Measurement resolution constrains fine-grained perturbation prediction

Bo Li[1,2,†], Chengyang Zhang[3,4,†], Mengran Li[5,6,†], Bob Zhang[1,7,✉], Lin Wang[8,9], Zhenchao Tang[10], Jun Liu[11], Chengliang Liu[1], Chen Wei[12], Yuhao Yi[3,4,✉], Jiancheng Lv[3], Yang Zhang[2,8,9,11,13,✉]

[1]PAMI Research Group, Department of Artificial Intelligence, University of Macau, Macau SAR; [2]Department of Computer Science, School of Computing, National University of Singapore, Singapore, Singapore; [3]College of Computer Science, Sichuan University, Chengdu, China; [4]Department of Pathology and Institute of Clinical Pathology, West China Hospital, Sichuan University, Chengdu, China; [5]Shenzhen Loop Area Institute, Shenzhen, Guangdong, China; [6]School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, China; [7]Centre for Artificial Intelligence and Robotics, Institute of Collaborative Innovation, University of Macau, Macau SAR; [8]Center for AI and Computational Biology, Institute of Systems Medicine, Chinese Academy of Medical Sciences, Suzhou, China; [9]State Key Laboratory of Common Mechanism Research for Major Diseases, Institute of Systems Medicine, Chinese Academy of Medical Sciences, Suzhou, China; [10]AI for Life Sciences Lab, Tencent, Shenzhen, China; [11]Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore; [12]School of Economics and Management, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi, China; [13]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore.

[†]These authors contributed equally to this work.
[✉]Corresponding authors: Bob Zhang (bobzhang@um.edu.mo), Yuhao Yi (yuhaoyi@scu.edu.cn) and Yang Zhang (zhang@nus.edu.sg).

Abstract

Virtual-cell models are increasingly evaluated at finer biological resolution, yet the experimental responses used as prediction targets may not reproducibly resolve the distinctions that models are asked to predict. Here we introduce PertResolve, a measurement-resolution framework that separates perturbation detection, perturbation identification and response prediction, and interprets model performance against the distinctions reproducibly supported by the measurement. We examine this problem using the allele-resolved arm of PertResolve-Bench, comprising 470 coding-variant conditions in TP53, KRAS, GATA1 and JAK1 and 321,043 cells. Feature-based predictors and released perturbation-prediction frameworks recovered substantial transcriptional direction, but this recovery was dominated by gene-shared structure and did not translate into reliable full-pool allele identification. Split-half references showed that similarly weak allele-identification scores can arise for different reasons: TP53 and KRAS were measurement-limited at the evaluated allele resolution, whereas JAK1 contained substantial reproducible allele-specific structure that current models recovered only partially. We then extended PertResolve-Bench to 31 configurations from 14 public perturbation resources, showing that detection depended on perturbation signal relative to within-condition variation, whereas identification depended on separation from nearby competitors. Repeated pilot–evaluation analyses provided an aggregate diagnostic for early detection triage, while reliable model ranking depended jointly on measurement reproducibility, model separation and benchmark size. Measurement resolution therefore constrains which biological distinctions fine-grained perturbation prediction can meaningfully evaluate and provides a practical basis for perturbation-benchmark design.

Introduction

Virtual-cell models aim to predict how cellular states change in response to genetic, chemical and environmental perturbations, with the broader goal of simulating cellular behaviour across molecular and cellular contexts[1–3]. Current approaches include generative models[4,5], graph-based predictors[6], flow-matching methods[7] and pretrained single-cell models[8–10]. For genetic perturbations, however, widely used benchmarks remain largely gene-centred, encoding an intervention by the targeted gene[11–13]. This

abstraction collapses protein-coding variants of the same gene even though such variants can differ in protein structure, biochemical activity, molecular interactions and downstream regulation[14–16]. A model can therefore recover a transcriptional programme shared across variants while capturing little of the response that distinguishes one variant from another. At variant-level resolution, evaluation thus depends not only on what a model predicts, but also on whether the experiment can reproducibly distinguish the variant-specific responses being compared.

Variant-resolved single-cell assays now make this problem experimentally accessible by measuring high-dimensional transcriptional responses to many coding variants in parallel[17–20]. Here, allele denotes an experimentally assigned protein-coding variant condition rather than a phased genomic haplotype. Whereas multiplexed assays of variant effect and deep mutational scanning typically summarize each variant by a fitness, activity or molecular-function score[21–23], allele-resolved perturbation prediction seeks to predict the variant's high-dimensional cellular response. At this resolution, a high response-similarity score can reflect recovery of a gene-shared programme without showing that the correct allele has been resolved. This motivates separating three questions: whether a perturbation can be reproducibly distinguished from its reference, whether it can be distinguished from sibling alleles, and whether the response of a held-out allele can be predicted from its molecular properties.

This separation exposes a prerequisite that perturbation benchmarks usually leave implicit: the experimental measurement must itself reproducibly support the biological distinction that the benchmark is intended to evaluate. Measured response profiles are finite-sample estimates whose reproducibility depends on experimental design and within-condition variation[24–26]. If the requested distinction cannot be reproduced across disjoint cell samples, evaluation at that resolution is measurement-limited. If the distinction is reproducibly present in the data but remains unrecovered by the predictor, the setting is model-limited. Measurement resolution therefore determines the level of biological specificity at which model scores can be meaningfully interpreted. Coding variants provide a stringent setting in which to expose this problem, but the same principle applies whenever perturbation benchmarks are asked to distinguish closely related perturbations or cellular states.

Here we introduce PertResolve, a measurement-resolution framework for fine-grained perturbation prediction, together with PertResolve-Bench and PertResolve-Eval. PertResolve-Bench combines an allele-resolved benchmark of 470 protein-coding variant conditions in TP53, KRAS, GATA1 and JAK1, comprising 321,043 cells profiled by three single-cell perturbation technologies, with a broader panel of 31 perturbation configurations from 14 public resources. In the allele-resolved setting, substantial recovery of transcriptional direction coexisted with near-chance allele identification, and similarly weak allele-identification scores arose for different reasons: some settings were limited by the measurements, whereas others contained reproducible allele-specific structure that current models recovered only partially. We then extended the same analysis beyond coding variants, showing that detection depends on perturbation signal relative to within-condition variation, whereas identification depends on separation from nearby competitors. Finally, we show how pilot measurements, candidate composition, sampling depth and benchmark size inform different aspects of perturbation-benchmark design. Together, these analyses establish measurement resolution as a prerequisite for interpreting fine-grained perturbation prediction and as an explicit component of perturbation-benchmark design.

Results

Allele-level model evaluation requires reproducibly resolved measurements

Allele-level evaluation separates three questions (Fig. 1a): whether a perturbation is reproducibly detectable relative to its reference, whether sibling alleles are distinguishable from one another, and whether a model can predict the response of a held-out allele from molecular features. Detection and identification characterize the resolution of the measurement, whereas prediction is the model-level problem. Model performance is interpreted only against distinctions that the measurement reproducibly supports.

The allele-resolved arm of PertResolve-Bench comprises 470 protein-coding variant conditions across TP53, KRAS, GATA1 and JAK1, together with their dataset-specific wild-type or control cells, totalling 321,043 cells profiled by Perturb-seq, base editing and scSNV-seq (Fig. 1b,c)[17–19]. The four datasets differ jointly in gene, assay and cellular context, and we therefore treat them as four measurement regimes

