the response-free input and its fallback for missing descriptors.

Dataset	Size	Outer splits and held-out set	Training and coverage	Descriptor
Norman, seeds 1/2/3	5,045 genes; 277 filtered conditions (284 raw)	train/val/test $139/31/107$, $140/20/117$, $133/35/109$; scored SEEN-0/SEEN-2 $9/19$, $12/16$, $4/19$	68 training singles per seed; eligible training pairs $39/35/28$; GO has 1,250 terms and 105 target rows	GO; zero-centered-response fallback
Norman, auxiliary panel, seeds 1/2/3	Strict both-unseen pairs	Test pairs $17/12/11$	Training pairs $36/36/45$; eligible pairs $32/34/42$	GO; zero-centered-response fallback
GSE200201 MOLM13	36,717 measured genes; 20 perturbed genes; 27 eligible pairs	Seven outcome-free graph splits; every eligible pair scored once	10–19 training pairs per split; 27 pooled unique test pairs	GO for 19/20 perturbed genes; zero-centered-effect fallback for one
GSE290153 (ionizing radiation)	Independent genetic screen; two response channels	Five target-disjoint panels; 92 held-out genes per panel	4,186 test pairs per panel; four-cell comparison of model class and supervision	5,387-dimensional sparse GO membership

Table A.4: Norman results on the three official splits, mean $\pm$ s.d. over splits. Lower RMSE and higher T20-Δ and CA are better. Bold marks the best value in each column.

	SEEN-0			SEEN-2		
Method	RMSE$\times$100	T20-Δ	CA	RMSE$\times$100	T20-Δ	CA
Perturbed Mean	6.70 ± 0.33	-0.35 ± 0.26	0.50 ± 0.00	7.15 ± 1.24	-0.30 ± 0.07	0.50 ± 0.00
Matching Mean	–	–	–	4.92 ± 0.89	0.83 ± 0.07	0.84 ± 0.07
Descriptor-free additive	6.53 ± 0.18	0.58 ± 0.14	0.50 ± 0.00	7.01 ± 1.19	0.39 ± 0.07	0.50 ± 0.00
GEARS 0.1.2	7.56 ± 0.37	0.40 ± 0.37	$\mathbf{0.66\pm0.11}$	6.84 ± 1.16	0.76 ± 0.08	0.89 ± 0.03
biolord	6.32 ± 0.27	0.57 ± 0.15	0.64 ± 0.15	4.58 ± 0.79	$\mathbf{0.90\pm0.08}$	0.95 ± 0.04
scGPT	7.00 ± 0.22	0.20 ± 0.08	0.55 ± 0.06	5.69 ± 0.99	0.76 ± 0.07	0.84 ± 0.10
Scouter	8.49 ± 0.66	0.39 ± 0.36	0.65 ± 0.06	6.28 ± 0.77	$\mathbf{0.90\pm0.06}$	$\mathbf{0.96\pm0.04}$
Response decomp., Ridge	6.79 ± 0.27	$\mathbf{0.65\pm0.12}$	0.51 ± 0.03	7.09 ± 1.96	0.53 ± 0.09	0.57 ± 0.06
Response decomp., MLP	6.92 ± 0.23	0.59 ± 0.16	0.52 ± 0.02	7.24 ± 2.10	0.50 ± 0.12	0.58 ± 0.04
DepMap Ridge	6.59 ± 0.20	0.63 ± 0.11	0.50 ± 0.01	6.77 ± 1.17	0.56 ± 0.07	0.53 ± 0.01
DepMap MLP	6.89 ± 0.52	0.42 ± 0.33	0.57 ± 0.09	5.98 ± 0.95	0.72 ± 0.02	0.76 ± 0.05
DirectJoint Ridge	5.98 ± 0.56	$\mathbf{0.65\pm0.15}$	0.62 ± 0.14	$\mathbf{4.12\pm1.07}$	$\mathbf{0.90\pm0.07}$	0.94 ± 0.06
DirectJoint MLP	5.92 ± 0.59	0.63 ± 0.18	0.65 ± 0.11	4.65 ± 0.58	0.84 ± 0.11	0.88 ± 0.06
DACL-S (singles)	5.94 ± 0.52	0.59 ± 0.18	$\mathbf{0.66\pm0.14}$	4.35 ± 0.64	$\mathbf{0.90\pm0.07}$	0.94 ± 0.04
DACL-C (singles + pairs)	$\mathbf{5.87\pm0.61}$	0.60 ± 0.17	0.65 ± 0.18	4.26 ± 0.94	$\mathbf{0.90\pm0.07}$	0.94 ± 0.05

A.6 GENE-CLUSTER INTERVALS FOR THE NORMAN CONTRASTS

Split-level intervals treat splits as independent, but the 70 further splits reuse the same genes. Table A.7 therefore also resamples genes. Repeated appearances of each unordered pair are averaged. Genes are resampled with replacement, and each pair is weighted by the product of the sampled multiplicities of its two genes. The 95% interval is the percentile interval of the weighted mean over 100,000 replicates (50,000 for the 70 further splits). The last column reports the range of the remaining mean after removing all pairs involving each gene in turn. This procedure also reproduces the interval reported for the pair-supervision increment on the official splits.

Table A.5: Norman results on the 70 further splits, mean ± s.d. over splits, as in Table A.4. Response decomposition was run on the official splits only.

	SEEN-0			SEEN-2		
Method	RMSE×100	T20-Δ	CA	RMSE×100	T20-Δ	CA
Perturbed Mean	7.15±1.60	−0.26±0.25	0.51±0.06	6.86±0.76	−0.26±0.15	0.50±0.00
Matching Mean	–	–	–	4.88±0.55	0.75±0.09	0.80±0.05
Descriptor-free additive	7.14±1.47	0.42±0.23	0.51±0.06	6.84±0.68	0.43±0.12	0.50±0.00
GEARS 0.1.2	8.66±1.80	0.39±0.22	0.56±0.08	7.18±1.44	0.70±0.10	0.82±0.08
biolord	7.69±1.52	0.38±0.23	0.57±0.14	4.51±0.53	**0.89±0.06**	**0.94±0.04**
scGPT	–	–	–	–	–	–
Scouter	8.83±1.05	0.40±0.21	**0.60±0.13**	6.35±0.50	0.88±0.07	**0.94±0.04**
Response decomp., Ridge	–	–	–	–	–	–
Response decomp., MLP	–	–	–	–	–	–
DepMap Ridge	7.33±1.41	0.38±0.22	0.51±0.07	6.69±0.69	0.56±0.11	0.53±0.03
DepMap MLP	7.53±1.44	0.29±0.23	0.52±0.09	5.88±0.69	0.68±0.08	0.76±0.07
DirectJoint Ridge	6.73±1.34	**0.50±0.20**	0.58±0.10	4.21±0.56	**0.89±0.06**	0.90±0.06
DirectJoint MLP	6.68±1.35	**0.50±0.20**	0.58±0.10	4.48±0.53	0.84±0.07	0.87±0.07
DACL-S (singles)	**6.67±1.37**	**0.50±0.19**	0.58±0.10	4.38±0.48	0.88±0.05	0.90±0.07
DACL-C (singles + pairs)	6.68±1.34	**0.50±0.19**	0.59±0.10	**4.12±0.45**	**0.89±0.06**	0.92±0.05

Table A.6: Recorded wall-clock time to fit one Norman split and predict its held-out pairs. Hardware differs between methods, so the times indicate orders of magnitude rather than a controlled benchmark.

Method	Hardware	Time per split
DACL-C, DACL-S, DirectJoint Ridge and descriptor-free, with model selection	4 CPU threads	2–4 s
biolord, released Norman recipe	8 CPU threads	22–31 s
Scouter, released Norman script	one A100 GPU	14–20 min
GEARS 0.1.2	RTX PRO 6000, shared by five runs	78–113 min
scGPT, perturbation tutorial	one A6000 or A4000 GPU	3.1–5.5 h

A.7 TARGET-DISJOINT PILOT TESTS

The pilot tests in Figure 1(a) were target-disjoint, so the held-out target's response was not used for fitting or model selection. Three tests asked whether additional response structure can be retrieved for a held-out target from variables available at test time. The fourth asked whether relearning the final composition step improves on an additive prediction that transfers to new targets. Pair supervision of the shared map is motivated analytically, by the pair-risk identity in Section 3.3. The table below gives the aggregate result behind each design constraint. The descriptor-by-capacity factorial is a separate mechanism analysis, reported in Table B.1.

A.8 MATCHED CONTRASTS OVER REPEATED SPLITS

The table lists every prespecified contrast. The gene-cluster analysis below quantifies one source of dependence between splits, but it does not turn repeated resplits into independent biological evidence.

