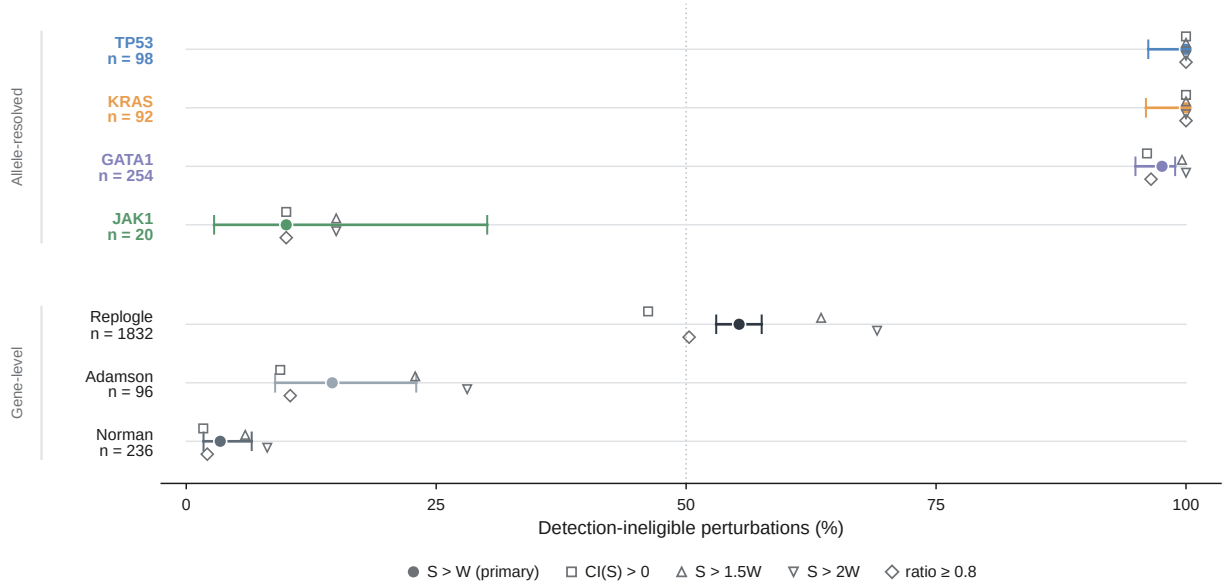
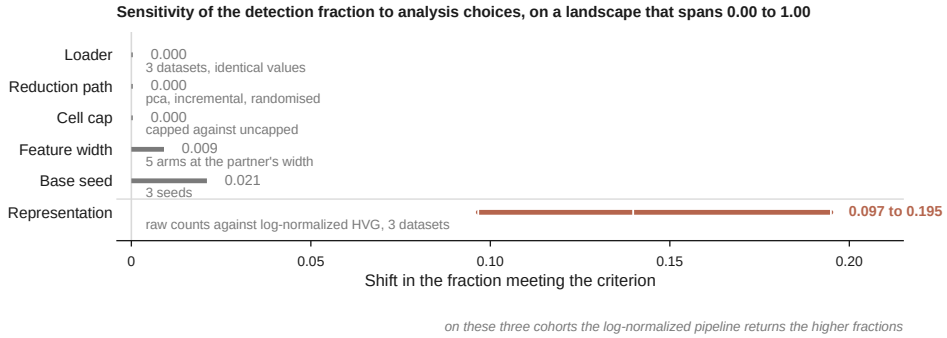
Supplementary Information

Measurement resolution constrains fine-grained perturbation prediction

This Supplementary Information provides additional methodological detail and supporting analyses for the variant representations, prediction models and measurement-resolution framework used in the main text.



Supplementary Figure 1: Detection regimes are stable across alternative decision criteria. Fraction of detection-ineligible perturbations in the four allele-resolved datasets and three gene-level Perturb-seq atlases under five prespecified criteria. The primary criterion compares perturbation-to-reference separation S with within-condition sampling variation W ($S > W$; filled circles). The alternatives (open symbols) apply a bootstrap lower bound on S, increased stringency ($S > 1.5W$ and $S > 2W$), or the distance ratio $D_{\text{self}}/D_{\text{null}} \geq 0.8$. Absolute fractions rise with stringency, by up to 23 percentage points for Replogle, but the measurement regimes are unchanged under every criterion: TP53 and KRAS remain at the detection floor, GATA1 remains weakly resolved and JAK1 remains substantially separated from its reference. Bars, Wilson 95% confidence intervals on the primary criterion; symbols are offset vertically within a row for clarity; n, perturbations per dataset; dotted line, 50%.



Supplementary Figure 2: Detection fractions are stable to routine pipeline choices and depend on the expression representation. Shift in the fraction of perturbations meeting the detection criterion under six analysis choices, each evaluated on cohorts held fixed. The data loader, the dimension-reduction path (exact, incremental and randomized), the per-perturbation cell cap, the feature width and the base random seed move the fraction by at most 0.021 on a landscape that spans 0.00 to 1.00. Changing the expression representation, from the unnormalized count space used for the broader perturbation panel to the log-normalized, highly-variable-gene space used for the split-half depth ladder, moves it by 0.097 to 0.195 on the three resources evaluated under both pipelines, and in the same direction in all three: the log-normalized pipeline returns the higher fractions. Absolute detection fractions are therefore specific to the representation and criterion under which they were measured, and are compared only within a pipeline. Tick marks on the representation bar are the three individual resources.

Supplementary Table 1: Structural and functional annotations used in the biophysical variant representation. Per-gene residue annotations for the position-dependent components of the feature vector θ. The fold-core term f_v was set to 1 within the listed structural cores and 0 otherwise. The functional-switch term s_v marks catalytic or coordinating residues. The hotspot/pathogenic term h_v was defined from external annotations without reference to the measured single-cell response.