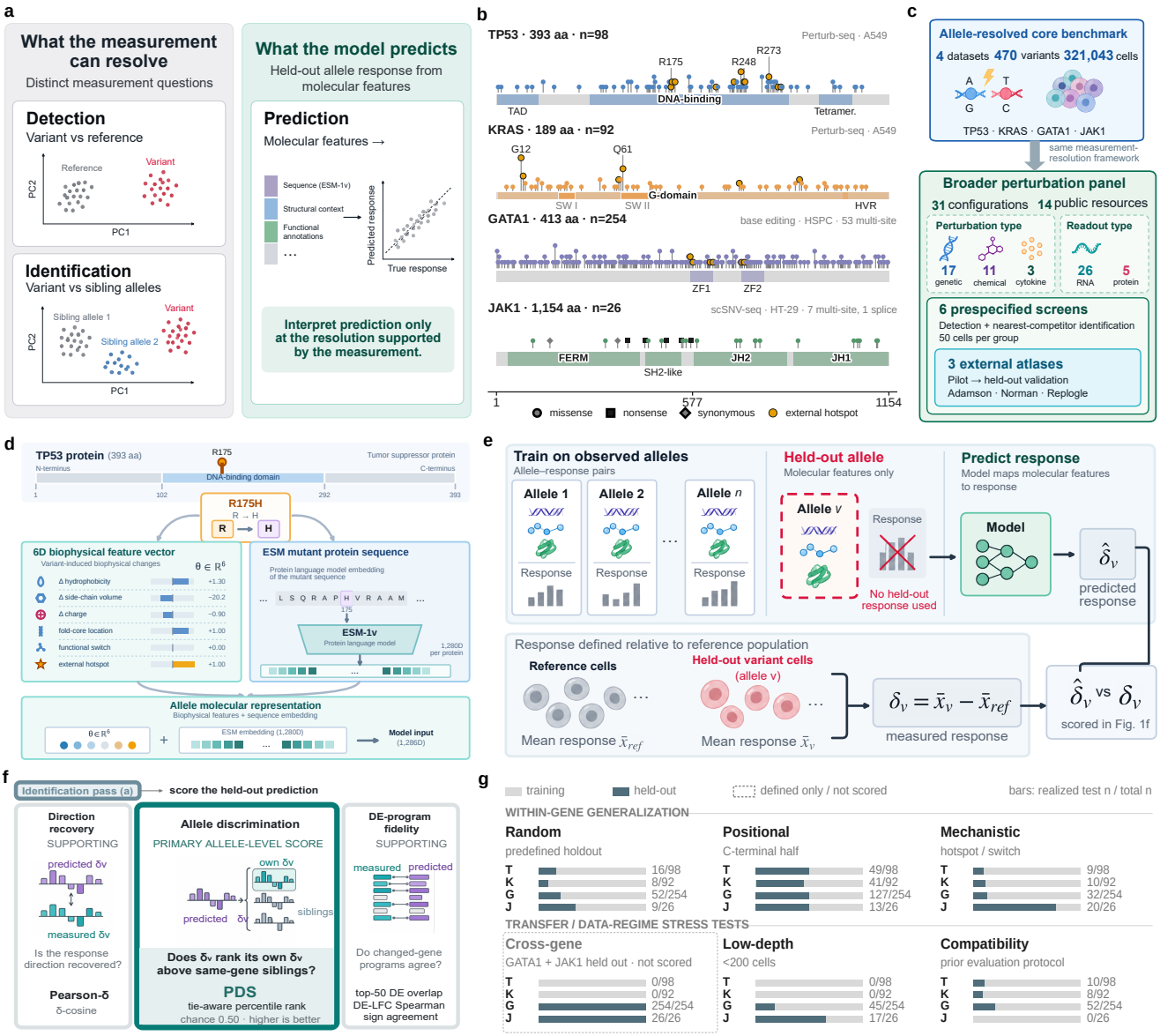

Figure 1: **PertResolve separates measurement resolution from fine-grained perturbation prediction. a**, PertResolve separates perturbation detection relative to the reference, identification among competing perturbations and prediction of a held-out perturbation. Detection and identification characterize measurement resolution and are answered separately; prediction is the model-level task. Together they define the biological distinctions that a comparison can support. **b**, The allele-resolved arm of PertResolve-Bench comprises 470 protein-coding variant conditions across TP53, KRAS, GATA1 and JAK1 profiled by Perturb-seq, base editing and scSNV-seq. Marker shape denotes variant consequence; amber marks external hotspot annotations. **c**, PertResolve-Bench comprises two complementary analysis arms: an allele-resolved benchmark of four datasets, 470 variants and 321,043 cells, and a broader perturbation panel of 31 configurations from 14 public resources, including six prespecified screens and the three gene-perturbation atlases used for pilot validation. **d**, Molecular representations of a held-out allele: a six-dimensional biophysical vector θ and protein-language-model embeddings derived from the mutant sequence. **e**, Prediction task. Variant and reference cells define the measured response δ_v; the model predicts $\hat{\delta}_v$ from molecular features without access to the held-out variant's response cells. **f**, PertResolve-Eval reports transcriptional direction recovery (Pearson-δ), full-pool allele identification (PDS) and differential-expression fidelity. PDS, the primary allele-level score, ranks the true measured response among same-gene candidates; the other two are supporting axes. **g**, Evaluation settings, grouped as within-gene generalization and transfer or data-regime stress tests. Four primary scored settings test random holdout, positional extrapolation, mechanistic extrapolation and low-depth evaluation; the compatibility setting reproduces a prior evaluation protocol, and cross-gene transfer is defined but not scored because the datasets share no common expression space (Methods). Bars show held-out fractions and counts.

rather than as four genes. PertResolve-Bench additionally includes a broader panel of public perturbation screens used later to evaluate measurement resolution beyond allele-resolved prediction (Fig. 1c; Supplementary Data 1).

The prediction task holds out a coding variant from model training and asks the model to predict

its population-average transcriptional response from molecular properties alone (Fig. 1d,e). Variants of the same gene remain distinct targets: TP53 R175H, R273C and R248W, for example, are predicted separately rather than pooled into a single TP53 response. For variant v, the target δ_v is the pseudobulk expression difference between variant and dataset-specific reference cells, and the model predicts $\hat{\delta}_v$ from biophysical or protein-language-model features without using the held-out variant's response cells (Methods).

Predictions are evaluated on complementary axes of response recovery and allele identification (Fig. 1f). Pearson-δ measures agreement with the observed transcriptional direction. The perturbation discrimination score (PDS) ranks the true measured response among all same-gene candidates and therefore tests full-pool allele identification; chance is 0.50. Differential-expression fidelity provides secondary biological readouts. Four primary scored settings test random holdout, positional extrapolation, mechanistic extrapolation and low-depth evaluation; a fifth compatibility setting provides a protocol comparison for PerturbNet (Fig. 1g; Methods). We first asked whether predictors recover allele-specific information beyond the gene-shared perturbation programme.

Perturbation direction is recovered while full-pool allele identification remains near chance

Scoring each held-out prediction on both axes separated direction recovery from allele identification (Fig. 2a). We evaluated 18 feature-based predictors spanning six model heads and three feature spaces, together with the Gene-mean baseline and the WT-null reference (Fig. 2b,c). Across the 19 non-null predictors, Pearson-δ ranged from 0.58 to 0.67, whereas PDS ranged from 0.47 to 0.49; no predictor showed reliable above-chance full-pool allele identification. Substantial recovery of transcriptional direction was thus accompanied by no reliable identification of the corresponding allele among same-gene variants.

Response decomposition explained the divergence. Within each cell partition, we separated the measured profiles into their mean across same-gene variants and the deviations from that mean. Averaged across partitions, the allele residual contributed 28%, 41%, 71% and 49% of the mean squared response norm in TP53, KRAS, GATA1 and JAK1, respectively (Fig. 2d; Methods). These descriptive shares include sampling variation and do not quantify reproducible allele signal. After measured and predicted profiles were centred separately within each evaluated gene cohort, residual Pearson-δ was near zero across predictors (-0.08 to 0.03). The Gene-mean baseline, which contains no allele-specific molecular information, matched the direction recovery of the best feature-based predictor (0.668 versus 0.673). Residual PDS, computed separately using a training-only gene mean, likewise remained close to its calibrated null (Fig. 2e). The direction recovered by the evaluated predictors was therefore dominated by the transcriptional programme shared across variants of the same gene.

Differential-expression fidelity showed the same loss of information with increasing biological specificity (Fig. 2f). Within the per-variant sets of 50 genes with the largest absolute Welch t-statistics, predicted and observed effects had the same sign for 68% of genes on average. Agreement was weaker for effect-size ordering (DE effect-size Spearman $\rho = 0.27$) and for the identity of responding genes (top-50 DE overlap, 15%). The dissociation held in every allele-resolved dataset: mean Pearson-δ ranged from 0.43 to 0.76, whereas mean PDS remained between 0.45 and 0.47. The evaluated predictors thus recover a coarse, gene-shared perturbation programme while capturing little of the structure needed for full-pool allele identification.

Full-pool allele identification remains weak across variant representations and model classes

The dissociation between direction recovery and full-pool identification also persisted across richer molecular representations. Protein-language-model embeddings, mutation-aware ESM2 representations and AlphaMissense features preserved direction recovery without producing reliable full-pool allele identification (Supplementary Note 2).

Released perturbation-prediction frameworks evaluated on the same held-out variants with the same allele-level scoring showed the same limit (Fig. 2g; Supplementary Table 3). Biolord and CellFlow, together with the allele-conditioned adaptation of scGen, yielded PDS values of 0.47–0.49 with confidence intervals overlapping 0.50. PerturbNet produced a PDS of 0.53 across five evaluation settings and 303 distinct variants. Its predictions and targets were scored in the model's 50-dimensional output representation, but output dimensionality does not itself alter the analytic PDS chance expectation. In a task-