Uncertainty for mechanism contrasts. Mechanism contrasts are resampled over the unit implied by the split design, namely held-out genes for the GSE116198 and GSE290153 factorials and gene clusters for the paired Norman analysis. Each interval therefore refers to the contrast being interpreted, not to an unpaired difference of aggregate means. We use the interval to state whether the data identify a mechanism, separate two settings or leave the comparison unresolved. Point estimates alone are not treated as evidence for an architecture.

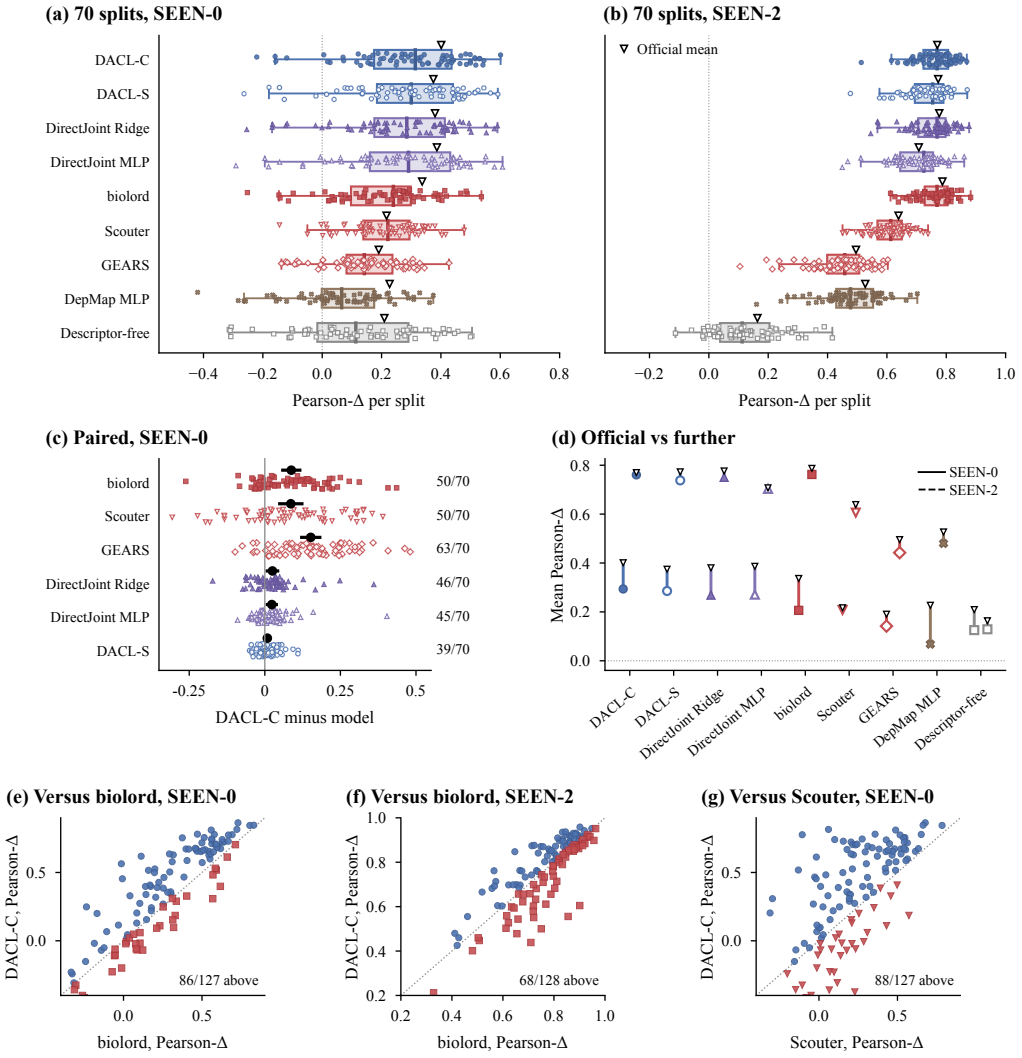


Figure A.1: Every Norman split and every unique held-out pair. (a, b) Pearson-Δ of each method on each of the 70 further splits (points, with boxes for quartiles), with the mean over the three official splits as an open triangle. (c) Per-split difference DACL-C minus each comparator on SEEN-0, with the mean and 95% t interval in black and the number of splits in which DACL-C is higher. (d) Mean over the official splits (triangle) against the mean over the further splits (coloured marker), for SEEN-0 (solid) and SEEN-2 (dashed). (e–g) DACL-C against biolord and Scouter for every unique pair of the further splits, averaged over repeated appearances. Pairs where DACL-C is higher are blue and counted in the corner. The others use the comparator's marker.

B DESCRIPTOR ASSIGNMENT AND ALTERNATIVE EXPLANATIONS

B.1 DESCRIPTOR AND CAPACITY CONTROLS

This experiment was prespecified as a $4 \times 2 \times 2$ factorial over descriptor, capacity of the constituent map and supervision. All rows use the same Norman evaluation, the three official splits and separate aggregation for SEEN-0 and SEEN-2, and held-out responses are not used for fitting or selection. Ridge uses a fixed α grid. The nonlinear variant is a shared one-hidden-layer constituent map with 32 hidden units and an 8-dimensional response basis, and its optimizer and weight decay are fixed. Pair supervision always enters as $f(q_a) + f(q_b)$, with no pair identifier or direct pair decoder. Values are mean $\pm$ sample standard deviation over splits. One descriptor level of the design permuted GO

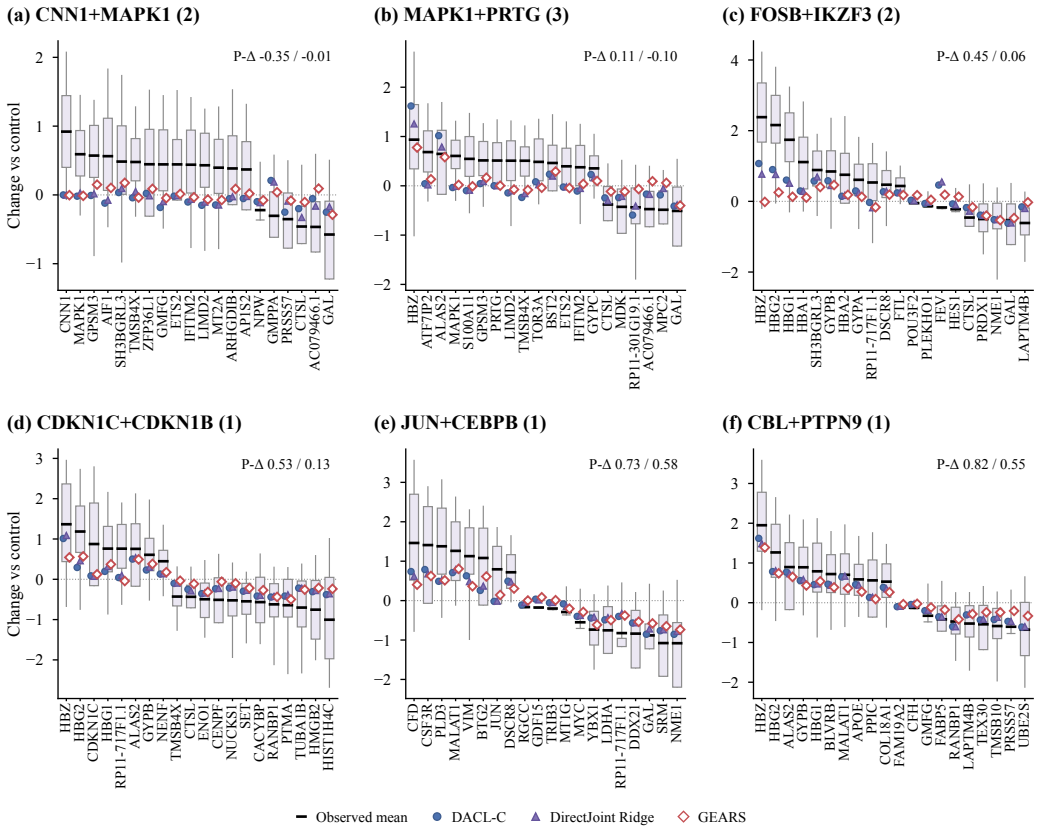


Figure A.2: Held-out pairs of unseen genes across the range of DACL-C scores. The panels are the official SEEN-0 instances ranked 1, 6, 11, 15, 20 and 25 of 25 by DACL-C Pearson-Δ (split in parentheses), and the median instance is in Figure 2(a). Each panel shows the 20 GEARS top-DE genes of the pair. Boxes and whiskers give the 25–75% and 5–95% ranges of the change from control across cells, bars the observed mean change and markers the predictions. The corner values are the all-gene Pearson-Δ of DACL-C and GEARS for the instance.

feature columns. This permutation does not break the assignment of terms to genes, so it is not a valid negative control. We exclude its rows and use the separately specified target-row permutation of Table B.4 as the control.