Gene	Structural core ($f_v = 1$)	Functional switch ($s_v = 1$)	External functional or hotspot annotation ($h_v = 1$)
TP53	DNA-binding domain, residues 94–293	–	COSMIC/IARC missense hotspot codons 175, 176, 179, 220, 238, 245, 248, 249, 273, 282 and 285
KRAS	G-domain, residues 1–169	G12, G13 and Q61	COSMIC/IARC missense hotspot codons 12, 13, 59, 61, 117 and 146
GATA1	Zinc fingers ZF1 (204–228) and ZF2 (258–282)	Zinc-coordinating cysteines	Zinc-coordinating cysteines 204, 207, 225, 228, 258, 261, 279 and 282; ClinVar pathogenic or likely-pathogenic annotations
JAK1	Kinase domain, residues 866–1154	–	JH2 pseudokinase (583–855) and JH1 kinase (875–1153) functional modules

Supplementary Table 2: Prediction-model and measurement-resolution parameters. The six regression heads were crossed with the three primary molecular feature spaces (θ, ESM-1v and ESM-1v+θ) to form the 18 feature-based predictors evaluated in the main model grid. The lower panel lists the principal sampling parameters used for the measurement-resolution and repeated-pilot analyses.

Regression head	Family	Settings
Ridge regression	Linear (L2)	$\alpha = 1.0$
Lasso regression	Linear (L1)	$\alpha = 0.01$; maximum 2,000 iterations
Random forest	Tree ensemble	50 trees; maximum depth 6
Gradient-boosted trees	Tree ensemble	50 estimators; maximum depth 3; multi-output wrapper
k-nearest-neighbours	Neighbour	$k = 3$
Multilayer perceptron	Neural network	hidden layers 128 and 64; ReLU activation; maximum 500 iterations

Analysis	Representation, depths and partitions
Split-half depth ladder (Fig. 4c,d, Fig. 6f)	dataset-native analysis spaces for the allele-resolved arm: processed 939-gene count spaces for TP53 and KRAS, the log-normalized 2,477-HVG space for GATA1, and the log-normalized 2,000-gene space for JAK1. The Replogle, Norman and Adamson gene-level screens use library-size normalization to 10^4 counts per cell, log transformation and 2,000 highly variable genes. A 50-component PCA is fitted once per gene or dataset; 50 random cell splits are used at depths of 30, 50, 100 and 150 cells per half for the allele-resolved datasets and 25, 50, 100, 200, 400 and 800 for the gene-level datasets. A perturbation enters a rung at twice that rung's cells; allele-resolved pools are capped at 300 cells
Nearest-competitor and controlled-geometry sweep (Fig. 4e, Fig. 6d)	$m \in \{25, 50, 100\}$ cells per group with 10 random cell partitions; the measured configurations take four disjoint groups per variant and the constructed configurations five, the extra block supplying the direction that is dialled; constructed candidate-set sizes, including the target, of 20, 45 and 90 variants
Broader-panel criterion (Fig. 5)	unnormalized counts with no log transformation and no gene selection, 50-component PCA fitted once per resource-by-stratum dataset; 50 cells per group; 8 random cell partitions; 1,000 bootstrap resamples; seed 0; a perturbation enters at 200 cells
Repeated-pilot detection association (Fig. 6b)	50-component PCA, or truncated SVD for feature matrices too wide for exact decomposition, fitted before partitioning; 10 random pilot and evaluation partitions; 25- or 50-cell pilot scores averaged across partitions; detection eligibility defined from the median and percentile summaries of ten 100-cell-per-half evaluation draws; perturbed pilot and evaluation cells disjoint within a partition but reusable across partitions; reference draws not mutually exclusive; one AUROC after aggregation; 2,000 bootstrap resamples over perturbations
Split-half reproducibility reference	15 random split-half subsamples per variant, each half capped at 300 cells, scored through the PDS harness used for the predictors
Sensitivity representations	gene-expression space; pathway-score space where available
Minimum cells for pseudobulk prediction	5 per variant condition

Note. Hyperparameters were fixed at the values above across genes, generalization settings and splits, and were not tuned to individual datasets; no hyperparameter search was performed. Linear, nearest-neighbour and neural-network models predicted expression profiles directly. For expression spaces exceeding 200 genes, random-forest and gradient-boosted models were fitted to a

PCA representation of the training responses and transformed back to gene space after prediction, with $q = \min(50, G, n_{\text{train}} - 1)$.

Supplementary Table 3: Released perturbation models and allele-conditioned adaptations under the unified evaluation. Except for the explicitly separated scVIDR sensitivity control, PDS and Pearson-δ are evaluated on held-out protein-coding variants without using the held-out response cells, using the same pseudobulk targets and scoring framework as the feature-based predictors. Repeated setting-level scores are first averaged within each gene–variant pair, and both metrics are then aggregated over distinct variants. PDS confidence intervals resample those distinct variants. The reported n is the number of PDS queries; finite prediction coverage and the Pearson-δ denominator differ only for scVIDR and are stated in its footnote.

Model family	Method	n	PDS (95% CI)	Pearson-δ
Response-blind allele-conditionable models				
Optimal-transport flow	CellFlow	303	0.49 (0.47–0.52)	0.16
Attribute disentanglement	Biolord	308	0.48 (0.45–0.51)	0.14
Latent arithmetic, allele-conditioned adaptation	scGen	308	0.47 (0.45–0.50)	0.20
Conditional-flow generative model	PerturbNet[†]	303	0.530 (0.504–0.555)	0.13
Feature regression	models in Fig. 2	305–308	0.47–0.49	0.58–0.67
Response-informed sensitivity control; excluded from response-blind inference				
Latent arithmetic, oracle-scaled adaptation	scVIDR[§]	308	0.486 (0.461–0.511)	0.18
Reference predictors				
Gene-level response	Gene-mean	308	0.48 (0.45–0.50)	0.67
No perturbation	WT-null	308	0.50	–
Models without an allele-resolved conditioning interface				
Foundation model	scGPT	–	n/a[‡]	–
Gene knowledge graph	GEARS	–	n/a[‡]	–
Virtual-cell model	STATE	–	n/a[‡]	–

[†]PerturbNet was run across all five scored within-gene settings, yielding 492 gene–setting–variant evaluations and 303 distinct gene–variant units after overlap across settings was removed. Predictions and candidates were compared in its 50-dimensional output representation. Uniform matching over the complete candidate pool still has exact expected PDS 0.5. A task-matched test that permuted only held-out labels within each gene and setting while retaining the full training-plus-held-out candidate columns had a restricted-label null mean of 0.515500; 1,000,000 permutations gave a one-sided plus-one $P = 0.069768$ for the reported PDS of 0.530143.