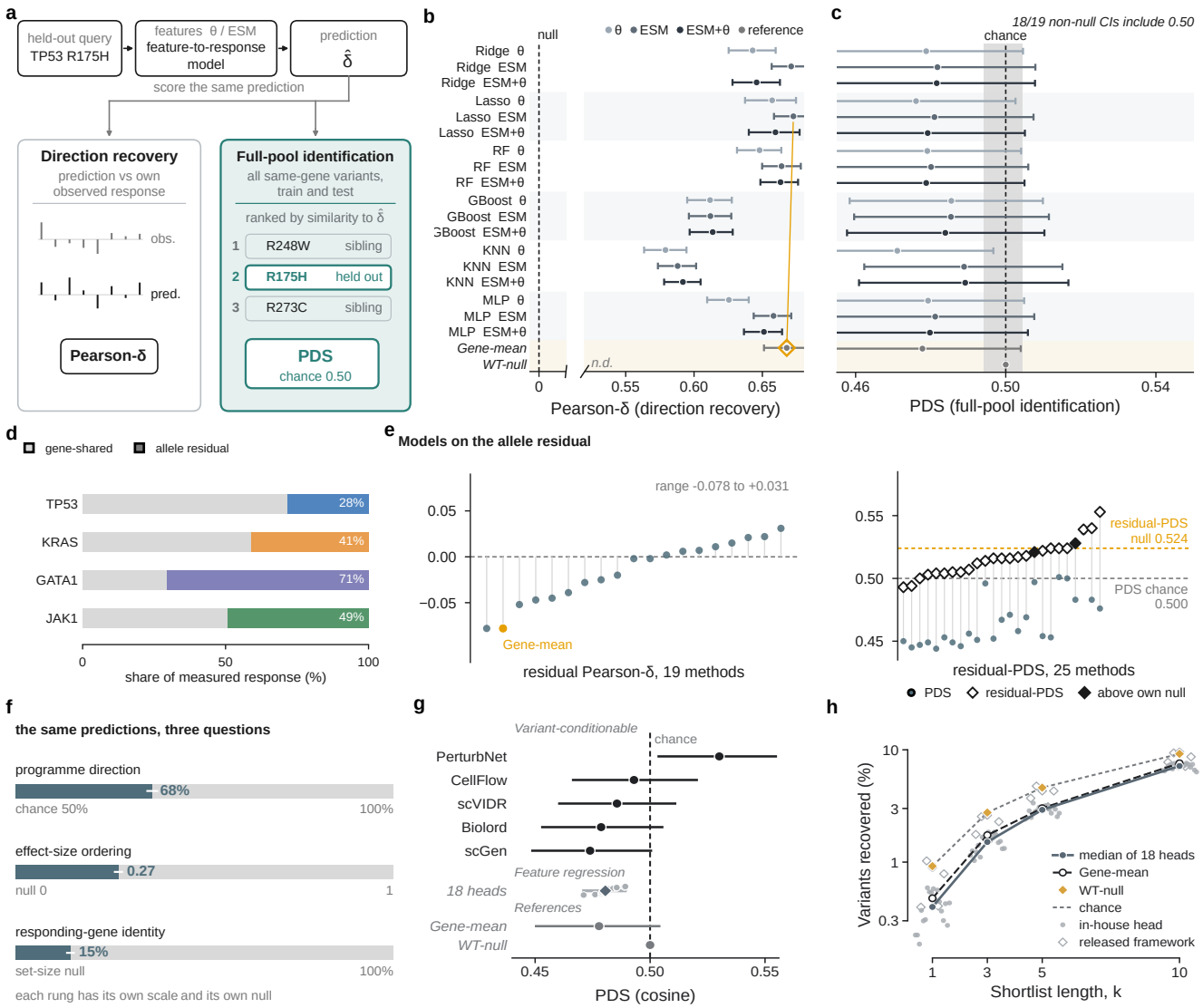

Figure 2: **Evaluated predictors recover gene-shared response while full-pool allele identification remains near chance. a**, Complementary evaluation by transcriptional direction recovery (Pearson-δ) and full-pool allele identification (PDS; chance, 0.50). **b**, Pearson-δ for 18 feature-based predictors and two references. Points, means; whiskers, 95% bootstrap confidence intervals over distinct held-out variants ($n = 308$, or 305 for sequence-dependent predictors). Gene-mean is highlighted; n.d., undefined for WT-null. **c**, PDS for the same predictors; intervals as in **b**. Grey band, range of held-out-only permutation means across method–gene–setting–distance combinations (0.494–0.505), used as an implementation check rather than a task-matched full-candidate inferential reference. **d**, Gene-shared and allele-residual shares of the mean squared measured-response norm, averaged over 15 cell partitions. Sample sizes are 98 TP53, 92 KRAS, 254 GATA1 and 26 JAK1 variants (Methods). **e**, Left, descriptive residual Pearson-δ after separate cohort centring. Right, full-response PDS (circles) and training-centred residual PDS (diamonds). Filled diamonds, unadjusted permutation $P < 0.05$; multiplicity adjustment and permutation resolution are described in Methods. Lines mark the respective nulls. **f**, Differential-expression fidelity: sign agreement, DE effect-size Spearman correlation and top-50 DE overlap. Each metric uses its own scale and null. Bars, means across 18 predictors; white ticks, ranges. **g**, Multi-seed PDS for five released-framework implementations; intervals bootstrap each method's distinct-variant query cohort, with coverage reported in Supplementary Table 3. Lower points, feature-based predictors; diamond, their median. The PerturbNet estimate aggregates five evaluation settings over 303 distinct variants; its task-matched restricted-label permutation test gives $P = 0.070$. The displayed scVIDR adaptation is a response-informed sensitivity control, not a response-blind held-out-variant result. **h**, Top-k allele shortlisting. Filled circles, feature-based predictors; open diamonds, released-framework implementations; line, feature-based median; references shown separately. The scVIDR point retains the response-informed sensitivity-control status specified in **g**. Dashed line, full 308-query reference-cohort mean of the query-specific chance $\min(k/N_v, 1)$, where N_v is candidate-pool size; inference uses each method's coverage-matched query-specific null. No response-blind method's paired 95% bootstrap interval exceeds that null (Methods).

matched test that permuted only held-out labels within each gene and setting while retaining all training and held-out candidate columns, the conditional null mean was 0.5155 and the one-sided permutation value was $P = 0.0698$; the observed PDS therefore did not exceed this task-matched conditional reference at the 0.05 level. The displayed scVIDR adaptation is not part of this response-blind comparison because its GATA1 scaling variable was computed from the held-out variant's measured response; it is retained only as a response-informed sensitivity control (Supplementary Note 3). Gene-keyed interfaces such as scGPT, GEARS and STATE condition on perturbed-gene identity, so sibling alleles receive the same perturbation representation and the 470 variant conditions collapse onto four conditioning keys. Their allele-level scores were not evaluated because their native conditioning interfaces do not encode variant-level differences (Supplementary Note 3).

A more operational task, placing the correct allele in a shortlist of k candidates, yielded the same conclusion (Fig. 2h). No response-blind method exceeded its query-matched chance expectation at $k = 1, 3, 5$ or 10; the response-informed scVIDR sensitivity control did not do so either. Weak full-pool identification is therefore not confined to a particular molecular representation, model class or ranking granularity. A near-chance model score alone, however, cannot distinguish failure to learn reproducible allele structure from a measurement that does not resolve that structure. We therefore quantified the allele-level resolution of the measurements themselves.

Measurement resolution separates measurement-limited from model-limited allele prediction

To place measurement reproducibility and model performance on the same allele-identification scale, we constructed a split-half reproducibility reference: each held-out variant's cells were divided into disjoint groups, and one half was scored against the other with the same full-pool PDS evaluation used for the prediction models (Fig. 3a,b).

The split-half reference revealed sharply different measurement regimes. TP53 and KRAS remained indistinguishable from chance, so these measurements did not reproducibly support full-pool allele identification at the evaluated depth. GATA1 showed modest reproducible structure, whereas JAK1 showed strong allele-level reproducibility; the measurement–model gap was 0.069 PDS in GATA1 and 0.359 in JAK1. The detection-window ratio lay at the sampling-variation boundary for TP53 and KRAS, was intermediate for GATA1 and was substantially lower for JAK1 (Fig. 3c). No sibling pairs were reproducibly separated in TP53 or KRAS, compared with 3.8% in GATA1 and 78.5% in JAK1, and 15.4% of JAK1 variants also resolved their nearest sibling (Fig. 3d; Supplementary Table 4).

Among measurement-resolvable sibling pairs, the evaluated response-blind models recovered substantial allele-specific information (Fig. 3e). The split-half measurement discriminated these pairs with accuracies of 0.988 in GATA1 and 1.000 in JAK1, whereas the highest response-blind point estimates were 0.780 for Ridge-ESM in GATA1 and 0.762 for Biolord in JAK1. The response-informed scVIDR sensitivity control reached 0.830 in GATA1 but is not a held-out-response-blind result. A variant-node bootstrap sensitivity analysis, which accounts for pairs sharing endpoints, widened the intervals but retained negative paired model-minus-measurement gaps for the highest response-blind point estimates: -0.208 (95% interval, -0.382 to -0.093) in GATA1 and -0.238 (-0.406 to -0.117) in JAK1 (Supplementary Note 2). Thus, response-blind models recover allele-specific structure where the measurement supports the comparison, without reaching the measurement reference. Pairwise discrimination tests one specified sibling contrast, whereas full-pool PDS requires ranking the target against all same-gene alternatives.

Between pairwise discrimination and full-pool identification lies the practical question of whether the measurement can place the correct allele in a useful shortlist (Fig. 3f). At $k = 10$, the split-half reference recovered 83.3% of JAK1 variants compared with a 38.5% chance rate, whereas TP53 and KRAS remained near chance. GATA1 illustrates the distinction between local and many-way resolution: although its small subset of resolvable sibling pairs was discriminated almost perfectly, the correct allele appeared in the top five for only 4.1% of variants, compared with 2.0% by chance. High discrimination on selected pairs therefore does not imply useful identification across the full candidate pool.

Similar near-chance model PDS values therefore arise in fundamentally different regimes (Fig. 3b,g). TP53 and KRAS are measurement-limited at allele resolution; JAK1 is model-limited, because reproducible allele structure substantially exceeds model performance; GATA1 is intermediate, with a small