The full factorial comprises 48 fits, of which the 36 with an interpretable descriptor are reported. The MLP has 328 effective parameters with the constant descriptor, 29,128–32,040 with GO and 37,992 with DepMap. For Ridge the number of effective coefficients varies with the kernel rank and is recorded per split. An evaluator-only oracle that substitutes the held-out single responses of both genes exists for 25 SEEN-0 pairs across splits. We do not report this oracle for SEEN-2, whose singles can occur in training or validation. For Ridge with true GO, trained on singles and pairs, the mean squared constituent error, interaction residual and pair error over these 25 pairs were 0.00468, 0.00170 and 0.00364, respectively, and the pair-error identity held to machine precision. The oracle was never used for fitting, model selection or choosing which model to report.

Descriptor augmentation. The descriptor-augmentation experiment tests whether adding information to the descriptor helps. It keeps the Norman evaluation, official splits, strata, Ridge map, inner-fold selection of α, pair weight, inference rule and evaluator fixed, and held-out responses are not used for fitting or selection. It compares GO alone, DepMap-24Q4 alone and the concatenated GO+DepMap descriptor. Missing DepMap profiles in the concatenated descriptor are replaced by the mean over available training genes, and GO coverage and the centered fallback are unchanged. The experiment comprises three descriptors $\times$ one model class $\times$ two supervision regimes $\times$ three splits

Table A.7: Paired contrasts on Norman with intervals over resampled genes. Values are Pearson-Δ differences of DACL-C minus the comparator, averaged over unique pairs, with both models using GenePT in the GenePT row. Positive counts unique pairs with a positive difference.

Comparator	Mean	95% interval	Positive	Leave one gene out
Official splits, SEEN-0				
GEARS	$+0.216$	$[+0.012, +0.397]$	19/23	$[+0.189, +0.281]$
biolord	$+0.075$	$[-0.086, +0.249]$	16/23	$[+0.032, +0.125]$
Scouter	$+0.175$	$[+0.001, +0.403]$	19/23	$[+0.117, +0.222]$
scGPT	$+0.340$	$[+0.061, +0.594]$	19/23	$[+0.306, +0.439]$
Descriptor-free	$+0.187$	$[+0.019, +0.440]$	18/23	$[+0.149, +0.238]$
DepMap MLP	$+0.222$	$[+0.016, +0.411]$	18/23	$[+0.193, +0.289]$
Response decomposition	$+0.181$	$[-0.041, +0.427]$	17/23	$[+0.137, +0.250]$
DirectJoint Ridge	$+0.005$	$[-0.055, +0.056]$	14/23	$[-0.008, +0.019]$
DirectJoint MLP	$+0.003$	$[-0.074, +0.093]$	13/23	$[-0.012, +0.027]$
DACL-S	$+0.026$	$[-0.037, +0.120]$	15/23	$[+0.010, +0.051]$
70 further splits, SEEN-0				
DirectJoint Ridge	$+0.035$	$[+0.009, +0.064]$	91/127	$[+0.031, +0.039]$
DirectJoint MLP	$+0.020$	$[+0.002, +0.041]$	77/127	$[+0.018, +0.023]$
biolord	$+0.079$	$[+0.016, +0.143]$	86/127	$[+0.069, +0.087]$
Scouter	$+0.095$	$[-0.023, +0.218]$	88/127	$[+0.070, +0.122]$
GEARS	$+0.166$	$[+0.076, +0.251]$	93/127	$[+0.152, +0.181]$
DepMap MLP	$+0.241$	$[+0.140, +0.348]$	106/127	$[+0.215, +0.255]$
Descriptor-free	$+0.163$	$[+0.085, +0.243]$	100/127	$[+0.146, +0.174]$
DACL-S	$+0.006$	$[-0.011, +0.024]$	70/127	$[+0.003, +0.009]$
DirectJoint Ridge (GenePT)	$+0.021$	$[-0.027, +0.057]$	84/127	$[+0.015, +0.030]$
70 further splits, SEEN-2				
DirectJoint Ridge	$+0.010$	$[-0.000, +0.020]$	74/128	$[+0.008, +0.011]$
DirectJoint MLP	$+0.057$	$[+0.038, +0.082]$	111/128	$[+0.053, +0.060]$
biolord	$+0.000$	$[-0.036, +0.032]$	68/128	$[-0.004, +0.005]$
Scouter	$+0.153$	$[+0.122, +0.183]$	123/128	$[+0.148, +0.157]$
GEARS	$+0.318$	$[+0.271, +0.366]$	127/128	$[+0.312, +0.326]$
DACL-S	$+0.025$	$[+0.003, +0.047]$	75/128	$[+0.019, +0.028]$

$\times$ two strata. The predictions with GO alone reproduce the main DACL predictions exactly. The combined descriptor does not improve absolute performance over GO alone, whereas its pair-supervision gain within the same map remains positive on average. This is a limited test of the descriptor error floor, not a proposed model.

With GO+DepMap, pair supervision adds $+0.0429$ in SEEN-0 and $+0.0905$ in SEEN-2, so the gain is positive in both strata. The lower absolute score of the concatenated descriptor shows that the population error floor does not guarantee that an arbitrary union of features helps at finite sample size. The positive gain within the same map suggests that pair supervision and descriptor quality contribute separately.

GenePT descriptor. The descriptor-swap experiment replaces the Gene Ontology membership vector by the 1536-dimensional GenePT embedding of each gene's National Center for Biotechnology Information (NCBI) summary (Chen & Zou, 2025). The embedding covers 104 of the 105 Norman target genes, and the uncovered gene uses the same zero-vector fallback as a gene without Gene Ontology terms. The kernel standardization, target-disjoint ridge selection, pair weight, inference rule, DirectJoint features and evaluator are unchanged. With the original Gene Ontology descriptor, the modified implementation reproduces the original DACL and DirectJoint predictions exactly, so the swap changes only the descriptor.

Target-assignment control. The target-assignment control applies a deterministic fixed-point-free permutation to the 100 GO target rows with annotations, while the five genes without annotations keep the zero fallback. It preserves the multiset of rows, row sparsities, column marginals and availability mask, but gives every target the terms of another gene. The training Gram matrix and the

Table A.8: Pilot tests behind the design constraints. In every test the held-out target's response was not used for fitting or model selection. Each result is shown in the direction used for the design decision.

Candidate	Target-disjoint question	Observed result	Design constraint
Learned target-specific shape	Whether a learned shape branch can be retrieved for a held-out target	Centered energy distance (lower is better): learned 0.1395 versus control shape 0.0991 on the target-disjoint test.	Do not freely predict higher-order response shape from the information available for the target.
Graph propagation	Whether graph proximity transfers as a way to locate responses	Propagation 0.1020 versus static GO 0.0945; $\Delta = -0.0076$.	Graph proximity is not itself a response label.
Susceptibility/dosage	Whether oracle headroom along a response axis yields a target-disjoint predictor	Gain falls from $+0.0077$ with the oracle to -0.0006 on held-out targets.	Oracle headroom alone does not make a target predictable at test time.
Flexible correction	Whether a learned correction improves a transferable additive prediction	Corrected 0.2316 versus additive 0.3963; $\Delta = -0.1647$.	Keep a strong transferable additive prediction unless a matched control supports relearning it.

Table A.9: Matched contrasts over the repeated splits of the robustness analyses. Values are the mean difference $\Delta \pm$ sample standard deviation, with the 95% t interval over resplits. The repeated splits reuse the same data and overlapping held-out sets, so they are not independent biological replicates.

Contrast	Mean $\Delta\pm$SD	95% interval
SEEN-0: $f_{\text{comp}} - f_{\text{single}}$	$+0.0083\pm0.0341$	$[0.0002, 0.0164]$
SEEN-2: $f_{\text{comp}} - f_{\text{single}}$	$+0.0225\pm0.0286$	$[0.0157, 0.0294]$
Norman auxiliary panel: $f_{\text{comp}} - f_{\text{single}}$	$+0.0137\pm0.0255$	$[0.0048, 0.0226]$
SEEN-0: $f_{\text{comp}} - \text{Ridge}$	$+0.0252\pm0.0728$	$[0.0078, 0.0425]$
SEEN-0: $f_{\text{comp}} - \text{MLP}$	$+0.0232\pm0.0628$	$[0.0082, 0.0382]$
SEEN-0: $f_{\text{comp}} - U_{\text{free}}$	$+0.0641\pm0.1232$	$[0.0348, 0.0935]$
SEEN-2: $f_{\text{comp}} - \text{Ridge}$	$+0.0098\pm0.0433$	$[-0.0006, 0.0201]$
SEEN-2: $f_{\text{comp}} - \text{MLP}$	$+0.0571\pm0.0518$	$[0.0447, 0.0694]$
SEEN-2: $f_{\text{comp}} - U_{\text{free}}$	$+0.4943\pm0.1218$	$[0.4653, 0.5234]$

test-to-training Gram matrix differ from those of true GO on every split. Table B.4 shows that the descriptor content is not sufficient without the correct assignment of descriptors to genes.