[‡]scGPT, GEARS and STATE condition on perturbed-gene identity and therefore do not define distinct predictions for variants of the same gene; allele-level PDS was not evaluated.

[§]The displayed scVIDR result is not a response-blind held-out-variant evaluation. Its scaling variable is an external AlphaMissense score for TP53 and KRAS, but for GATA1 it is the measured variant-to-WT expression distance normalized by the maximum over the full GATA1 cohort. It therefore uses the held-out GATA1 response and full-cohort response information at prediction time. The variable is missing for JAK1, so the adaptation returned no finite prediction for 26 JAK1 queries. To retain the common PDS query cohort, those queries receive the stated no-ranking value of 0.5, giving 308 PDS queries with finite prediction coverage of 282; the Pearson-δ value excludes the non-finite JAK1 vectors and is aggregated over those 282 finite predictions. This row is reported only as a response-informed sensitivity control.

Supplementary Table 4: Measurement resolution and prediction performance across the four allele-resolved datasets. The split-half ratio $D_{\text{self}}/D_{\text{null}}$ compares within-variant sampling variation with variant-to-reference separation. The reproducibility reference measures full-pool allele identification between disjoint cell samples using the same PDS framework as the predictors, aggregated on the distinct protein-coding variant, which is the primary statistical unit defined in Methods; n is the number of distinct variants scored. These are the same values used in Fig. 3. Sibling pairs resolvable gives the fraction of same-gene variant pairs separated beyond the mean of their two split-half distances, and nearest-sibling identifiable the stricter fraction of variants whose closest sibling remains separated beyond that variant's own split-half distance. Detection eligibility gives the fraction of variants distinguishable from the corresponding reference at native depth.

Gene	$D_{\text{self}}/D_{\text{null}}$	Reproducibility reference (95% CI), n	Sibling pairs resolvable	Nearest-sibling identifiable	Detection-eligible	Best feature-based PDS
TP53	0.96	0.472 (0.441–0.505), 59	0%	0%	0%	0.479
KRAS	1.00	0.481 (0.431–0.533), 51	0%	0%	0%	0.505
GATA1	0.88	0.578 (0.552–0.604), 172	3.8%	0%	2.4%	0.509
JAK1	0.21	0.807 (0.729–0.880), 26	78.5%	15.4%	90%	0.448

Note. Native depth denotes the deepest evaluated split-half rung supported by each variant, capped at 150 cells per half for these datasets; it does not denote use of all available cells. The final column reports the highest PDS among the feature-regression models for each gene on the primary scoring path. Released perturbation models are reported separately in Supplementary Table 3.

Supplementary Table 5: Three distinct measurement resolutions across six prespecified perturbation screens. The datasets, preprocessing parameters and operating points were specified before evaluation as described in Methods. The three quantities are different biological comparisons rather than successive levels of one scale: detection is separation from the reference condition, nearest-competitor identification is separation from the closest competing perturbation, and the reproducibility reference with $P(\text{order})$ describe benchmark-level ranking resolution, the first as split-half perturbation ranking and the second as the bootstrap probability of recovering the predefined ordering of the controlled predictor family. A screen can support one and not another, and the table is read across the row rather than as a pass or fail sequence.

Dataset	n	Detection (%)	Nearest-competitor identification (%)	Reproducibility reference	$P(\text{order})$
McFarland	10	90.0	20.0	0.928	0.418
Tahoe-100M	7	71.4	42.9	0.929	0.106
Norman 2019	195	68.7	0.5	0.898	0.997
Replogle 2022	414	51.2	1.4	0.865	0.959
Adamson 2016‡	93	48.4	3.2	0.853	0.977
Virtual Cell Challenge	139	25.9	4.3	0.806	0.903

Note. The prespecified operating points were $> 50\%$ of perturbations detected, $> 50\%$ identified against the nearest competitor and a split-half reproducibility reference > 0.65. They were fixed before the screens were evaluated.

‡An unassigned `<NA>` level in the Adamson perturbation field was excluded as an observation rather than a perturbation, leaving 93 scored perturbations.

Supplementary Table 6: Public datasets assessed for inclusion in the primary allele-resolved benchmark. Datasets were assessed against three operational requirements: **R1**, an accessible genotype-linked single-cell expression representation; **R2**, sufficient sibling coding variants of the same gene, each with at least five assigned cells, for within-gene held-out evaluation; and **R3**, compatibility with a common held-out-variant prediction protocol. The table lists the public datasets assessed for this study.

Study (accession)	Variants assayed	R1	R2	R3	Outcome
Ursu et al. 2022 (GSE161824)	*TP53* 98; *KRAS* 92	yes	yes	yes	included
Martin-Rufino et al. 2023 (GSE215253)	*GATA1* 254	yes	yes	yes	included
Cooper et al. 2024 (PRJEB48915)	*JAK1* 26	yes	yes	yes	included
Kim et al. 2024, TISCC-seq (PRJNA880341)	*TP53*, > 100	no	–	–	no released cell-to-genotype expression mapping
Hollek et al. 2026 (GSE299651)	50 transgenes across five cell lines	yes	no	–	insufficient sibling alleles per gene
Xu et al. 2023, preprint	*CDKN2A*, *TP53*, *SOD1*	no	–	–	no accessible expression dataset identified

Note. Dataset availability was assessed on 11 August 2026. R1 concerns the data products available for unified reanalysis rather than whether genotype-linked expression could in principle be reconstructed from raw sequencing data.

Supplementary Note 1: Robustness of measurement-resolution estimates

Alternative detection criteria. The relative measurement regimes were stable across alternative definitions of detection eligibility (Supplementary Fig. 1). Increasing the stringency of the threshold on perturbation-to-reference separation relative to within-condition sampling variation changed the absolute fraction of eligible perturbations, by up to 23 percentage points for Replogle, but did not alter the separation between the weakly resolved TP53, KRAS and GATA1 datasets and the substantially resolved JAK1 dataset. What is stable is the regime each dataset falls in, not the eligibility value.