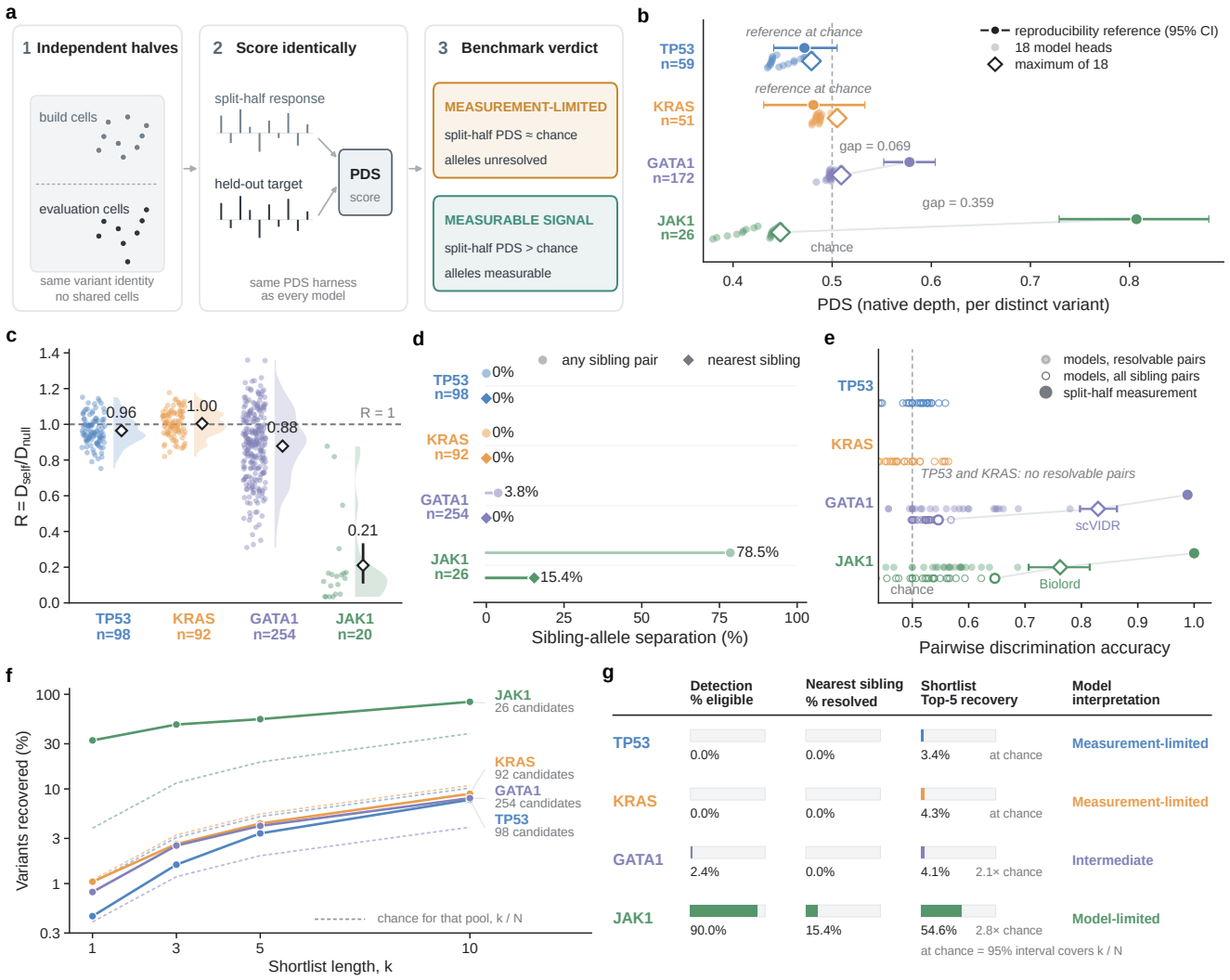

Figure 3: **Split-half measurement resolution distinguishes measurement-limited from model-limited allele prediction. a**, Split-half reproducibility reference: disjoint cell groups supply the query and target responses for the same full-pool PDS evaluation used for models. **b**, Reference and 18 feature-based predictors at native depth. Reference whiskers, 95% bootstrap confidence intervals over distinct held-out variants (n at left); open diamond, best predictor; dashed line, chance PDS (0.50). **c**, Detection-window ratio $R = D_{\text{self}}/D_{\text{null}}$, relating within-variant variation to variant-to-reference separation. Diamonds, dataset means; whiskers, 95% bootstrap confidence intervals; n, variants evaluable at native depth; $R = 1$, equal variation and signal. **d**, Fractions of sibling pairs separated beyond within-variant variation (circles) and variants resolving their nearest sibling (diamonds). **e**, Pairwise discrimination accuracy. Small filled markers show models on pairs independently identified as measurement-resolvable; small open markers show models on all scored pairs. The large filled circle shows the split-half measurement on resolved pairs. The labelled diamond marks the highest displayed implementation on resolved pairs, joined to its all-pairs value; the GATA1 diamond is the response-informed scVIDR sensitivity control and is not part of the response-blind comparison. Whiskers, descriptive 95% percentile-bootstrap intervals over 425 GATA1 and 116 JAK1 resolved pairs; pairs can share variants, so a variant-node bootstrap sensitivity analysis is reported in Supplementary Note 2. Chance, 0.50. TP53 and KRAS have no resolved pairs and show only all-pairs scores. **f**, Split-half top-k recovery: fraction of held-out variants whose true allele ranks among the top k same-gene candidates. Dashed curves, chance k/N for each dataset (N = 26–254 candidates); logarithmic axis. **g**, Interpretation across identification tasks. *At chance* denotes a 95% recovery interval covering k/N. TP53 and KRAS are measurement-limited, GATA1 intermediate and JAK1 model-limited at the evaluated depth.

resolvable subset but weak full-pool resolution. Direction recovery orders the datasets almost oppositely: mean Pearson-δ decreases from 0.76 in TP53 to 0.43 in JAK1, while the fraction of reproducibly resolved sibling pairs increases from 0% to 78.5%. The dataset on which predictions appear strongest by direction recovery is the one that provides the least allele-level measurement resolution.

Measurement resolvability depends on perturbation signal relative to within-condition variation

We next asked what determines these measurement regimes. Detection depends on the separation between a perturbation and its reference relative to the variation obtained by repeatedly measuring the perturbation itself. We summarize this window by $R = D_{\text{self}}/D_{\text{null}}$, where D_{self} is the split-half within-perturbation distance and D_{null} the perturbation-to-reference distance (Fig. 4a). Values near $R = 1$ indicate perturbation signal comparable to within-condition sampling variation, and smaller values a more clearly resolved response. Representative TP53 and JAK1 variants illustrate closed and open measurement windows (Fig. 4b).

Across the four allele-resolved datasets, nominal cell count did not order measurement resolution. Larger pseudobulk responses were associated with lower detection-window ratios (Fig. 4c). TP53 and KRAS had median depths of 929 and 1,000 cells per variant yet no detection-eligible variants, whereas JAK1 had a median of 104 cells per variant and 90% of its variants were detection-eligible; GATA1 was intermediate at 2.4% (Fig. 4d). Subsampling 30–150 cells per half with perturbation identity held fixed preserved these regimes: TP53 and KRAS remained unresolved, GATA1 weakly resolved, and JAK1 was strongly resolved at the shallowest tested depth. Sampling depth over this range therefore did not explain the between-dataset differences in measurement resolution.

Detection and identification depend on different response separations. Detection asks whether a perturbation separates from its reference, whereas allele identification depends on separation from nearby competing perturbations. Estimated cross-fitted nearest-competitor separation relative to within-condition sampling variation followed the same ordering (Fig. 4e). TP53 and KRAS showed no reproducibly positive nearest-competitor separation at 25, 50 or 100 cells per group, whereas JAK1 remained positive at all three depths; GATA1 lay near the boundary. The datasets that differed in detection therefore also differed in the local response geometry required for allele identification.

The relative measurement regimes were preserved across five distance families and alternative detection criteria (Fig. 4f; Supplementary Fig. 1; Supplementary Note 1). Because detection and nearest-competitor resolution are defined from perturbation responses and their references rather than allele-specific annotations, PertResolve applies beyond coding variants. We therefore evaluated the broader perturbation arm of PertResolve-Bench across public genetic, chemical and cytokine perturbation resources.

Measurement resolution varies widely across public perturbation screens

We applied the same prespecified detection criterion to 31 perturbation configurations nested within 14 public resources, spanning genetic, chemical and cytokine perturbations and RNA and surface-protein readouts (Fig. 5a; Methods); several resources contribute more than one context or readout, so the configurations are nested within resources rather than independent. The fraction of perturbations meeting the detection criterion spanned the full range from 0 to 1 (Fig. 5b), placing the allele-resolved datasets within a broad continuum of perturbation measurement resolution.

Across configurations, nominal sampling depth did not order detection. Among the 25 configurations with recorded cell counts, detection was not ordered by cells per perturbation (Fig. 5c): the deepest configuration, with 13,080 cells per perturbation, had a detection fraction of 0.256, whereas a configuration with 1,587 cells per perturbation reached 1.000. Matched RNA–protein comparisons showed no consistent ordering by readout modality, with protein resolving more perturbations in two comparisons and RNA in three (Fig. 5d). Screen size did not order detection either. Across all 31 configurations the number of perturbations spanned 7 to 1,500 (Fig. 5e): a configuration of 96 perturbations reached a detection fraction of 1.000, whereas among the ten largest configurations, which carry 414 to 1,500 perturbations, the fraction ranged from 0.008 to 0.873. Neither axis of screen scale is a proxy for whether a screen can be evaluated.