Target-assignment control on external screens. We applied the same target-row derangement to every original Wessels and MCF10A split. The permutation has no fixed points and preserves the multiset of GO rows and the column marginals, and the model class, held-out responses and evaluator are unchanged. The predictions with true GO reproduce the original predictions exactly. This control was run after the main external analyses, so we treat it as supporting evidence rather than as an independent confirmation.

Under single supervision, which gives the primary contrast, the stratified condition-bootstrap 95% intervals are $[0.0157, 0.0427]$ on Wessels and $[-0.0041, 0.0121]$ on MCF10A, so the effect is clearer on Wessels than on MCF10A.

Pair-information controls. The pair-information control fixes true GO, Ridge, the selection of α on training singles, the identities and number of eligible training pairs and the total weight of the pair block in the objective. For the three arms with pair rows, only the target attached to each pair row in the loss changes. Observed targets reproduce the DACL-C result exactly. The shuffled

Table B.1: Factorial comparison of descriptor, capacity and supervision on Norman, mean ± sample standard deviation over the official splits. P-Δ is Pearson-Δ relative to the perturbation mean, averaged over conditions (higher is better), and RMSE is lower-is-better. Rows for the feature-column permutation are excluded because that permutation is not a valid control. The DepMap rows belong to this factorial and are not the DepMap baselines of Table 1, which use a different fitting recipe.

Descriptor	Capacity	Supervision	SEEN-0		SEEN-2	
			P-Δ	RMSE	P-Δ	RMSE
Constant	Ridge	Single	0.2095±0.1815	0.0653±0.0018	0.1637±0.0337	0.0701±0.0119
Constant	Ridge	Single+Pair	0.2095±0.1815	0.0653±0.0018	0.1637±0.0337	0.0701±0.0119
Constant	MLP	Single	0.2095±0.1815	0.0653±0.0018	0.1637±0.0337	0.0701±0.0119
Constant	MLP	Single+Pair	0.1792±0.1975	0.0654±0.0021	0.1879±0.0486	0.0692±0.0114
True GO	Ridge	Single	0.3746±0.2102	0.0594±0.0052	0.7729±0.0688	0.0435±0.0064
True GO	Ridge	Single+Pair	0.4010±0.2112	0.0587±0.0061	0.7699±0.0747	0.0426±0.0094
True GO	MLP	Single	0.3402±0.2015	0.0618±0.0050	0.6813±0.0601	0.0511±0.0053
True GO	MLP	Single+Pair	0.3341±0.2621	0.0609±0.0064	0.6562±0.0761	0.0498±0.0082
DepMap-24Q4	Ridge	Single	0.2143±0.1660	0.0657±0.0018	0.2232±0.0384	0.0691±0.0121
DepMap-24Q4	Ridge	Single+Pair	0.3059±0.1462	0.0637±0.0035	0.3217±0.0768	0.0663±0.0142
DepMap-24Q4	MLP	Single	0.1617±0.0736	0.0763±0.0032	0.3851±0.0315	0.0672±0.0121
DepMap-24Q4	MLP	Single+Pair	0.1476±0.0667	0.0805±0.0100	0.2976±0.1624	0.0685±0.0180

Table B.2: Descriptor augmentation on Norman. Values are mean ± sample standard deviation over the three official splits. Pearson-Δ is higher-is-better and RMSE is lower-is-better. The prespecified success criterion, that GO+DepMap with single supervision exceeds GO alone in both strata, was not met.

Descriptor	Supervision	SEEN-0		SEEN-2	
		P-Δ	RMSE	P-Δ	RMSE
GO	Single	0.3746±0.2102	0.0594±0.0052	0.7729±0.0688	0.0435±0.0064
GO	Single+Pair	0.4010±0.2112	0.0587±0.0061	0.7699±0.0747	0.0426±0.0094
DepMap-24Q4	Single	0.2143±0.1660	0.0657±0.0018	0.2232±0.0384	0.0691±0.0121
DepMap-24Q4	Single+Pair	0.3059±0.1462	0.0637±0.0035	0.3217±0.0768	0.0663±0.0142
GO+DepMap	Single	0.3234±0.2312	0.0603±0.0054	0.5795±0.2243	0.0509±0.0056
GO+DepMap	Single+Pair	0.3662±0.2330	0.0590±0.0062	0.6700±0.1021	0.0481±0.0034

control reassigns targets across pairs by a deterministic fixed-point-free derangement and preserves the multiset of targets. The synthetic control uses $\beta_a + \beta_b$ computed from training singles.

Sensitivity to the pair weight and effective capacity. This analysis varies only the pair weight over $\lambda \in \{0, 0.25, 0.5, 1, 2\}$ on the three official Norman splits. GO features, training-pair identities and targets, target-disjoint selection of α, support, inference and evaluation are fixed. Every level is reported, and no level is selected from these results. Table B.7 gives Pearson-Δ and RMSE. A positive pair weight improves SEEN-0 Pearson-Δ at all four values tested, but the gain levels off and then declines. In SEEN-2, small weights improve Pearson-Δ whereas $\lambda = 1$ and 2 do not, so performance does not improve monotonically with the pair weight.

The capacity analysis separates the number of parameters from the complexity of the regularized fit. For splits 1/2/3, the effective ranks of the kernel and of the joint design remain $67/65/64$ at all five weights, the parameter counts remain $338{,}015/327{,}925/322{,}880$ and the selected α values remain $0.3/0.3/0.1$. The effective degrees of freedom are $42.49/42.60/52.12$ at $\lambda = 0$, $45.83/45.75/53.92$ at 0.25, $47.32/47.08/54.62$ at 0.5, $48.88/48.49/55.33$ at 1, and $50.32/49.82/55.97$ at 2. Pair supervision therefore changes the fitted solution and its effective degrees of freedom without adding feature columns or parameters. The fitted solutions satisfy the closed-form update in Eq. (6) to a maximum absolute error below 1.1×10^{-14}, and no test response was accessed during fitting.

Table B.3: Gene Ontology and GenePT descriptors on Norman. Pearson-Δ, mean $\pm$ s.d. over splits. Contrasts are paired over the 70 further splits with 95% t intervals.

	Official splits ($n = 3$)		Further splits ($n = 70$)	
Method	SEEN-0	SEEN-2	SEEN-0	SEEN-2
Descriptor-free additive	0.209±0.182	0.164±0.034	0.125±0.207	0.129±0.111
Gene Ontology				
DirectJoint Ridge	0.380±0.221	0.776±0.092	0.269±0.189	0.751±0.072
DACL-S	0.375±0.210	0.773±0.069	0.285±0.186	0.738±0.070
DACL-C	0.401±0.211	0.770±0.075	0.294±0.186	0.760±0.065
GenePT				
DirectJoint Ridge	0.302±0.221	0.515±0.050	0.225±0.167	0.618±0.146
DACL-S	0.322±0.272	0.513±0.083	0.212±0.200	0.503±0.155
DACL-C	0.338±0.237	0.646±0.019	0.250±0.189	0.635±0.107
GenePT contrasts, 70 further splits				
DACL-C − DirectJoint Ridge			+0.025 [0.004, 0.046]	+0.017 [−0.012, 0.046]
DACL-C − DACL-S			+0.039 [0.024, 0.053]	+0.132 [0.115, 0.149]
Gene Ontology contrasts, 70 further splits				
DACL-C − DirectJoint Ridge			+0.025 [0.008, 0.043]	+0.010 [−0.001, 0.020]

Table B.4: Target-assignment control on Norman. Values are Pearson-Δ, mean $\pm$ sample standard deviation over the three official splits.

Descriptor	Supervision	SEEN-0	SEEN-2
True GO	Single	0.3746±0.2102	0.7729±0.0688
Target-row-permuted GO	Single	0.2043±0.1765	0.3200±0.0437
True GO	Single+Pair	0.4010±0.2112	0.7699±0.0747
Target-row-permuted GO	Single+Pair	0.1930±0.1577	0.5168±0.0859

C COMPARATORS, TRANSFER AND THE PRIMARY MECHANISM SCREEN

C.1 RELEASED COMPARATORS AND TRANSFER AFTER REFITTING

Table A.4 reports the RMSE of these models from the same evaluator as every other row, averaged over conditions.