Distance metric and expression representation. The detection regime was not specific to energy distance. At 50 cells per half in the canonical PCA space, the ratio $D_{\text{self}}/D_{\text{null}}$ remained near one for TP53, KRAS and GATA1 and substantially lower for JAK1 across energy distance, RBF maximum-mean-discrepancy, sliced-Wasserstein distance, cosine distance and Euclidean distance. The corresponding ranges were 1.01–1.03 for TP53, 0.95–0.99 for KRAS, 0.90–1.04 for GATA1 and 0.05–0.40 for JAK1.

Repeating the native-depth analysis directly in gene-expression space gave the same dataset-level interpretation. Detection eligibility was 0% for TP53 and KRAS, 1.2% for GATA1 and 90% for JAK1, compared with 0%, 0%, 2.4% and 90%, respectively, in the primary PCA representation. Pathway-score representations, available for GATA1 and the three gene-level Perturb-seq datasets, generally produced lower absolute eligibility fractions but did not create a new measurement regime.

Batch-disjoint split-half sensitivity analysis. The split-half reference draws both groups from a pooled shuffle, so both carry the same batch composition. Where per-cell experimental batch was recoverable, we replaced the pooled shuffle with two disjoint batch groups at matched depth through the identical harness. At 50 cells per group the dataset-level reference moved by at most

0.038 PDS (largest in sci-Plex, 0.619 versus 0.581; smallest in Replogle, 0.789 versus 0.784). TP53 and KRAS remained at chance under both definitions, and the dataset-level regimes were preserved throughout. These analyses assess sensitivity to within-experiment batch structure; no dataset analysed here contains independently repeated experiments, so they do not constitute replication. The JAK1 genotyped screen is a single sequencing lane and could not be tested this way.

Reference-population variability. We tested whether weak perturbation-to-reference separation could instead arise from heterogeneity of the reference population. At matched sampling depth, we compared within-variant split-half distance D_{self}, variant-to-reference distance D_{null} and the distance between independent reference samples D_{wtwt}. For TP53, KRAS and GATA1, $D_{\text{wtwt}}/D_{\text{null}}$ ranged from 0.91 to 1.09 across 50–300 cells per half, indicating that variant-to-reference separation was comparable to variation between independent reference samples. In JAK1, by contrast, D_{null} exceeded D_{wtwt} by 7.4- to 31-fold across the tested depths. The weak detection regime in the first three datasets therefore reflects limited perturbation separation relative to sampling variation rather than an unusually heterogeneous reference population.

Supplementary Note 2: Robustness of allele-prediction conclusions

Dependence among variants at the same residue. Several benchmark residues contain more than one amino-acid substitution, so variant-level observations are not completely independent. We therefore repeated the primary uncertainty analysis by resampling residue-level clusters rather than individual variants. Twenty-three of the 24 predictors with confidence intervals continued to overlap the analytic full-pool PDS expectation of 0.5, with PerturbNet remaining the sole exception by that descriptive comparison, and Pearson-δ remained positive for every feature-based regression model. For PerturbNet, the matched restricted-label permutation test for the current 303-variant estimator gives $P = 0.0698$, as described below. The separation between direction recovery and full-pool allele identification is therefore unchanged when substitutions at the same residue are treated as a single resampling cluster.

Dependence among resolved sibling pairs. The distinct-pair bootstrap displayed in Fig. 3e does not make pairs independent when two pairs share a variant. We therefore performed a variant-node bootstrap sensitivity analysis on the fixed cross-fitted resolved-pair cohorts. GATA1 contained 425 edges incident to 131 variants and JAK1 contained 116 edges incident to 21 variants. In each of 1,000,000 draws, the n incident variants were sampled with replacement, an original edge (i, j) received weight $c_i c_j$ from the sampled endpoint multiplicities, and its pair score entered a weighted mean. The same node draw was applied to the measurement, each model and their paired difference. Repeated copies of one original variant did not create an unobserved self-edge. For GATA1, the split-half measurement accuracy was 0.988 (95% node-bootstrap interval, 0.972–0.998) and the highest response-blind point estimate, Ridge-ESM, was 0.780 (0.603–0.901), with a paired model-minus-measurement difference of -0.208 (-0.382 to -0.093). For JAK1, the corresponding values were 1.000 (1.000–1.000) for the measurement and 0.762 (0.594–0.883) for Biolord, with a difference of -0.238 (-0.406 to -0.117). The response-informed scVIDR sensitivity control had the largest displayed GATA1 point estimate, 0.830 (0.634–0.937), and a difference of -0.159 (-0.356 to -0.050). Thus the intervals widen when shared endpoints are respected, but every fitted predictor's paired interval remained below the split-half measurement in these fixed cohorts. These conditional intervals hold the resolved-pair selection and cell partitions fixed and are not simultaneous or winner-selection-adjusted intervals.

Alternative molecular representations. We evaluated the same held-out-variant task using the

biophysical representation θ, ESM-1v, four ESM2-derived representations and AlphaMissense. The ESM2 representations were built from ESM2-650M (`esm2_t33_650M_UR50D`) final-layer per-residue hidden states. For each variant the mutant sequence was formed by applying every listed substitution to the reference sequence, with each substitution validated against the wild-type residue at that position. Sequences exceeding the model's 1,022-residue limit, which applies to JAK1 at 1,154 residues, were cropped to a 1,022-residue window centred on the mutated positions, and the wild-type sequence was cropped identically so that every difference is taken between matched windows. Writing $\mathbf{h}^{\mathrm{mut}}$ and $\mathbf{h}^{\mathrm{wt}}$ for the two hidden-state matrices and p for a mutated position, the four representations are the mutant global embedding, the residue mean of $\mathbf{h}^{\mathrm{mut}}$; the global difference, the residue mean of $\mathbf{h}^{\mathrm{mut}}$ minus that of $\mathbf{h}^{\mathrm{wt}}$; the mutated-residue difference, $\mathbf{h}^{\mathrm{mut}}_p - \mathbf{h}^{\mathrm{wt}}_p$; and the local-window difference, the mean of $\mathbf{h}^{\mathrm{mut}} - \mathbf{h}^{\mathrm{wt}}$ over residues $p \pm 16$. For variants carrying more than one substitution, the last two are averaged over the mutated positions. AlphaMissense entered as its per-variant pathogenicity score, matched on wild-type residue, position and mutant residue; variants carrying more than one substitution have no such entry and were excluded from that representation. Across these representations, full-pool allele identification remained near the null while transcriptional direction remained recoverable. Representation coverage differed because some molecular annotations were unavailable for individual variants; AlphaMissense scores were available for 374 of the 470 benchmark variants, compared with 470 for θ, 467 for ESM-1v and the corresponding available sets for the ESM2 representations. These analyses therefore test whether the principal result depends on a particular molecular encoding rather than providing a direct ranking of representations with unequal variant coverage.