The resolution assigned to a screen also depends on how its expression measurements are repre-

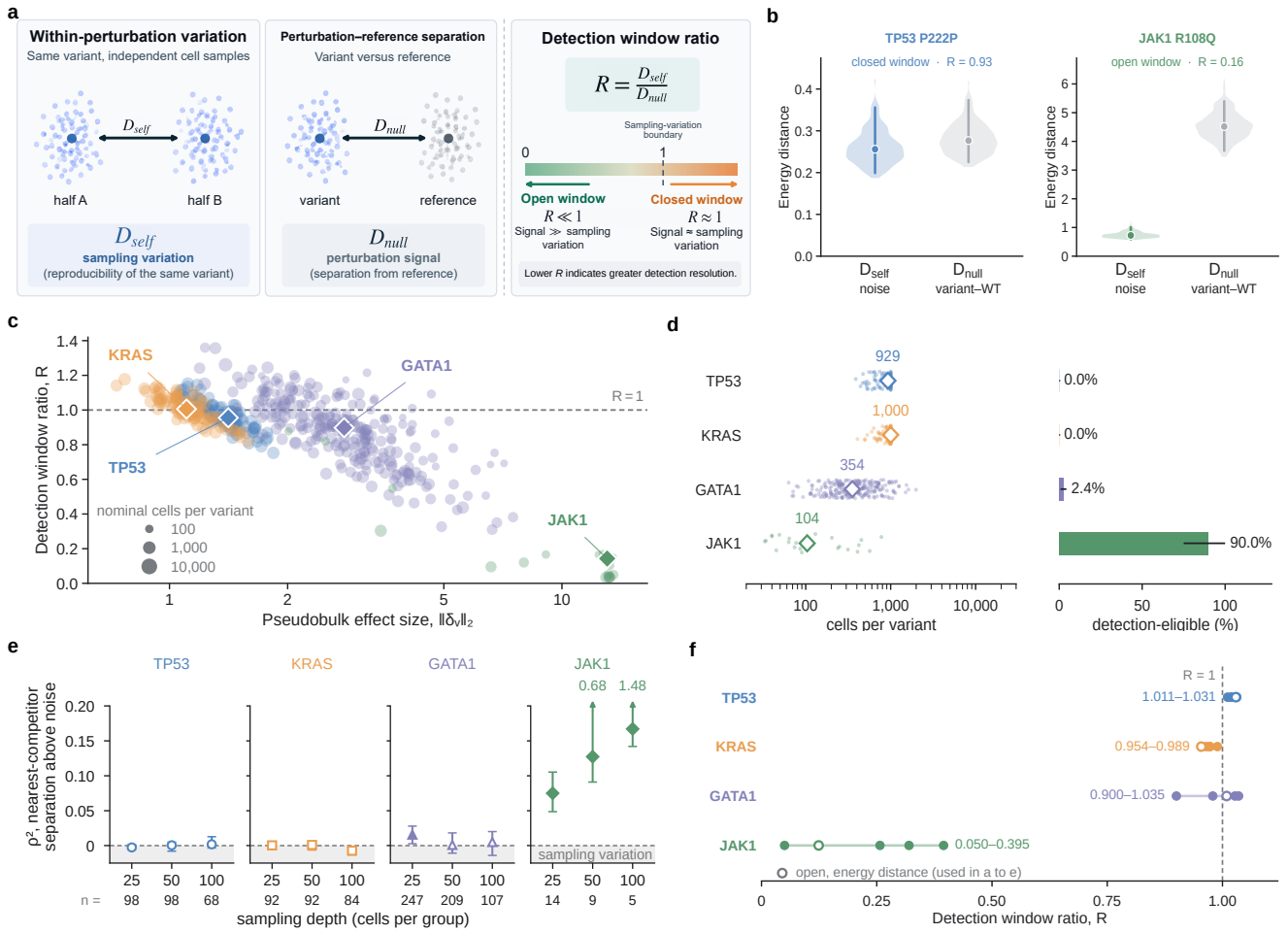

Figure 4: **Measurement resolvability depends on perturbation signal relative to within-condition variation.** **a**, Detection-window ratio $R = D_{\text{self}}/D_{\text{null}}$, comparing within-perturbation split-half variation with perturbation-to-reference separation. $R = 1$ indicates equal variation and signal; smaller values indicate greater separation. Detection eligibility uses the prespecified $S > W$ criterion (Methods). **b**, Representative TP53 P222P and JAK1 R108Q windows. Distributions show distances across 50 split-half draws; y-axis scales differ. **c**, Detection-window ratio against pseudobulk response magnitude $\|\delta_v\|_2$. Marker area, nominal cell count; diamonds, dataset medians. **d**, Available cells per variant and detection-eligible fractions at native depth. Diamonds, median cell counts; bars and whiskers, fractions and 95% Wilson confidence intervals. Denominators are variants evaluable at native depth (JAK1, 20 of 26). Native depth is each variant's deepest supported evaluated rung, capped at 150 cells per half. **e**, Cross-fitted nearest-competitor separation relative to within-condition sampling variation at 25, 50 and 100 cells per group. Points, medians across variants; whiskers, 95% bootstrap confidence intervals; filled symbols, intervals entirely above zero. Zero is the nominal no-signal reference under the exchangeable, common-variance sampling-error conditions in Methods; heteroskedasticity can shift the centring. Lower numbers, evaluable variants; upper JAK1 numbers, confidence limits beyond the axis. Depth-dependent cell requirements change cohort membership, so these are not fixed-cohort depth trajectories. **f**, Detection-window ratios under five distance families: energy distance, maximum mean discrepancy, sliced Wasserstein, cosine and Euclidean distance. Rows, datasets; printed intervals, ranges across families; open symbols, energy distance used in **a**–**d** (Supplementary Note 1).

sented. On the three resources evaluated under both preprocessing pipelines, changing from the count-based representation used for the public panel to the log-normalized, highly-variable-gene representation used in the depth analysis shifted the detection fraction by 0.097–0.195 (Supplementary Fig. 2; Supplementary Note 5). Measurement resolution is therefore a property of a resource under a specified representation and criterion, not of the resource alone.

Detection and identification separated even more strongly when both gates were evaluated in the six prespecified external screens at 50 cells per group (Fig. 5f; Supplementary Table 5). Between 26% and 90% of perturbations were separable from their reference, whereas no more than 43% were separable from their nearest competitor, and none of the screens exceeded the prespecified 0.50 operating point for identification. Candidate-pool size and composition differ across screens, so identification fractions are read within each screen. Identification is also depth-dependent: on a fixed cohort, increasing one GSE306429 configuration from 50 to 200 cells per group raised its identification fraction from 0.085 to

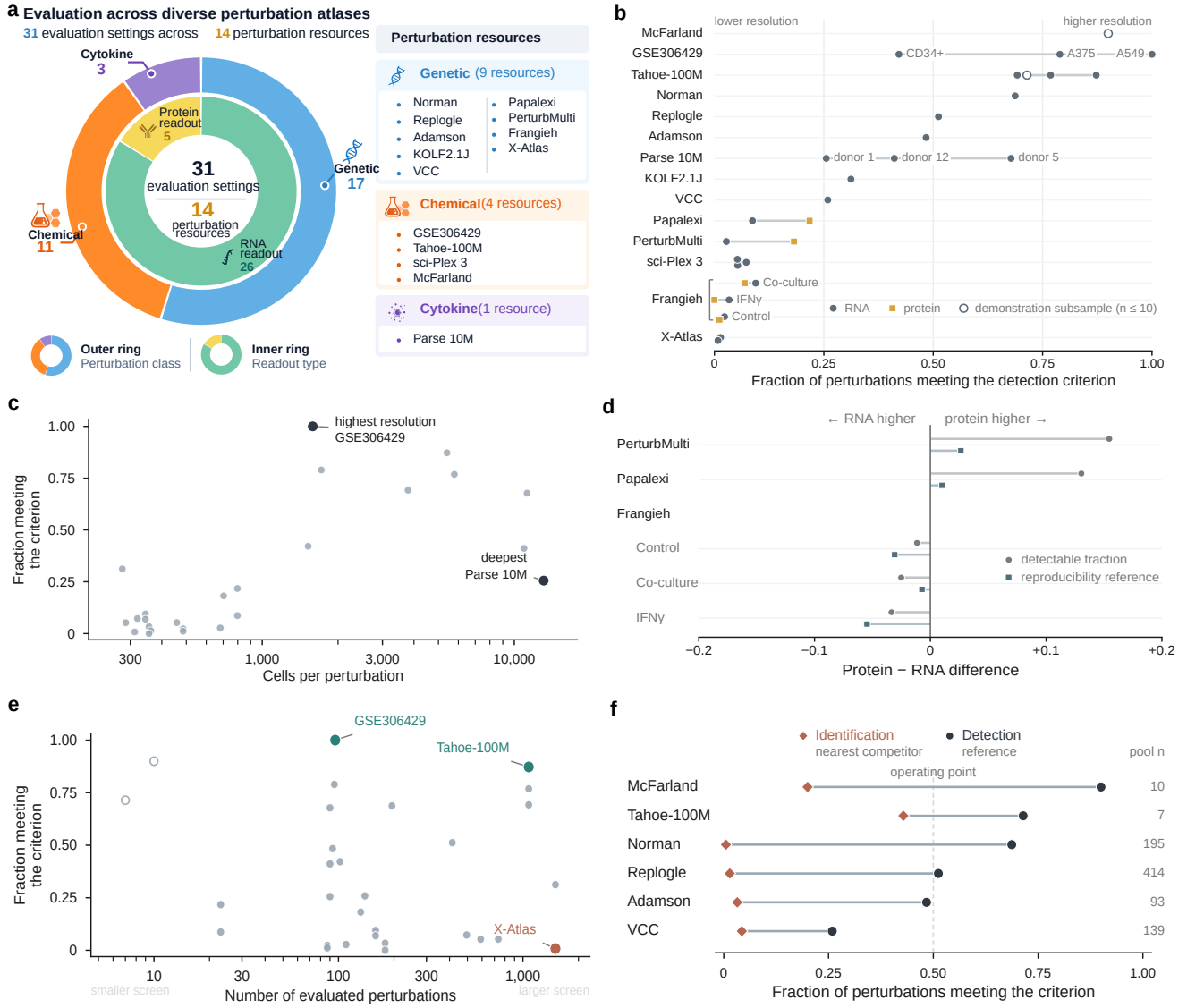

Figure 5: **Public perturbation screens span broad measurement-resolution regimes.** **a**, Analysis scope: 31 configurations nested within 14 resources, spanning genetic, chemical and cytokine perturbations and RNA and surface-protein readouts. **b**, Detection fractions grouped by resource. Rows are ordered by median; horizontal rules span within-row ranges. Frangieh 2021 contexts are shown separately. RNA fractions use 50 principal components of unnormalized counts and are specific to that representation (Methods). **c**, Cells per perturbation against detection fraction for the 25 configurations with recorded cell counts. **d**, Protein-minus-RNA differences in detection fraction and split-half reproducibility reference for five matched context–resource pairs from three resources. These are exploratory comparisons; the PerturbMulti reproducibility comparison uses different candidate pools and is descriptive. **e**, Evaluated perturbation count against detection fraction for all 31 configurations; logarithmic count axis. Coloured symbols identify the highest-resolution configuration and the highest- and lowest-resolution configurations among the ten largest. Open symbols, demonstration subsamples. **f**, Detection and nearest-competitor identification in six prespecified screens at 50 cells per group. Circles, fractions separable from reference; diamonds, fractions separable from the nearest competitor; connectors, within-screen differences; dashed line, 0.50 operating point. Rows are ordered by detection. Candidate pools differ across screens, so identification fractions are interpreted within rows; pool sizes appear at right. McFarland and Tahoe-100M are demonstration subsamples of at most ten perturbations.