C.2 STRICT BOTH-UNSEEN RESULT ON GSE116198

The GSE116198 analysis tests the constituent map and treats pair supervision as a separate refinement. GSE116198 is a human CRISPRi genetic-interaction screen with 467 retained genes, 108,799 distinct matched gene pairs and the matched single-gene conditions paired with negative controls. Its five splits hold out 20% of genes, chosen by a fixed pseudorandom rule that does not use outcomes. They remove every training pair that involves a held-out gene and score only pairs whose two genes are both held out. The primary response is the released Triple Sequencing phenotype, with two replicates in each of Jurkat and K562 cells. GO covers 430 of the 467 genes, and the remaining genes stay in the evaluation with a zero-vector fallback. Predictions were exported before scoring.

Table B.5: External target-assignment control. Values are Pearson-Δ, mean $\pm$ sample standard deviation over the original splits. Δ_q is true GO minus target-row-permuted GO.

Panel	Supervision	True GO	Permuted GO	Δ_q
Wessels	Single	−0.1782±0.1579	−0.2067±0.1302	+0.0284±0.0333
Wessels	Single+Pair	−0.1649±0.1664	−0.2089±0.1292	+0.0440±0.0484
MCF10A	Single	0.0354±0.0410	0.0313±0.0359	+0.0041±0.0051
MCF10A	Single+Pair	0.0336±0.0449	0.0292±0.0370	+0.0044±0.0099

Table B.6: Pair-information controls at fixed pair support and weight. Scores are Pearson-Δ, mean $\pm$ sample standard deviation, and Δ is the change relative to training on singles only.

Pair target in training loss	SEEN-0 score	SEEN-0 Δ	SEEN-2 score	SEEN-2 Δ
None	0.3746 ± 0.2102	0.0000	0.7729 ± 0.0688	0.0000
Observed pair response	0.4010 ± 0.2112	$+0.0265 \pm 0.0026$	0.7699 ± 0.0747	-0.0030 ± 0.0063
Shuffled pair response	0.2895 ± 0.1305	-0.0851 ± 0.1153	0.3119 ± 0.1261	-0.4610 ± 0.0684
Synthetic additive response	0.3840 ± 0.1801	$+0.0094 \pm 0.0327$	0.7859 ± 0.0602	$+0.0130 \pm 0.0088$

Table B.7: Sensitivity to the pair weight on Norman at fixed support. Values are mean $\pm$ sample standard deviation over the three official splits. Δ_0 is the paired change in Pearson-Δ relative to $\lambda = 0$. All weights in the grid are shown.

λ	Pearson-Δ	Δ_0	RMSE
SEEN-0			
0	0.3746 ± 0.2102	0.0000 ± 0.0000	0.0594 ± 0.0052
0.25	0.3971 ± 0.2120	$+0.0225 \pm 0.0047$	0.0587 ± 0.0058
0.5	0.4009 ± 0.2110	$+0.0263 \pm 0.0026$	0.0586 ± 0.0059
1	0.4010 ± 0.2112	$+0.0265 \pm 0.0026$	0.0587 ± 0.0061
2	0.3987 ± 0.2127	$+0.0241 \pm 0.0086$	0.0589 ± 0.0063
SEEN-2			
0	0.7729 ± 0.0688	0.0000 ± 0.0000	0.0435 ± 0.0064
0.25	0.7911 ± 0.0761	$+0.0182 \pm 0.0084$	0.0409 ± 0.0080
0.5	0.7839 ± 0.0761	$+0.0111 \pm 0.0079$	0.0414 ± 0.0087
1	0.7699 ± 0.0747	-0.0030 ± 0.0063	0.0426 ± 0.0094
2	0.7520 ± 0.0731	-0.0209 ± 0.0060	0.0444 ± 0.0096

D SUPERVISION REUSE AND PAIR-SPECIFIC CAPACITY

Training-only diagnostic. The diagnostic of Section 3.3 was computed on the ten GSE116198 panels and seven GSE200201 splits from training responses only, with target-disjoint folds. Gene Ontology predicts held-out centered single effects better than the fold mean in both screens. The MSE reduction relative to the fold mean (positive values favor the descriptor) is 0.0382 ± 0.0401 on GSE116198 (8 of 10 panels positive) and 0.0151 ± 0.0353 on GSE200201 (5 of 7). Pair residuals are not predictable from the descriptors in either screen, with MSE changes of -1.2177 ± 0.2474 and -0.0536 ± 0.0745 relative to the fold mean. Pair support differs sharply, with 52,271.9 and 13.1 training pairs per panel whose two genes have descriptors. The diagnostic therefore separates the two screens by pair support only. It does not show a descriptor-predictable residual on GSE200201, where the direct model is ahead, and it is not used to select models.

In the main comparison on this screen, DACL-S has a higher pooled Pearson than DACL-C. The contrast between DACL-C and DirectJoint Ridge is slightly negative in one cell-line-by-split panel (-0.0076), and no panel, split or cell line was removed. The results support a difference in how the two model classes reuse single supervision in this setting. They leave open whether DACL is better

Table C.1: Released response-decomposition models run by us on the official Norman splits. Values are Pearson-Δ, mean $\pm$ sample standard deviation over the three official splits. These are our runs of the released code, not published scores, and they are distinct from the DepMap rows of the main Norman table.

Method	SEEN-0	SEEN-2	DACL-C $-$ method (SEEN-0)	DACL-C $-$ method (SEEN-2)
Response decomp., Ridge	0.2418 ± 0.0999	0.2623 ± 0.0885	$+0.1592$	$+0.5076$
Response decomp., MLP	0.1985 ± 0.1054	0.2305 ± 0.1219	$+0.2026$	$+0.5394$
DACL-C	$\mathbf{0.4010 \pm 0.2112}$	$\mathbf{0.7699 \pm 0.0747}$	–	–

Table C.2: ComboSciPlex after refitting on this dataset, scored on seven held-out drug-combination conditions. Pearson-Δ is computed relative to the mean training response. The paired difference $f_{\text{comp}} - f_{\text{single}}$ is $+0.1549$, with condition-bootstrap 95% CI $[0.0160, 0.3180]$.

Method	Pearson-Δ	Pearson	RMSE
Train mean	0.0000	0.7030	0.0845
DACL-S	0.0150	0.7180	0.0554
DirectJoint Ridge	0.1505	0.8534	0.0427
DACL-C	**0.1699**	**0.8728**	**0.0402**

Table C.3: Strict both-unseen results on GSE116198. Pearson is averaged over the two Triple Sequencing replicate channels within each panel (cell line by split) and then over all ten panels. Every method uses the same GO descriptor, held-out-gene splits, target-disjoint selection, training responses and evaluator. Intervals are from a 10,000-replicate bootstrap over held-out genes.

Method	Pooled Pearson (channel mean)	Difference from DirectJoint Ridge	95% interval
DACL-S	0.168660	$+0.135421$	$[+0.085551, +0.190039]$
DACL-C	0.149730	$+0.116492$	$[+0.069079, +0.169439]$
DirectJoint Ridge	0.033238	–	–

in absolute error and whether it dominates direct predictors in general. The raw-feature nonlinear control and the matched direct head are reported in the main text.

The two additive models therefore gain almost equally from single supervision, and the difference in gain is small with an interval that includes zero. This control distinguishes the additive path through which single supervision reaches pair predictions from factorization itself. The richer DirectJoint comparison above shows that a direct model with additional interaction capacity can reach a comparable final Pearson without gaining the same amount from single rows. Trained on singles and pairs, the factorized model exceeds the sum-only model by $+0.009$, with a 95% bootstrap interval over held-out genes of $[-0.009, +0.028]$. The corresponding difference in gain, $+0.001$, has the separate interval reported in the table. The two intervals are not interchangeable, since one compares final predictors and the other compares the gain from single supervision.

With the same encoder, the pooled Pearson difference between the factorized and the direct model is slightly negative with an interval that crosses zero, and the direct model has lower pooled RMSE and MAE. The benefit reported for DACL therefore comes from how the two model classes reuse single responses, not from an advantage of factorization that holds regardless of architecture.

Table D.1: Choosing between the constituent and the direct model from training data. The selector compares DACL-C with DirectJoint Ridge using only training responses and target-disjoint inner folds, and test responses are read only after the selected predictions are exported. GSE200201 uses the prespecified aggregation over primary test conditions, weighted by primary pairs.

Screen	Always constituent	Always direct	Selector	Selector $-$ direct	Selected model
GSE116198	0.149730	0.033239	0.149730	$+0.116491$	constituent
GSE200201	0.088019	0.161802	0.108797	-0.053005	mixed

The selector chooses the constituent model throughout GSE116198 and both models across GSE200201 panels. In those panels, its aggregate score remains below that of always choosing the direct model. The selector is therefore a training-only model-choice rule for these settings, not a general-purpose selector. Differences in this and other tables are computed before rounding, so a value repeated across analyses can differ in its last printed digit.