Numerical convergence of Lasso-ESM to the gene-constant reference. For Lasso-ESM, held-out predictions in TP53, KRAS and GATA1 converged numerically to the Gene-mean reference, agreeing to a maximum relative difference of 4×10^{-7}, which is the rounding scale of the single-precision prediction arrays; JAK1 did not show this equivalence (maximum relative difference 7.4). One of the 18 feature-based predictors therefore coincides with the gene-constant reference in three of the four datasets. Lasso-θ and Lasso-ESM+θ did not converge in this way (maximum absolute differences 6×10^{-2} to 9×10^{-2}).

Top-k shortlisting: treatment of predictions with no ranking. A prediction whose distance to its target is not finite has no ordering over the candidate set, so in the primary analysis it is assigned that query's chance expectation, the same convention by which such predictions receive PDS 0.5. Scoring them as misses instead would record an absent ordering as a wrong one. Exactly one implementation is affected: the response-informed scVIDR sensitivity control, in 885 of its 7,770 scored rows, all in JAK1. Under the primary convention its shortlisting rates are 0.0079, 0.0229, 0.0430 and 0.0865 at $k = 1, 3, 5$ and 10; scoring the same rows as misses gives 0.0046, 0.0132, 0.0267 and 0.0541. No response-blind method moves, and because the sensitivity arm can only lower the scVIDR rate, the conclusion that no response-blind method exceeds its query-matched chance expectation is unchanged. Both conventions are reported in the released summary table.

Residual-axis permutation calibration. Uniform matching over the complete residual candidate pool retains the exact analytic PDS expectation of 0.5. The residual-axis inference instead used a task-matched restricted-label permutation: prediction rows were reassigned among held-out labels within a gene and setting, while the full training-plus-held-out residual candidate columns were retained. Because the permuted label set is only a subset of the ranking pool, this conditional reference need not equal 0.5. Permutations were run in two stages for economy: 50 permutations for every method, followed by 1,000 permutations for the methods whose preliminary permutation P fell below 0.05. Residual-axis P values were then Bonferroni-adjusted across the 25 scored methods. The released table reports unadjusted values of 0.0020 for scGen and 0.0080 for Biolord, corresponding to adjusted values of 0.050 and 0.200 at the reported precision. With 1,000 permu-

tations, the smallest attainable P is $1/1001$. The scGen result lies at the corrected threshold at this reporting precision and is treated as unresolved, rather than as evidence below that threshold.

Evaluation setting, scoring geometry and standardization. The separation between direction recovery and full-pool allele identification was examined under each component of the evaluation that could plausibly produce it.

For any fixed prediction row, the tie-aware mid-ranks over all n candidate columns sum to $n(n-1)/2$. Uniform matching over the complete candidate pool therefore has expected PDS exactly 0.5, regardless of finite n, ties or the dimension of the response representation. The empirical check used for the grey band in Fig. 2c instead constructed held-out-only square distance matrices within each gene and evaluation setting and permuted the prediction rows. Because both its labels and candidate columns were restricted to the same held-out set, this implementation check also has expectation 0.5. Across the 1,020 feature-predictor, gene, setting and distance combinations, its permutation means averaged 0.5000 and ranged from 0.494 to 0.505.

For task-matched restricted-label permutations, only the held-out target labels were reassigned while the complete training-plus-held-out candidate columns used by the reported PDS were retained. This conditional reference can differ from 0.5 because the permuted labels form a subset of the ranking pool. For the current PerturbNet estimator, the observed PDS is 0.530143 across five settings and 303 distinct variants. The exact restricted-label null mean is 0.515500; 1,000,000 permutations gave a central 95% null interval of 0.496171–0.534942 and a one-sided plus-one value of $P = 0.069768$. The mapping was held fixed across the 15 cell-subsample seeds, after which scores were averaged over seeds, then across settings for each distinct variant, and finally over variants. The shift of this conditional null above 0.5 is caused by the restricted held-out label set, not by PerturbNet's 50-dimensional output representation.

The same predictors were then scored under each evaluation setting separately, so that a dissociation present in only one generalization regime would be visible as such. Averaged over the 19 non-null predictors, mean PDS and mean Pearson-δ were 0.493 and 0.709 under random holdout, 0.473 and 0.661 under positional extrapolation, 0.494 and 0.503 under mechanistic extrapolation, 0.511 and 0.636 under the compatibility protocol, and 0.240 and 0.547 under low-depth evaluation. Full-pool identification therefore stayed at or below chance in every setting. Low-depth evaluation weakened both axes rather than revealing identification, and drove PDS well below chance, which is expected when the measured targets themselves are the noisiest: a prediction can then sit closer to a better-sampled competitor than to its own target.

Scoring geometry was varied by recomputing the identical predictions and targets under L1 and L2 distances alongside the cosine distance used throughout. The per-head mean PDS moved by at most 0.012 under L1 and 0.015 under L2, the largest shift in each case belonging to the same head, so no predictor crossed from chance to reliable identification under a different geometry.

Finally, the training-only per-gene standardization used in the primary analysis was replaced with all-cell standardization. Relaxing the leakage control raised scores for 18 of 20 predictors on PDS and 19 of 20 on Pearson-δ, by at most 0.033 PDS and 0.030 Pearson-δ, both for the same gradient-boosted head. The convention used throughout is therefore the conservative one, and the reported dissociation is not produced by it.