0.557. The key distinction is therefore which biological comparison a measurement supports, perturbation versus reference or perturbation versus nearby competitors. Across public resources, measurement resolution was not ordered by cell number, screen size, perturbation class or readout modality, and detection of a perturbation relative to its reference did not imply resolution among nearby competitors. We next asked whether these quantities can be estimated early enough to guide benchmark design.

Measurement resolution informs benchmark design

Benchmark design involves three distinct decisions: which perturbations are likely to be detectably measured, which competitors define the identification problem, and how many perturbations are needed to distinguish models (Fig. 6a). Sampling depth influences measurement resolution but replaces neither the biological distinction being scored nor the benchmark size required for model comparison. We examined these components separately.

Repeated pilot–evaluation partitions provided an aggregate assessment of detection triage. For each of ten random partitions, the perturbed cells used for the pilot score were disjoint from those used for the evaluation quantities. The 25- or 50-cell pilot scores were then averaged across partitions, the evaluation quantities were separately aggregated across the same ten partitions to define detection eligibility, and one AUROC was calculated per dataset (Fig. 6b). Cells could reappear in different roles across repetitions, and the independently sampled reference-cell sets were not constrained to be disjoint. Across Adamson, Norman and Replogle, the repeated-pilot aggregates from a learned effect-size score and a training-free signal-to-variation rule were associated with the aggregate detection label, with AUROCs of 0.85–0.99 and 0.84–0.98, respectively. These values characterize the repeated-pilot aggregate and are not estimates of performance from a single 25- or 50-cell pilot. The analysis concerns detection triage; sibling-level identification must be measured directly.

Candidate-set size and candidate composition had different effects on attainable identification (Fig. 6c). Uniformly subsampling same-gene competitors left the mean split-half PDS unchanged from $K = 5$ to the full candidate pool, as expected for a percentile-ranked score. Restricting evaluation to the nearest competitors sharply reduced attainable identification: JAK1 fell from 0.807 over the full pool to 0.445 against its four nearest competitors, and GATA1 from 0.578 to 0.495. Controlled response geometries confirmed this dependence on local geometry. Across 1,440 configurations in which mean signal was controlled while response geometry varied, together with 12 measured gene-by-depth conditions, the split-half reproducibility reference tracked cross-fitted nearest-competitor separation more closely than mean pairwise separation (Spearman 0.961 versus 0.767; Fig. 6d). Local competitor separation, rather than average perturbation separation or candidate count, is therefore the more informative measure of attainable identification.

Pilot measurements and local competitor geometry can also inform sampling-depth projections (Fig. 6e). A calibration linking nearest-competitor separation to the split-half reproducibility reference returned finite projections for JAK1 and GATA1 but not for TP53 or KRAS. Because these estimates extrapolate beyond the candidate-set sizes used for calibration, they are illustrative design estimates rather than validated depth prescriptions (Supplementary Note 5).

Sampling depth remained consequential when perturbation identity was held fixed. Subsampling Replogle, Norman and Adamson to 50 cells per half reduced the detection-eligible fraction by 2.8, 8.5 and 24.0 percentage points, respectively (Fig. 6f). A common cell budget therefore does not produce a common measurement resolution, because screens differ in how much additional depth improves perturbation-to-reference separation.

Benchmark size formed a separate design axis (Fig. 6g). Among predictor pairs whose design ordering was realized in the full cohort, the smallest tested benchmark achieving a 0.9 ordering-recovery probability ranged from ≤ 5 perturbations for well-resolved measurements with large model differences to 160 perturbations when measurement reproducibility was near chance and the realized PDS difference was below 0.01. The number of perturbations required for model comparison therefore depends jointly on measurement reproducibility and the performance difference to be resolved.

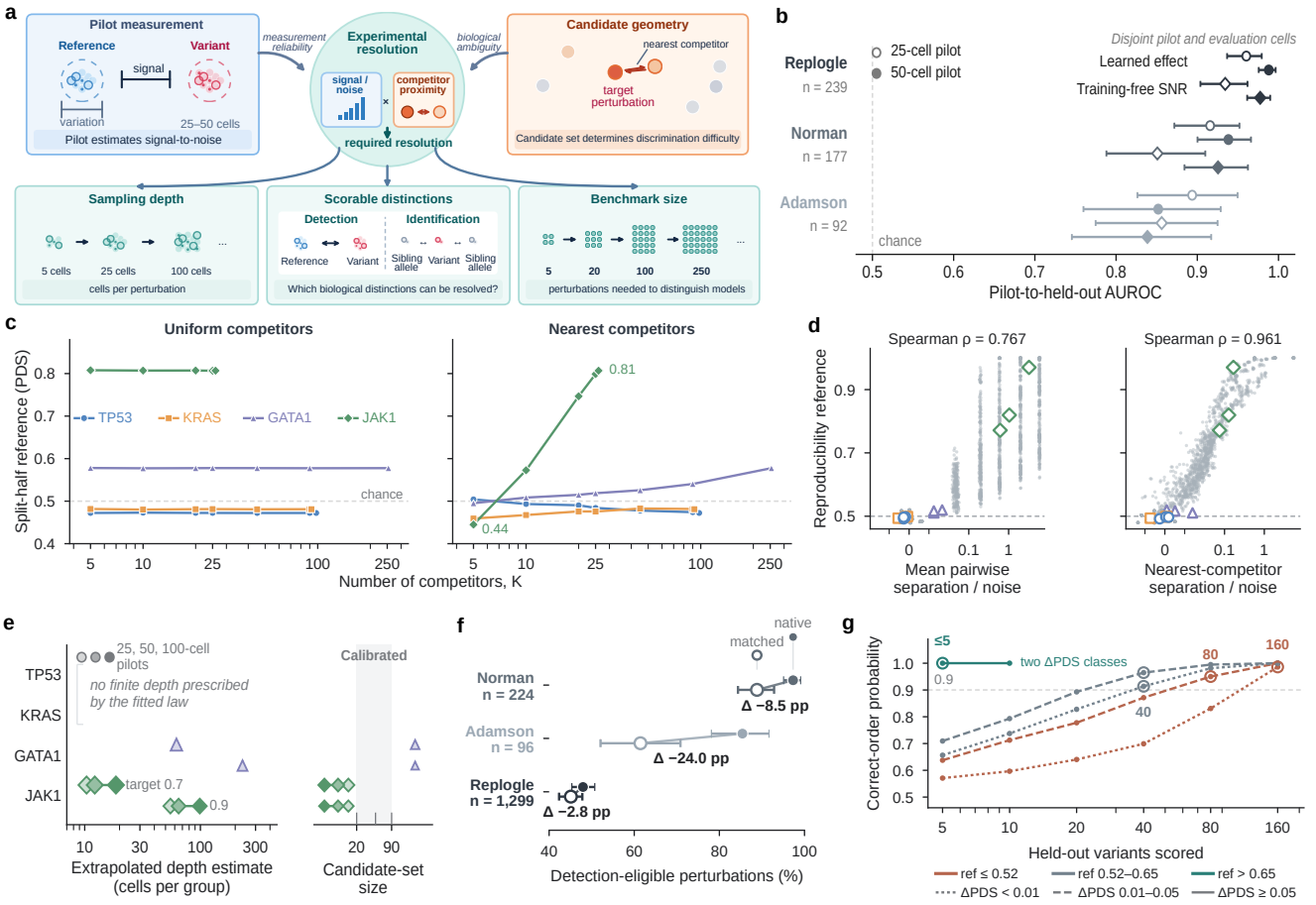

Figure 6: **Measurement resolution informs benchmark design.** **a**, Design framework linking pilot measurements, competitor geometry, sampling depth and benchmark size. **b**, Detection association based on ten-repeat aggregates of 25- or 50-cell pilot scores and evaluation labels. Perturbed pilot and evaluation cells are disjoint within each partition but can recur across partitions; reference-cell draws are not mutually exclusive. AUROC is calculated after aggregation, rather than within each partition. Symbols distinguish learned effect-size and training-free signal-to-variation scores. Whiskers, 95% bootstrap confidence intervals over perturbations; dashed line, chance AUROC; n, evaluated perturbations. **c**, Split-half reproducibility reference under uniformly sampled competitors (left) and nearest-competitor restriction (right), varying candidate-set size K, including the target. **d**, Reproducibility reference against mean pairwise separation (left) and cross-fitted nearest-competitor separation (right). Grey points, 1,440 controlled response geometries; open symbols, 12 measured gene-by-depth conditions. Reported coefficients are Spearman correlations. **e**, Model-based depth projections from pilot nearest-competitor separation, with candidate-set size selecting the calibration stratum. Eight of 24 gene–target–pilot combinations return finite projections. The companion panel compares their candidate-set sizes with the calibration range (20–90). All projections extrapolate beyond that range and are not measured depth requirements. **f**, Native versus depth-matched detection at 50 cells per half in Norman, Adamson and Replogle. Whiskers, 95% paired-bootstrap confidence intervals; n, perturbations evaluated at both depths. **g**, Probability of recovering the controlled predictors' design ordering as benchmark size increases, grouped by measurement reproducibility and realized PDS difference. Recovery probabilities are summarized over gene–depth–predictor-pair conditions with a positive realized full-cohort PDS difference; deterministic full-cohort evaluations are excluded from curves (Methods). Labels mark the smallest tested size reaching 0.9 recovery; ≤ 5 is an upper bound. Contributing conditions can change with benchmark size.