Table D.2: Model class × supervision factorial on GSE116198. All four configurations use the same GO encoder, target-disjoint selection, optimizer and paired initialization. Pair-only models are fitted on pair rows only, and the other models add the same single rows. The reuse interaction is the difference-in-differences.

Model	Supervision	Pearson	RMSE	MAE
Factorized	Pair-only	0.119990	0.210062	0.167129
Direct	Pair-only	0.148036	0.163314	0.129109
Factorized	Single + pair	0.153972	0.193075	0.152666
Direct	Single + pair	0.148357	0.163188	0.128468
Factorized single-row gain	–	+0.033982	–	–
Direct single-row gain	–	+0.000321	–	–
Reuse interaction	–	+0.033660	–	–

The gain from single rows has a 95% bootstrap interval over held-out genes of $[+0.018, +0.048]$ for the factorized model and $[-0.010, +0.011]$ for the richer direct model. The reuse interaction has interval $[+0.016, +0.050]$. The final difference between the factorized and the richer direct model is $+0.006$ with interval $[-0.023, +0.034]$, and the direct model has lower pooled RMSE and MAE. The interaction and final difference are different estimands. The interaction measures the change caused by adding singles, whereas the final difference compares the two predictors trained on singles and pairs.

Table D.3: Parameter-matched sum-only direct control on GSE116198. The control uses the same 2263-dimensional GO input, 48-unit encoder, 108,770 trainable parameters, target-disjoint selection and paired initialization as the factorized model. Intervals are 95% bootstrap intervals over held-out genes, computed over the same ten panels.

Model	Pair-only Pearson	Single+pair Pearson	Gain from singles	95% interval for gain
Factorized	0.119990	0.153972	+0.033982	[+0.018404, +0.048423]
Sum-only Direct	0.112478	0.145284	+0.032807	[+0.015993, +0.045584]
Factorized − Sum-only Direct	+0.007512	+0.008687	+0.001175	[−0.015244, +0.020914]

D.1 Factorization and reuse on GSE290153

GSE290153 is an independent screen of responses to ionizing radiation (IR) with its own fixed splits, evaluated with the same four-cell comparison after the predictions were exported. Each of five target-disjoint panels contains 4,186 test pairs.

Residual-target permutation. To test pair-specific structure rather than extra capacity, we reassigned training-pair residual targets by 32 fixed-point-free cyclic shifts within each training panel. The additive predictions, residual features, panels and ridge penalty remained fixed. Relative to DACL-S, the true residual changes Pearson-Δ on GSE200201 by $+0.094$, whereas the permuted residuals change it by -0.010 on average over the 32 permutations. On GSE116198 the true residual lowers Pearson by 0.104 and the permuted residuals lower it by 0.093 on average. Reassigned residuals therefore give no mean gain in either screen. On GSE200201, only the correctly assigned residual helps.

E External screens and scope

E.1 Calibration on training data

The map sets the response pattern, but its raw output can have an incorrect amplitude or offset for a given screen. Where absolute error matters, we optionally apply a channel-wise affine map

$$\widetilde{\mu}_{ab,j} = s_j \widehat{\mu}_{ab,j} + b_j, \qquad 0 \le s_j \le 2, \quad |b_j| \le 1, \tag{E.1}$$

fitted on out-of-fold pair predictions from gene-disjoint inner folds of the training data. For $s_j > 0$ it leaves each channel's Pearson correlation across held-out conditions unchanged and changes absolute error (Table E.4).

Table D.4: Factorized and direct models with the same encoder on GSE116198. Both arms use the same 48-unit GO encoder, response targets, optimizer, target-disjoint selection and paired initialization, and they differ only in model class and output head. This is a separate experiment from the factorial of Table D.2, fitted in its own runs, so its pair-only scores differ from the factorial's (0.127 and 0.129 against 0.120 and 0.148).

Method or contrast	Pearson	RMSE	MAE	Pearson difference	95% interval
Factorized twin	0.126703	0.156815	0.125659	–	–
Direct pair twin	0.129414	0.143860	0.114225	–	–
Factorized − direct	−0.002712	–	–	−0.002712	[−0.033208, +0.027033]

Table D.5: Factorization and reuse on GSE290153 under the strict both-unseen rule. Pearson is averaged over channels within each panel and then equally over five IR panels. Intervals are 95% bootstrap intervals over held-out genes, and the last column counts the panels with a positive contrast.

Method or contrast	Pearson	Difference / interaction	95% interval	Positive panels
Factorized, pair-only	0.049805	–	–	–
Direct, pair-only	0.031970	–	–	–
Factorized, single + pair	0.029120	–	–	–
Direct, single + pair	0.024644	–	–	–
Pair-only model contrast	–	+0.017835	[−0.002227, +0.036583]	4/5
Single + pair model contrast	–	+0.004476	[−0.020011, +0.029487]	2/5
Reuse interaction	–	−0.013359	[−0.041619, +0.017703]	2/5

With pairs only, the factorized model has a higher point estimate than the matched direct model, but the interval crosses zero. Adding single supervision yields a negative point estimate of the reuse interaction, with an interval that also includes positive values. We report this external result to delimit the positive mechanism result on GSE116198, and it does not show that factorization is better in general.

E.2 GSE284018, WESSELS, GSE200201 AND MCF10A

Tables E.2–E.7 give the matched results behind Figure 3(d) and Table E.1 for these screens.

On the primary Wessels endpoint, the matched contrasts are $f_{\mathrm{comp}} - f_{\mathrm{single}} = +0.0133$ (95% CI $[+0.0062, +0.0210]$), $f_{\mathrm{comp}} - \mathrm{DirectJoint\ Ridge} = +0.0167$ (95% CI $[+0.0036, +0.0290]$) and $f_{\mathrm{comp}} - \mathrm{DirectJoint\ MLP} = -0.0200$ (95% CI $[-0.0337, -0.0073]$). Pair supervision thus improves the shared map, and DACL-C exceeds the linear direct control, but the nonlinear direct model remains better on this screen. This external comparison is limited in scope and does not establish an advantage of factorization.

Table D.6: Pair-residual diagnostic. A constituent map is fitted on training singles. A residual head fits training-pair residuals from the symmetric GO pair descriptor. Held-out responses are read only for the final score. GSE116198 reports Pearson and GSE200201 reports Pearson-Δ. The residual head is a diagnostic, not a proposed model. Negative ΔRMSE and ΔMAE favor the residual head.

Screen	DACL-S	DACL + residual	DirectJoint Ridge	Residual − single (correlation)	ΔRMSE	ΔMAE
GSE116198	0.168660	0.064932	0.033238	−0.103727	−0.011891	−0.022610
GSE200201	0.065381	0.159775	0.161802	+0.094395	+0.002085	+0.000203

The diagnostic separates the constituent component, which carries the correlation, from a pair-specific residual. On GSE116198 the residual lowers Pearson while improving RMSE and MAE. On GSE200201 it brings Pearson-Δ nearly to the level of DirectJoint Ridge while slightly worsening absolute error. The direction of the effect therefore depends on the screen. The training-only diagnostic does not predict it, because residuals are not predictable from the descriptors in either screen (Appendix D).

Table E.1: DACL and direct pair models across screens under the strict both-unseen rule. Values are the primary association endpoint of each screen (Pearson-Δ for expression readouts, Pearson for scalar readouts), averaged over that screen's splits or panels. Direct models receive the same Gene Ontology descriptors. Other direct is the best remaining direct model. It is the DirectJoint MLP on Norman, GSE200201 and Wessels, the richer direct head of the neural factorial on GSE116198 (whose factorized head reaches 0.154) and the sum-only direct model on GSE284018. On PRJNA691742 and Shen2017 it is the richer DirectJoint model of the prospective tests, where DACL has 20 to 30 times its normalized pair error (Appendix F). The last column names the model with the higher point estimate. The label on par marks a point-estimate difference below 0.01, not an equivalence test. The rows mix endpoints and model classes, so they are not independent replications of one effect. Scalar (2) marks two readout channels.

Screen	Readout	DACL-S	DACL-C	DirectJoint Ridge	Other direct	Higher
Norman SEEN-0	Expression	0.375	**0.401**	0.380	0.386	DACL
Norman SEEN-2	Expression	0.773	0.770	**0.776**	0.707	on par
GSE116198	Scalar (2)	**0.169**	0.150	0.033	0.148	DACL
GSE284018 Gamma	Scalar	**0.376**	0.357	0.322	0.332	DACL
GSE284018 Tau	Scalar	**0.314**	0.272	0.214	0.235	DACL
GSE200201	Expression	0.065	0.088	**0.162**	0.155	Direct
Wessels	Expression	−0.178	−0.165	−0.182	**−0.145**	Direct
PRJNA691742	Scalar	–	0.006	–	**0.017**	Direct
Shen2017	Scalar	–	−0.132	–	**0.307**	Direct

Table E.2: Model development on GSE284018 Gamma. The five strict both-unseen gene splits were fixed before any phenotype was read and were evaluated with the same GO descriptor. Values are means over splits, and the primary contrast is DACL-C minus the full DirectJoint Ridge.