Training-response subspace control in JAK1. The JAK1 dataset provides the clearest setting in which the measurement itself contains reproducible allele-level information but the evaluated predictors remain close to chance. To ask whether that gap could arise because no prediction expressible in the space of measured training responses could do better, each held-out JAK1 response was projected onto the span of the measured training-variant responses. This projection reached PDS 0.690 (95% CI, 0.638–0.742), against a best feature-based model PDS of 0.448 and a split-half reproducibility reference of 0.792 measured on the same four cell splits as the projection. That

reference is scored per cell split rather than on the distinct-variant unit used in Supplementary Table 4, which is why it reads 0.792 here and 0.807 there; the two are the same quantity under two aggregations, and the control is reported against the one it was matched to. The span of measured training-variant responses therefore contains substantially more allele-discriminative structure than the fitted feature-based models recover. The projection is an upper reference for that subspace and not a score any particular architecture is guaranteed to reach.

Exact identification remained substantially more difficult than ranking-level discrimination: the projection achieved top-1 accuracy 0.049, whereas the split-half measurement reference reached 0.247. The JAK1 model–measurement gap is therefore interpreted at the level of within-gene ranking rather than exact nearest-allele identification.

Supplementary Note 3: Implementation of the released perturbation models

All scored released-framework implementations were run on the same arrays, cell caps and variant partitions as the feature-based predictors: per-gene standardized expression matrices, at most 300 cells sampled without replacement per variant condition under a fixed per-gene seed, and the five scored evaluation settings. Every model predicts a pseudobulk response, formed as the mean of its predicted profiles over the reference cells minus the measured reference mean, so that all scored entries of Supplementary Table 3 are scored through the identical PDS harness. Held-out variants contributed no cells to any parameter fit. Except for the explicitly separated scVIDR sensitivity control described below, held-out response cells also contributed no information to prediction-time conditioning or scaling.

scGen and the response-informed scVIDR sensitivity control. Both were implemented with one latent-arithmetic pipeline and differ in how the latent shift is scaled. A variational autoencoder with encoder $d \to 256 \to 256$, a 32-dimensional latent space and a mirrored decoder was trained per gene and per split on the training-variant cells together with the reference cells, minimizing mean squared reconstruction error plus a Kullback–Leibler term weighted at 0.01, with Adam at learning rate 10^{-3}, batch size 512, 120 epochs, no early stopping and a fixed seed. For each training variant the latent shift was the difference between the mean latent code of that variant's cells and the mean latent code of the reference cells. A ridge regression with $\alpha = 1$ was then fitted from the six-dimensional biophysical representation θ to those latent shifts, which is what makes the adaptation allele-conditioned: the original methods index a categorical perturbation, whereas here the shift is a continuous function of molecular variant features and is therefore defined for a variant never seen in training. For a held-out variant the predicted shift was added to the latent code of every reference cell and decoded. **scGen** uses that shift unchanged and is the response-blind entry in the main comparison.

scVIDR additionally rescales the fitted shift by $\max(r_v^{\text{scale}}, 0.05) / \text{mean}_{u \in \text{train}} |r_u^{\text{scale}}|$. The provenance of this model-specific scalar differs by gene. For TP53 and KRAS, r_v^{scale} is the AlphaMissense pathogenicity score; for all 84 TP53 and 78 KRAS nonsynonymous single substitutions, the stored values agree with the corresponding AlphaMissense entries to maximum absolute errors of 2.6×10^{-7} and 2.0×10^{-7}, respectively, and synonymous or wild-type entries are set to zero. These are response-independent external attributes. For GATA1, however, the stored value is

$$r_v^{\text{scale}} = \frac{\|\bar{X}_v - \bar{X}_{\text{WT}}\|_2}{\max_w \|\bar{X}_w - \bar{X}_{\text{WT}}\|_2},$$

where the means are calculated from the measured GATA1 expression matrix and the maximum

uses the full GATA1 cohort. Direct reconstruction over all 255 GATA1 conditions gave Pearson correlation 0.99999999997, Spearman correlation 1.0 and maximum absolute reconstruction error 5.24×10^{-8} after normalization. Thus a held-out GATA1 variant's response cells and a full-cohort response statistic enter its scaling at prediction time. All 172 distinct scored GATA1 variants in the displayed scVIDR result are affected. Accordingly, we report scVIDR as an oracle-scaled, response-informed sensitivity control rather than a valid result for prediction from molecular attributes alone. The r^{scale} field is unpopulated for every JAK1 variant, so the scaling factor is non-finite and the implementation returns no finite JAK1 prediction. The displayed PDS therefore contains 308 queries, with the 26 JAK1 queries assigned the no-ranking value of 0.5, whereas finite prediction coverage and the Pearson-δ denominator are 282 variants. A response-blind scVIDR comparison would require a consistently defined external r_v^{scale} available for every held-out variant and a new run.

Biolord. Biolord was run through its released implementation with the six θ components registered as ordered attributes and no categorical attributes, so the reference cells enter as the all-zero attribute vector. One model was trained per gene and split with a 32-dimensional latent space, a decoder of width 256 and depth 2, attribute networks of width 128 and depth 2, a 16-dimensional ordered-attribute latent, reconstruction penalty 10^4, unknown-attribute penalty 10 and unknown-attribute noise 10^{-1}; training used 100 epochs at batch size 512 with a cosine schedule, learning rate 10^{-3} for the latent, decoder and attribute networks, and a random 10% of cells held out as a validation split. No classifier heads were trained and early stopping was disabled, so the reported model is the final epoch rather than a selected checkpoint. Prediction for a held-out variant set that variant's θ on every reference cell and read the model's predicted expression.