Discussion

Our results show that fine-grained perturbation prediction can only be interpreted when the experiment itself can reproducibly distinguish the biological states being compared. We first observed this problem in allele-specific prediction. Models often recovered the overall transcriptional response shared by variants of the same gene, but this did not mean that they could distinguish one allele from another. In our allele-resolved benchmarks, substantial recovery of transcriptional direction coexisted with near-chance identification of the correct allele among all same-gene candidates. Importantly, weak allele-level performance had different explanations in different datasets. In TP53 and KRAS, the measurements themselves did not reproducibly separate the alleles at the evaluated resolution. In JAK1, by contrast, reproducible allele-specific differences were present, but the evaluated models recovered them only partially. Thus,

a weak model score can mean either that the experiment does not resolve the requested distinction or that the distinction is measurable but has not yet been learned by the model. This interpretation also depends on the comparison being made: models could distinguish some reproducibly separated sibling alleles even when they could not identify the correct allele across the full candidate pool.

The same issue extends beyond coding variants. A perturbation benchmark can ask at least two different measurement questions: whether a perturbation can be distinguished from its reference, and whether it can be distinguished from other, similar perturbations. These questions depend on different properties of the data. Detection depends mainly on the size of the perturbation response relative to within-condition variation, whereas identification also depends on how close the most similar competing perturbations are. Across 31 configurations from 14 public perturbation resources, we observed wide variation in both forms of resolution across genetic, chemical and cytokine perturbations and across RNA and surface-protein readouts. More cells or more perturbations did not consistently imply better resolution, and a perturbation that was clearly different from its reference was not necessarily distinguishable from its nearest competitors. The apparent resolution also changed with the expression representation and the set of perturbations used for comparison. A perturbation dataset should therefore not be assigned a single, context-free notion of resolution; the relevant resolution depends on the biological question, the readout and the comparison being evaluated.

These findings also suggest a practical way to design perturbation benchmarks. More cells improve the precision with which each perturbation response is measured, but they do not by themselves resolve ambiguity between biologically similar perturbations. Likewise, the number of perturbations in a benchmark does not determine how difficult identification will be; this depends more directly on which competing perturbations are included. Benchmark size addresses a separate question: how many perturbations are needed to distinguish the performance of two models reliably. These design choices should therefore be considered separately. Repeated pilot–evaluation analyses can provide an initial indication of which perturbations are likely to be detectably measured, whereas identification requires explicit assessment of nearby competitors. Benchmark size should then be chosen according to the model difference that the study aims to resolve. Our pilot analysis is an aggregate over repeated re-partitions rather than a prospective evaluation of one fixed pilot, and should be interpreted accordingly. For model development, the most informative settings are those in which the measurement reproducibly supports a distinction but current models still fail to recover it, as illustrated by the JAK1 allele-level regime. By contrast, poor model performance on an experimentally unresolved comparison should not be interpreted as evidence of insufficient model capacity alone.

Several limitations define the scope of these conclusions. The allele-resolved analysis includes four dataset–assay–context configurations in which gene identity, experimental technology and cellular context vary together. The observed regimes therefore should not be interpreted as intrinsic properties of TP53, KRAS, GATA1 or JAK1. The broader perturbation panel is a selected set of eligible configurations rather than an exhaustive survey of public perturbation datasets, and several configurations come from the same resource. Our split-half analyses measure reproducibility across cell samples from the same experiment, not across independent experiments. The split-half reference is also an empirical reference rather than a hard upper bound, because models that denoise measurements or borrow information across conditions could in principle exceed it. Measurement resolution is specific to the readout and representation used for evaluation, and biological effects that are weak in transcriptional space may be clearer in other molecular or phenotypic readouts. Finally, the sampling-depth projections are model-based extrapolations rather than validated prescriptions for experimental design. Within these limits, the central conclusion is straightforward: before interpreting a model score at fine biological resolution, one should first establish that the experiment itself can reproducibly resolve that distinction. Reporting measurement resolution alongside model performance separates what the experiment cannot distinguish from what the model has not yet learned.

Methods

PertResolve-Bench: datasets and preprocessing

PertResolve-Bench comprises two complementary components: an allele-resolved benchmark for evaluating fine-grained response prediction and a broader perturbation panel for characterizing measurement

resolution across perturbation classes, cellular contexts and readout modalities. The allele-resolved component contains genotype-linked single-cell expression datasets with multiple coding variants of the same gene sufficient for within-gene evaluation, whereas the broader panel is defined by the eligibility criteria described below.

Allele-resolved benchmark. All analyses used publicly available single-cell perturbation datasets; no new experiments were performed. The primary unit of prediction in the allele-resolved benchmark was the protein-coding variant condition. Here, *allele* denotes an experimentally defined protein-coding variant condition rather than a phased genomic haplotype. The allele-resolved component included datasets with genotype-linked single-cell expression, multiple coding variants of the same gene sufficient for within-gene evaluation, and compatibility with a common held-out-variant protocol. TP53, KRAS, GATA1 and JAK1 met these criteria (Supplementary Table 6).

Data sources and eligibility. The allele-resolved component of PertResolve-Bench integrates single-cell perturbation data for four protein-coding genes across three experimental technologies. TP53 (98 variants) and KRAS (92 variants) were obtained from a Perturb-seq screen of coding variants in A549 lung adenocarcinoma cells (Gene Expression Omnibus GSE161824)[17]. GATA1 (254 variants) was obtained from a base-editing screen in human haematopoietic stem and progenitor cells[18]; we used the processed expression matrix distributed with PerturbNet[27] (Hugging Face repository `cyclopeta/PerturbNet_reproduce`). JAK1 (26 variants) was obtained from an scSNV-seq experiment in HT-29 colorectal cancer cells stimulated with interferon-gamma (European Nucleotide Archive PRJEB48915; processed objects from Zenodo record 10418435)[19]. The four datasets comprise 470 protein-coding variant conditions and 321,043 cells, including the dataset-specific wild-type or control cells retained after preprocessing.

Expression preprocessing. For TP53 and KRAS, processed count matrices, cell barcodes and variant-to-cell assignment tables were loaded from GEO, yielding a shared 939-gene expression space for the two A549 datasets. For GATA1, we used the PerturbNet-distributed AnnData object containing log-normalized expression over 2,477 highly variable genes. For JAK1, the released SingleCellExperiment object was loaded in R (Bioconductor). Cells were restricted to homozygous genotyped calls with unambiguous coding consequences (missense, splice-site, stop-gained or synonymous). The 2,000 most highly variable genes were selected using `scran::modelGeneVar` followed by `getTopHVGs`, and log-normalized counts were exported for downstream analysis.

Each dataset was analysed in its native expression space; expression genes were therefore not required to match across datasets. Within each gene and generalization split, expression features were standardized using only cells assigned to training variants and to the corresponding reference population, excluding all held-out-variant cells. The fitted transformation was then applied unchanged to training and held-out cells. The dataset-specific reference population, wild-type or control cells, served as the common reference for constructing perturbation-response profiles, and is denoted the reference throughout.

Variant representations. Each variant v was represented by molecular features rather than by a variant identifier, allowing prediction for variants not used during model fitting. The primary representation was a six-dimensional biophysical vector

$$\theta_v = (\Delta H_v,\ \Delta V_v,\ \Delta Q_v,\ f_v,\ s_v,\ h_v)\,,$$

where the first three components describe mutant-minus-wild-type differences in Kyte–Doolittle hydrophobicity, side-chain volume and formal charge at pH 7.0. For a residue property $p(\cdot)$ and mutant and wild-type amino acids a_v^{mut} and a_v^{wt},

$$\Delta p_v = p(a_v^{\mathrm{mut}}) - p(a_v^{\mathrm{wt}}).$$

These differences were standardized before model fitting within each generalization setting, using the mean and standard deviation of that setting's training variants alone, so held-out variants do not influence the feature scale. The three position annotations below are on a common 0–1 scale and were not rescaled.

The remaining components encode position-level prior annotations, assigned according to the pre-specified gene-specific rules in Supplementary Table 1. The fold-core term f_v was set to 1.0 for residues within an annotated structural core and 0.0 otherwise; the functional-switch term s_v to 1.0 for catalytic or metal-coordinating residues; and the hotspot or pathogenic marker h_v to 1.0 from external annotations alone, without reference to the single-cell response, using COSMIC and IARC missense hotspot codons, ClinVar pathogenic annotations and gene-specific functional-module membership.

The physicochemical difference terms are naturally defined for single amino-acid substitutions. Conditions without a single wild-type-to-mutant substitution were assigned zero for these three terms while retaining their position-based annotations; conditions naming multiple substitutions used the mean of the corresponding site-level differences. Because this representation is not injective, distinct variants can share the same feature vector and therefore receive identical predictions from a feature-based model. Such ties are handled explicitly by the tie-aware discrimination score defined below; sensitivity to alternative molecular representations is reported in Supplementary Note 2.