Method	Pearson	RMSE	MAE	Role
DACL-S	0.375939	0.089399	0.072533	constituent model
DACL-C	0.356992	0.089782	0.072690	constituent model refined with pairs
Direct sum-only Ridge	0.331945	0.090050	0.071229	additive direct control
DirectJoint Ridge	0.321855	0.091575	0.072267	full direct control
DirectJoint MLP	0.032711	0.257927	0.189557	nonlinear direct control
Training-pair mean	≈ 0	0.091895	0.075642	response-free reference
DACL-C − DirectJoint Ridge	+0.035137	–	–	95% CI $[-0.007960, +0.078635]$

DACL-S has the highest Pearson, whereas the sum-only direct model has the lowest MAE. DACL-C differs from DACL-S by -0.019, so this screen supports the constituent model as a predictor but not a general benefit of pair supervision or an advantage on every metric.

The matched controls vary only the predictor's form while retaining the descriptor and training responses. Relative to DACL-S, DACL-C improves Pearson-Δ by $+0.0226$ and changes RMSE by less than 0.001. DirectJoint Ridge and MLP remain higher in Pearson-Δ by 0.0738 and 0.0666, respectively, and they stay ahead after removing any single gene (Table E.7). On this small pair graph, pair supervision therefore improves the shared map, whereas the direct pair models are the better predictors.

For MCF10A, the stratified pair-bootstrap 95% CI for the pair-supervision gain is $[-0.0054, 0.0015]$ over 29 test-pair instances. For GSE90063, the pair-supervision gain is $+0.0013 \pm 0.0083$ over ten splits, five of them positive. These panels delimit the external scope, and they do not show that the pair-supervision gain replicates reliably on independent genetic screens.

F PROSPECTIVE CAPACITY TESTS

PRJNA691742 measures combinatorial CRISPR genetic interactions in MCF10A cells grown in full, minimal and TGF-β1 media (Zhao et al., 2021). Shen2017 contains combinatorial CRISPR growth screens in HeLa, A549 and 293T cells (Shen et al., 2017). The readout in both screens is

Table E.3: Confirmation on the held-out GSE284018 Tau endpoint. Values are means over five fixed held-out-gene panels. The predictor was selected on Gamma, and all Tau predictions were exported before the held-out Tau responses were made available for scoring. DACL-S with affine calibration is the prespecified primary model, and offset-only calibration is a matched comparator.

Method	Pearson (channel mean)	RMSE	MAE
DACL-S	0.313992	0.148036	0.119045
DACL-S + offset	0.313992	**0.134478**	**0.105148**
DACL-S + affine calibration	**0.313992**	0.134681	0.105255
DACL-C	0.272162	0.149469	0.119547
DirectJoint Ridge	0.214086	0.151698	0.119236
Direct sum-only Ridge	0.234787	0.147703	0.116106
DirectJoint MLP	0.028161	0.278458	0.206464
Training-pair mean	≈ 0	0.139697	0.109590

The calibrated primary model exceeds the full DirectJoint Ridge by $+0.100$ Pearson (95% CI $[+0.042, +0.151]$ from a bootstrap over shared held-out genes) and the sum-only Ridge by $+0.079$ ($[+0.023, +0.128]$). Both contrasts favor it in $5/5$ panels. Compared with uncalibrated DACL-S, affine calibration improves RMSE in $4/5$ panels and MAE in $5/5$. Offset-only calibration is slightly better on pooled RMSE and MAE, so the data do not show a benefit from learning a slope. Tau comes from the same screen as Gamma, so this confirms the predictor on a new endpoint and is not a replication on an independent dataset.

Table E.4: Train-only amplitude calibration on GSE116198 after its held-out results had been examined. The affine map is fitted per response channel on out-of-fold pair predictions from gene-disjoint inner folds. Test responses are read only after predictions are exported. Pearson, reported as a check, is computed within each response channel across held-out conditions and then averaged over channels. A positive slope leaves each channel's Pearson correlation unchanged whenever the predictions have nonzero variance.

Method	Pearson	RMSE	MAE
DACL-C	0.149730	0.184520	0.158151
DACL-C, offset only	0.149730	0.130966	0.105216
DACL-C, affine	0.149730	0.128796	0.104170
DirectJoint raw MLP	0.027798	0.155459	0.126224
DirectJoint Ridge	0.033239	0.164875	0.133361

Relative to uncalibrated DACL-C, affine calibration reduces RMSE by 0.056 (30.20%) and MAE by 0.054 (34.13%), and all ten panels improve. Bootstrap intervals over held-out genes are $[0.048, 0.063]$ and $[0.047, 0.062]$, respectively. Affine and offset-only calibration differ by only 0.002 in RMSE and 0.001 in MAE, so calibration corrects amplitude and offset and does not add a new mechanism for the response pattern.

a scalar pair phenotype. Both screens were analyzed after the models and the selection rule were fixed. Outer splits hold out whole genes. Each of the three PRJNA691742 media has 7 gene-disjoint blocks (21 panels, 504 test pairs), and each of the three Shen2017 cell lines has 8 blocks (24 panels, 891 test pairs). In every panel, three candidates sharing the same Gene Ontology descriptor were fitted on training responses only. These were DACL-C, a sum-only additive model and a richer

Table E.5: Matched controls with the same information on Wessels. Values are means over conditions $\pm$ sample standard deviation over the three fixed strict both-unseen splits. Every row shares the GO descriptor, outer splits, training responses and evaluator. Higher Pearson-Δ and lower RMSE are better.

Method	Pearson-Δ	RMSE	Control-reference Δ	Top-20 Δ	Centroid accuracy
DACL-S	-0.1782 ± 0.1579	0.0599 ± 0.0051	0.5504 ± 0.0363	0.1381 ± 0.2169	0.4880 ± 0.0281
DACL-C	-0.1649 ± 0.1664	0.0586 ± 0.0058	0.5589 ± 0.0406	0.1396 ± 0.2380	0.4925 ± 0.0388
DirectJoint Ridge	-0.1816 ± 0.1517	0.0525 ± 0.0044	0.5630 ± 0.0408	0.1240 ± 0.1940	0.4998 ± 0.0248
DirectJoint MLP	-0.1449 ± 0.1811	0.0516 ± 0.0045	0.5669 ± 0.0410	0.1400 ± 0.2434	0.4894 ± 0.0263

Table E.6: Matched baselines with the same information on GSE200201. Values are mean ± sample standard deviation over the 27 unique primary pairs. Higher Pearson-Δ and lower RMSE are better. All methods use the same GO descriptor, training responses, graph splits and evaluator.

Method	Pearson-Δ	RMSE	MSE×10^3	Control-reference Δ	Top-20 Δ
Matching perturbed mean	0.0000 ± 0.0000	0.0531 ± 0.0094	2.904 ± 1.137	0.5172	0.0000
Response-aligned mean	0.0276 ± 0.0533	0.0551 ± 0.0085	3.104 ± 1.051	0.4606	0.1758
DACL-S	0.0654 ± 0.0929	0.0547 ± 0.0083	3.059 ± 1.019	0.4671	0.1903
DACL-C	0.0880 ± 0.1433	0.0548 ± 0.0074	3.057 ± 0.902	0.4660	0.2214
DirectJoint Ridge	0.1618 ± 0.2833	0.0526 ± 0.0077	2.827 ± 0.892	0.5262	0.3773
DirectJoint MLP	0.1546 ± 0.2879	0.0524 ± 0.0075	2.796 ± 0.869	0.5261	0.3769

Table E.7: Leave-one-gene-out sensitivity on GSE200201. Ranges are over all 20 gene removals, and positive Pearson and negative RMSE differences favor DACL-C.

Comparison	Full Pearson	LOO Pearson range	Full RMSE
$f_{\text{comp}} - f_{\text{single}}$	$+0.0226$	$[+0.0101, +0.0361]$	$+0.000096$
$f_{\text{comp}} -$ DirectJoint Ridge	-0.0738	$[-0.0997, -0.0296]$	$+0.002183$
$f_{\text{comp}} -$ DirectJoint MLP	-0.0666	$[-0.0938, -0.0307]$	$+0.002445$

DirectJoint model. The prespecified rule selected the candidate with the lower normalized pair MSE on target-disjoint inner folds. The normalization divides the MSE by the variance of the training pair phenotypes. All predictions were exported before the outer responses were read. The only query gene without a descriptor, MLLT4 in PRJNA691742, used the zero vector. Table F.1 reports the outer results with 95% intervals from a paired bootstrap over blocks.