CellFlow. CellFlow was run as an optimal-transport flow-matching model on the same arrays, with the perturbation covariate declared as a single categorical variant column whose representation is the variant's 1,280-dimensional ESM-1v embedding, and with the reference cells labelled as the control condition carrying a zero embedding. The embedding dictionary passed to the model includes the held-out variants' embeddings, which is what allows the trained model to resolve a condition it never saw during training; the held-out variants' cells are not present. The condition embedding had dimension 64, the condition encoder hidden dimensions were $512, 512, 64$ with mean pooling, and the decoder dimensions were $256, 256$; training ran for 8,000 iterations at batch size 256. Prediction pushed the reference cells forward under the held-out variant's condition.

PerturbNet. PerturbNet was evaluated using ESM-1v conditioning under all five scored within-gene settings: random holdout, positional extrapolation, mechanistic extrapolation, low-depth evaluation and compatibility. The respective numbers of gene–setting–variant evaluations were 83, 229, 51, 59 and 70, for 492 evaluations in total. Because variants can appear in more than one setting, these collapse to 303 distinct gene–variant units: 59 TP53, 51 KRAS, 172 GATA1 and 21 JAK1 variants. Its outputs and measured candidates were compared in the same 50-dimensional principal-component output representation. Scores were averaged over 15 cell-subsample seeds within each gene–setting–variant unit and then across settings for each distinct variant before aggregation. The full-candidate uniform-matching expectation remains exactly 0.5 in this representation. For the task-matched restricted-label test, prediction rows were reassigned only among held-out labels within each gene and setting while all training-plus-held-out candidate columns were retained, and the same mapping was used across the 15 seeds. The observed PDS was 0.530143, the exact conditional null mean was 0.515500 and 1,000,000 permutations gave a one-sided plus-one $P = 0.069768$.

CPA, attempted and not scored. The CPA adaptation replaced the categorical perturbation representation with the continuous six-dimensional θ encoding, leaving the out-of-distribution conditioning path otherwise intact. Its adversarial autoencoder returned a non-finite loss and no

converged model, so no prediction score is reported. The failure is specific to this adaptation.

Models without an allele-resolved conditioning interface. scGPT, GEARS and STATE condition on perturbed-gene identity through a gene token, a knowledge-graph node and a one-hot perturbation identity respectively. None can ingest a held-out variant's molecular features, and all variants of a gene collapse onto a single prediction, so the 470 variant conditions reduce to four conditioning keys. Allele-level discrimination was not evaluated because their native conditioning interfaces do not encode variant-level differences; this is an interface scope decision rather than a low or mathematically undefined PDS.

Supplementary Note 4: Controlled response geometry

Substrate and noise. The configurations compared in Fig. 6d are built on measured TP53 and KRAS cells rather than on simulated expression. Across the depths examined, the between-variant signal of these two datasets satisfies $|\rho^2| \leq 0.005$, where ρ^2 denotes the noise-corrected mean pairwise squared separation between variant profiles divided by twice the per-profile sampling noise and is defined in full below, with every interval covering zero, so they supply realistic cell-sampling noise without contributing allele-specific signal of their own. Rescaling measured profiles would not serve as a control, because it changes signal and noise together and leaves their ratio fixed; the construction therefore acts on the signal alone.

Cell partition. Each variant's cells are divided, from a single shuffle, into five disjoint groups of m cells: a direction block, a query block Q, a target block T, and two signal blocks S_1 and S_2. Q and T each subtract their own wild-type half-mean, matching the convention used for the empirical split-half reference, while S_1 and S_2 share one, which therefore cancels from their difference. The reproducibility reference is measured on Q and T and the separation statistics on S_1 and S_2, so the two axes of every point come from different cells. The direction block enters no score; a direction correlated with the noise it is later added to would produce the relationship by construction rather than measure it.

Constructed separation. Directions are the mean profile of each variant's direction block, centred across variants so that the injected signal contains no component shared across all variants, and normalized to unit length. Their geometry is then reshaped in two ways: the direction set is truncated to a rank $r \in \{\text{full}, 5, 3, 2, 1\}$ by singular value decomposition, and each variant's direction u_v is multiplied by an independent lognormal amplitude a_v with $\log a_v \sim \mathcal{N}(0, s^2)$, $s \in \{0, 1, 2\}$, so $s = 0$ leaves the amplitudes equal. The reshaped set is then rescaled so that its mean pairwise squared separation equals exactly 2, which fixes the mean while leaving the geometry free, and multiplied by $a = \sqrt{\tau\, \hat{\eta}_0^2}$, where $\hat{\eta}_0^2$ is the substrate's own per-profile noise measured before any signal is added and $\tau \in \{0, 0.05, 0.2, 0.6, 2, 6\}$ is the target. The resulting mean pairwise squared separation is $2\tau\hat{\eta}_0^2$, so the configuration is constructed at $\rho^2 = \tau$ by construction. The same vector is added to Q, T, S_1 and S_2.

Factor grid. Two substrates, three sampling depths (25, 50 and 100 cells per group), three candidate-set sizes including the target (20, 45 and 90 variants), six signal levels, five ranks and three amplitude dispersions define 1,620 configurations. Neither substrate contains 90 variants with five disjoint groups of 100 cells, so 180 configurations are unavailable and 1,440 configurations remain. Each configuration is scored with the same tie-aware PDS harness used for every model and for the empirical split-half reference, comparing Q against T, and averaged over ten random cell splits.

Normalized separation metrics. Both statistics are defined in full in Methods and are restated here in the notation of this note. With $H = 2$ signal blocks $S_{v,1}, S_{v,2}$ per variant, block mean $\bar{S}_v$ and

n variants, the per-profile noise is $\hat{\eta}^2 = n^{-1}\sum_v (H-1)^{-1}\sum_h \|S_{v,h} - \bar{S}_v\|^2$. Mean separation relative to noise is

$$\hat{\delta}^2 = \binom{n}{2}^{-1} \sum_{i<j} \|\bar{S}_i - \bar{S}_j\|^2 - \frac{2\hat{\eta}^2}{H}, \qquad \rho^2 = \frac{\hat{\delta}^2}{2\hat{\eta}^2}.$$