The rule selected DirectJoint in all 45 panels, and the sign of the inner-fold gap between DirectJoint and DACL matched the sign of the outer gap in 21 of 21 and 24 of 24 panels. The rule predicted the direction of the gap but not its size. With the gap defined as the normalized MSE of DirectJoint minus that of DACL-C, the mean inner and outer gaps are -32.9 and -36.2 on PRJNA691742 and -37.4 and -32.7 on Shen2017. The outer minus inner difference averages -3.27 ($[-27.49, 19.06]$) and 4.77 ($[-4.73, 13.67]$). On these screens, DACL has about 20 to 30 times the normalized pair MSE of the direct model. On Shen2017 the direct model is also better in correlation, whereas on PRJNA691742 neither model has a correlation distinguishable from zero. Because the rule chose the direct model everywhere, it coincides with always choosing DirectJoint and cannot improve on that policy. In these two screens, training data alone identified the better model class, but switching between classes remains unvalidated. The 45 panels share genes within each screen and are not independent experiments.

Table F.1: Prospective capacity tests. Outer normalized pair MSE (lower is better) and Pearson correlation across held-out pairs (higher is better), averaged over panels, with 95% paired block-bootstrap intervals. The last column counts panels in which the Pearson of DirectJoint exceeds that of DACL-C. The selection rule chose DirectJoint in every panel.

Screen	Model	Normalized pair MSE	Pearson	Pearson above DACL
PRJNA691742	DACL-C	$37.998\,[18.183, 62.072]$	$0.006\,[-0.262, 0.268]$	–
PRJNA691742	Sum-only additive	$42.647\,[26.932, 60.359]$	$-0.079\,[-0.156, -0.003]$	–
PRJNA691742	DirectJoint	$\mathbf{1.843}\,[0.947, 3.055]$	$0.017\,[-0.255, 0.312]$	13/21
Shen2017	DACL-C	$33.887\,[25.904, 42.517]$	$-0.132\,[-0.281, 0.015]$	–
Shen2017	Sum-only additive	$123.150\,[94.577, 152.227]$	$-0.029\,[-0.153, 0.080]$	–
Shen2017	DirectJoint	$\mathbf{1.209}\,[0.703, 1.789]$	$\mathbf{0.307}\,[0.164, 0.463]$	19/24

Table G.1: Execution of each comparator. Conceptually related methods are kept separate from controls run under our protocol. A published method receives a benchmark score only if the information it uses and its evaluation units match the strict both-unseen protocol. Our runs use the fixed splits, response basis, target-disjoint selection and evaluator of this paper.

Method	Its own protocol compared with ours	Execution in this study	Role in the argument
biolord	Predicts single-gene effects from Gene Ontology attributes and adds them for pairs, using measured singles where available, without pair supervision of the constituent map	Our runs of the released Norman recipe on the official and 70 further splits, fitted on training singles only	Shows that addition by itself does not produce the gain on pairs of unseen genes
PerturbNet	Reports a setting in which both genes of a pair are unseen, with GO-based genotype encoding and distributional outputs	Not run, and no score from its own protocol copied	Places the choice of a response-free descriptor in context
Scouter	Its paper uses GenePT embeddings of gene descriptions, whereas its released Norman script loads scELMo GPT-3.5 embeddings (Liu et al., 2026), which we used; trained on single and pair perturbations	Our runs of the released code and Norman script, which restores last-epoch weights, with test cells removed before fitting	Tests whether a richer response-free gene representation closes the gap without constituent reuse
scGPT	Fine-tunes a pretrained transformer on the training perturbations and flags the perturbed genes in the input	Our runs of the released perturbation tutorial on the official splits, epochs selected by validation error, test responses unused	Tests whether pretrained gene representations retrieve responses for unseen genes
Response decomposition	Maps summed DepMap profiles of the two genes to the pair response, trained on pairs, and its published results do not use the present strict split	Our runs of the released Ridge and MLP pair models on the official Norman splits	Direct-pair reference with a different descriptor
DirectJoint Ridge/MLP	Same response-free GO information, target-disjoint selection and evaluator as DACL in the mechanism analyses	Run on Norman, GSE116198, GSE200201, GSE284018 and Wessels	Tests whether a weak or under-parameterized direct baseline explains the reuse effect
DACL-S/DACL-C	Shared response-free constituent map, supervised by singles with optional pair refinement, and explicit addition	DACL-S and DACL-C on Norman and the external screens, DACL-C in the prospective tests, and a neural factorized variant in the GSE116198 and GSE290153 factorials	Tests reuse through a descriptor-indexed map and when direct pair capacity is preferable

G Comparators and the questions they test

Table G.1 lists, for each comparator, how its own protocol differs from ours, how it was run in this study and which question of the argument it addresses.

H Theory and metric geometry

Descriptor floor and pair-risk identity. For $f^\star(q) = \mathbb{E}[\beta \mid q]$, write $\beta - f(q) = (\beta - f^\star(q)) + (f^\star(q) - f(q))$. The first term is orthogonal in L_2 to every q-measurable function, so the cross term vanishes and Eq. (1) follows. For the pair identity, write $Y_{ab} = \beta_a + \beta_b + \eta_{ab}$ and $e_g = f(q_g) - \beta_g$, and direct expansion gives Eq. (7). The cross terms between e_a, e_b and η_{ab} are not determined by the marginal single-gene risk, so equal single risk need not imply equal pair risk.

Shared-map update. The single normal equations are $(X^\top X + \rho I)W_0 = X^\top B$. Adding the weighted pair block gives $(X^\top X + \omega_\lambda Z^\top Z + \rho I)W_\lambda = X^\top B + \omega_\lambda Z^\top P$. Subtracting the single normal equations and collecting terms in $W_\lambda - W_0$ yields the exact update in Eq. (6). Submultiplicativity then gives the operator-norm bound

$$\|W_\lambda - W_0\|_F \le \omega_\lambda \|(X^\top X + \omega_\lambda Z^\top Z + \rho I)^{-1} Z^\top\|_2 \|P - ZW_0\|_F.$$

No pair identifier or new feature column is introduced.

Constituent versus direct pair models. Let $m(Q) = \mathbb{E}[Y \mid Q]$, $R(h) = \mathbb{E}\|Y - h(Q)\|^2$, and $R_{\text{info}} = \mathbb{E}\|Y - m(Q)\|^2$. For a class $\mathcal{H}_M$, choose $h^*_M \in \arg\min_{h \in \mathcal{H}_M} R(h)$ when the minimum is attained, and define $A_M = R(h^*_M) - R_{\text{info}}$ and $X_M = R(\widehat{h}_M) - R(h^*_M)$. Then $R(\widehat{h}_M) = R_{\text{info}} + A_M + X_M$ by definition, so subtracting the constituent and direct identities gives Eq. (9) without an orthogonality assumption. For a closed linear or affine subspace, the class minimizer is an L_2 projection, and X_M can additionally be written as $\mathbb{E}\|\widehat{h}_M - h^*_M\|^2$. A general nonlinear MLP class does not justify this replacement, because its cross term need not vanish. For a loss bounded by B and two fixed candidates evaluated on n_V independent and identically distributed validation samples, Hoeffding's inequality and a union bound give $|\widehat{\mathcal{R}}_V(M) - \mathcal{R}_M| \le B\sqrt{\log(4/\delta)/(2n_V)}$ simultaneously. The selected candidate is therefore within twice this margin of the better candidate, which gives the selection margin

$$\Delta_{\text{select}} = 2B\sqrt{\frac{\log(4/\delta)}{2n_V}}. \tag{H.1}$$

For clustered screens, the bound requires a count of independent clusters or a concentration or bootstrap argument that is valid under clustering, because raw rows do not satisfy the iid assumption.

Correlation and amplitude. Expanding half the squared norm of the difference between the two centered unit vectors gives $1 - y_c^\top \hat{y}_c/(\|y_c\|\|\hat{y}_c\|)$. Equivalently,

$$\text{Pearson}(y, \widehat{y}) = 1 - \frac{1}{2}\left\| \frac{y_c}{\|y_c\|_2} - \frac{\widehat{y}_c}{\|\widehat{y}_c\|_2} \right\|_2^2. \tag{H.2}$$

A residual head can therefore improve the scale of a prediction, and with it RMSE or MAE, while rotating the normalized response direction that Pearson measures. For the same response vector, the squared loss used in the risk identities is the MSE in response space, $\text{MSE}(y, \hat{y}) = d^{-1}\|y - \hat{y}\|_2^2$, with $\text{RMSE} = \sqrt{\text{MSE}}$. The matched controls report these quantities alongside Pearson-Δ, which links the theory to absolute error without implying a guarantee on Pearson.