Nearest-competitor separation replaces the mean over pairs with each variant's closest competitor, and the selection is cross-fitted in both directions and averaged:

$$s_v = \tfrac{1}{2}\left(\|S_{v,2} - S_{j_v^{(1)},2}\|^2 + \|S_{v,1} - S_{j_v^{(2)},1}\|^2\right) - 2\hat{\eta}^2, \qquad j_v^{(h)} = \arg\min_{u \neq v} \|S_{v,h} - S_{u,h}\|^2.$$

Selecting the competitor on one block and reading its distance on the other removes the downward bias of taking a minimum and measuring it on the same data. The subtracted term is $2\hat{\eta}^2$ here and $2\hat{\eta}^2/H$ for the mean statistic, because the nearest-competitor distances are between single measurements whereas the pairwise distances are between block means; the two corrections are not interchangeable. Cross-fitting prevents reuse of the same cell-sampling fluctuation for selection and evaluation but does not by itself make the two estimates statistically independent or ensure zero expectation after subtracting one pooled noise term. Zero is the nominal no-signal expectation only under a working model in which per-profile sampling errors are independent, mean-zero and exchangeable with common variance across variants. Under variant-specific heteroskedasticity, the pooled correction need not centre every selected pair. The s_v are averaged across the ten cell splits before being summarized as $\rho^2_{\mathrm{nn}} = \mathrm{med}_v(s_v)/(2\hat{\eta}^2)$. Noise correction can send an individual separation below zero near the measurement floor; the median tolerates this, whereas a geometric mean does not and is reported as undefined rather than floored.

Scope of the reported correlations. The Spearman correlations quoted for this comparison are computed over all 1,452 plotted points, that is the 1,440 constructed configurations together with the twelve measured gene and depth conditions. Restricted to the 1,440 constructed configurations alone they are 0.763 for mean separation and 0.961 for nearest-competitor separation.

Supplementary Note 5: Broader-panel sensitivity and the sampling-depth projection

Sensitivity of the external perturbation panel. The arms below test the sensitivity of the panel's detection fractions to analysis choices (Supplementary Fig. 2). Varying the data loader, the dimension-reduction path (exact, incremental and randomized), the per-perturbation cell cap, the feature width and the base random seed moved the detection fraction by at most 0.021 on the same cohorts, and left every configuration on the same side of the prespecified operating points. The choice of expression representation moved it by 0.097 to 0.195, an order of magnitude more than any of those arms. The direction of that difference matters for how the panel should be read: on the three cohorts evaluated under both pipelines, the log-normalized, highly-variable-gene pipeline returned the *higher* detection fractions, so on those cohorts the prespecified count-space criterion used throughout Fig. 5 returned the lower ones. Detection fractions are therefore compared only within a pipeline, and the representation is stated as part of the criterion.

The sampling-depth projection law. The projection reported in Fig. 6e inverts the controlled-geometry sweep of Supplementary Note 4. This note gives the calibration in full; the construction is summarized in Methods.

Choice of axis. Four summaries of measured response geometry were scored by Spearman correlation with the split-half reproducibility reference over all 1,452 sweep configurations: mean

pairwise separation ρ^2 (0.767), median nearest-competitor separation ρ^2_{nn} (0.961), its geometric-mean counterpart (0.818) and the fraction of perturbations whose nearest competitor exceeds the noise (0.950). The median nearest-competitor separation was taken as the axis on that evidence rather than by assumption.

Monotonicity screen. Three quantities were candidates for inversion: the split-half reproducibility reference, the probability of recovering the complete design-defined ordering of the graded predictor family, and the probability of selecting its best member. Each was scored by its Spearman correlation with the axis restricted to configurations above the median axis value. Only the reproducibility reference was retained ($+0.918$). Both recovery probabilities turn over in that range (-0.280 and -0.248, with the turnover near $\rho^2_{\mathrm{nn}} = 0.043$, at a reference of about 0.705), so a monotone calibration fitted to either would return incorrect depths in the high-resolution regime. We therefore did not invert either recovery probability; only the split-half reproducibility reference was retained as a monotone calibration outcome.

Calibration. Configurations were stratified by candidate-set size into $n \leq 30$, $30 < n \leq 70$ and $n > 70$, holding 543, 541 and 368 configurations, a stratum being dropped if it held fewer than 20. No functional form was assumed within a stratum. The axis was cut at its empirical quantiles at levels $0, 1/24, \ldots, 1$; each bin holding at least three configurations contributed a knot at the median axis value and the median reference of that bin; and the knot ordinates were then replaced by their running maximum. All three strata yielded 24 knots, reaching maximum references of 1.000, 1.000 and 0.997 respectively. The strata contain the constructed configurations together with the twelve measured gene-by-depth configurations, so the projections reported for those datasets are not strictly out of sample.

Inversion and refusal. For a target τ the required axis value $x^*(n, \tau)$ is the smallest knot abscissa whose monotonized ordinate reaches τ, with no interpolation between knots, and the depth projected from a pilot at m_0 cells per group is $\hat{m}(\tau) = m_0\, x^*(n, \tau)/\rho^2_{\mathrm{nn}}(m_0)$. No depth is returned when the stratum held fewer than 20 configurations, when the calibration never reaches τ, when the pilot estimate is not positive, or when the pilot's lower confidence bound does not exceed zero. Applied to the four datasets at 25, 50 and 100 cells per group and at $\tau \in \{0.7, 0.9\}$, the law returned a finite depth for 8 of 24 combinations: all six for JAK1 (10.4 and 55.1 cells per group from the 25-cell pilot, 12.2 and 64.9 from the 50-cell pilot, 18.6 and 98.9 from the 100-cell pilot) and two for GATA1 (62.1 and 231.5, both from the 25-cell pilot), and none for TP53 or KRAS. In every refused combination the pilot's lower confidence bound failed to clear zero. The first two refusal conditions never applied.

Domain and uncertainty. The constructed configurations span candidate-set sizes of 20, 45 and 90, whereas the eight finite projections were read off at 247 for GATA1 and at 14, 9 and 5 for JAK1, so every reported projection extrapolates in candidate-set size. No interval is placed on a projected depth: the pilot lower bound acts as a gate and is not propagated. A projection from this law is therefore a model-based design estimate rather than a measured requirement, pending extension of the calibration grid to the candidate-set sizes at which it is read.