Protein-language-model embeddings. As an alternative molecular representation, per-variant embeddings were extracted from ESM-1v (`esm1v_t33_650M_UR90S_1`)[28]. For the mutant sequence of variant v, let $\mathbf{H}_v \in \mathbb{R}^{L_v \times 1280}$ denote the final-layer hidden states over L_v residues. The sequence embedding was obtained by mean pooling across residue positions,

$$e_v = \frac{1}{L_v} \sum_{\ell=1}^{L_v} \mathbf{H}_v[\ell, :] \in \mathbb{R}^{1280},$$

and was used directly without dimensionality reduction. Synonymous variants were represented by the corresponding wild-type protein sequence. Three JAK1 conditions for which a well-formed mutant protein sequence could not be constructed were excluded from ESM-based analyses, leaving 467 of the 470 benchmark variants embedded. Mutant sequences longer than 1,000 residues were reduced to a 1,000-residue window starting 500 residues before the first listed substitution, clamped to lie inside the sequence, and every sequence was then truncated to its first 1,022 residues to fit the model's 1,024-token limit; wild-type reference sequences were embedded without windowing. Only JAK1 (1,154 residues) exceeds either limit, so its conditions are embedded over windows that differ between conditions and JAK1 ESM-1v features carry a condition-dependent crop. Matched-window mutation representations were evaluated separately with ESM2 (Supplementary Note 2).

Allele-specific prediction

Prediction target. For variant v, let $\mathcal{X}_v$ and $\mathcal{X}_{\text{ref}}$ denote the corresponding variant and reference cells, and let $x_i \in \mathbb{R}^G$ denote the standardized expression profile of cell i over G genes. The measured perturbation response was defined as the pseudobulk difference

$$\delta_v = \frac{1}{|\mathcal{X}_v|} \sum_{i \in \mathcal{X}_v} x_i - \frac{1}{|\mathcal{X}_{\text{ref}}|} \sum_{j \in \mathcal{X}_{\text{ref}}} x_j.$$

For each profile, up to 300 cells were sampled without replacement from the variant and reference groups; variants represented by fewer than five cells were excluded. Each predictor maps molecular features of a held-out variant to an estimated response $\hat{\delta}_v$. The prediction target throughout is therefore the population-average transcriptional response relative to the reference. Single-cell distributional analyses described below are used only to quantify measurement resolution and do not define an alternative prediction target.

Generalization settings. We evaluated four primary forms of within-gene generalization. In each setting, models were fitted only to the training variants of a gene and evaluated on its held-out variants.

1. **Random holdout.** We used the predefined per-dataset partitions distributed with the allele-resolved component of PertResolve-Bench, with GATA1 following the PerturbNet partition. The held-out sets contained 16 of 98 TP53 variants, 8 of 92 KRAS variants, 52 of 254 GATA1 variants and 9 of 26 JAK1 variants.
2. **Positional extrapolation.** Variants located in the C-terminal half of each protein were held out, testing generalization to protein positions not represented by the corresponding training partition.

3. **Mechanistic extrapolation.** Variants at externally annotated hotspot or functional-switch residues were held out, testing generalization to functionally prioritized positions.
4. **Low-depth evaluation.** Variants represented by fewer than 200 cells were held out to evaluate prediction in the shallow-sampling regime. This setting applies to GATA1 (45 variants) and JAK1 (17 variants); all TP53 and KRAS variants exceed this depth.

A fifth *compatibility* setting was retained only for comparison with the PerturbNet evaluation protocol and was not treated as an independent generalization axis. It uses the same 52 held-out GATA1 variants as the random setting together with predefined TP53 and KRAS holdouts. All variants evaluated under this setting are contained within the random-holdout evaluation.

A sixth *cross-gene transfer* setting is distributed with the benchmark and holds out GATA1 and JAK1 as whole genes, leaving TP53 and KRAS as the training genes (Fig. 1g). It is assigned no prediction scores, because the four datasets are measured in different cell types over different gene panels and therefore share no common expression space in which a prediction trained on one gene could be scored on another.

Feature-based predictors. The primary model grid comprised 18 feature–model combinations obtained by crossing six regression heads with three molecular feature spaces. Each model learns

$$g_\phi : \mathbb{R}^p \to \mathbb{R}^G$$

from a p-dimensional variant representation to the G-dimensional perturbation response and was fitted on training variants by minimizing

$$\min_\phi \sum_{v \in \text{train}} \left\| g_\phi(\text{feat}_v) - \delta_v \right\|_2^2,$$

with model-specific regularization where applicable.

The six regression heads were ridge regression, lasso regression, random forest, gradient-boosted trees, k-nearest-neighbours regression and a multilayer perceptron. Hyperparameters were fixed at the values listed in Supplementary Table 2 and were not tuned by gene, by generalization setting or by split; no hyperparameter search was performed at any point. Linear, neighbour and neural-network models predicted the expression profile directly. For expression spaces exceeding 200 genes, random-forest and gradient-boosted models were fitted to a principal-component representation of the training responses and transformed back to gene space after prediction. Principal components were fitted using training responses only. All primary model fits used fixed random seeds.

The three primary feature spaces were the six-dimensional biophysical representation θ, the 1,280-dimensional ESM-1v embedding, and their concatenation. Additional representation controls included ESM2-based mutant embeddings, mutant-minus-wild-type ESM2 representations at global, mutated-residue and local-window scales, and AlphaMissense pathogenicity scores. These controls used the same train–test partitions and evaluation metrics; representation-specific construction details are given in Supplementary Note 2.

Reference predictors. The WT-null reference predicts

$$\hat{\delta}_v = 0$$

for every variant and represents prediction without perturbation information. The Gene-mean reference predicts the mean response of the training variants of the held-out variant's gene,

$$\hat{\delta}_v = \frac{1}{|\text{train}(g_v)|} \sum_{w \in \text{train}(g_v)} \delta_w.$$

Because this reference uses gene identity but no allele-specific molecular information, it was used to quantify the response recoverable from gene-shared structure rather than as a competing allele-level predictor.

Released-model evaluation. Released perturbation-prediction frameworks were evaluated through the same pseudobulk targets, held-out variants and PDS definition as the feature-based predictors (Supplementary Table 3). Biolord[29] and CellFlow[7] were run using their released implementations; variant ESM-1v embeddings were used as the CellFlow condition. scGen[4] natively uses categorical perturbation identities rather than continuous variant features, so we evaluated an allele-conditioned implementation of its latent-shift framework in which the latent shift is a fitted continuous function of molecular variant attributes and is therefore defined for a variant never seen in training. A corresponding scVIDR[30] implementation used the same fitted latent shift but additionally scaled it by a per-variant variable. That variable was an external AlphaMissense pathogenicity score for TP53 and KRAS[31], but it was derived from measured response cells for GATA1. The scVIDR result is therefore retained only as a response-informed sensitivity control and is excluded from inference about response-blind held-out-variant prediction. These entries evaluate adaptations rather than the released gene-level methods. Architectures, conditioning interfaces, training configurations, scaling-variable provenance and coverage are given in Supplementary Note 3.

scGPT[8], GEARS[6] and STATE[32] condition on perturbed-gene identity and consequently produce no within-gene allele distinction. Their allele-level scores were not evaluated because their native conditioning interfaces do not encode variant-level differences. PerturbNet[27] was evaluated using ESM-1v conditioning across all five scored within-gene settings: random holdout, positional extrapolation, mechanistic extrapolation, low-depth evaluation and compatibility. These runs yielded 492 gene–setting–variant evaluations and 303 distinct gene–variant units after overlap across settings was removed. Predictions and candidate responses were compared within PerturbNet's 50-dimensional principal-component output representation. Scores were first averaged over 15 cell-subsample seeds within a gene–setting–variant unit, then across settings for each gene–variant, before aggregation and bootstrap over the 303 distinct variants. The matching restricted-label permutation procedure is specified in Statistical analysis. Model coverage and conditioning interfaces are summarized in Supplementary Table 3.

Prediction evaluation

PertResolve-Eval assesses complementary properties of perturbation prediction rather than reducing performance to a single similarity score. For the allele-resolved benchmark, the primary metric, perturbation discrimination score (PDS), measures whether a prediction is specific to the correct allele. Direction-recovery metrics quantify agreement with the overall transcriptional response, and differential-expression metrics assess agreement in the genes and effect sizes underlying that response. Unless otherwise stated, metrics are calculated for each held-out variant and then aggregated over distinct held-out variants. The primary PDS estimate first averages 15 random cell subsamples within each gene–setting–variant unit and then averages repeated evaluations of the same gene–variant across settings; secondary response-recovery metrics use a fixed pseudobulk draw.

Perturbation discrimination score. For a held-out variant v of gene g, the predicted response $\hat{\delta}_v$ was compared with the measured responses of all variants of the same gene. The candidate set

$$\mathcal{C}_g = \mathrm{train}(g) \cup \mathrm{test}(g)$$

therefore contains both training and held-out variants that pass the five-cell filter. Let $n = |\mathcal{C}_g|$, let

$$\ell_v = \left|\left\{w \in \mathcal{C}_g : d(\hat{\delta}_v, \delta_w) < d(\hat{\delta}_v, \delta_v)\right\}\right|$$

be the number of candidates strictly closer to the prediction than its true target, and let

$$e_v = \left|\left\{w \in \mathcal{C}_g : d(\hat{\delta}_v, \delta_w) = d(\hat{\delta}_v, \delta_v)\right\}\right|$$

be the number tied with the true target, including v itself. We define the tie-aware score

$$\mathrm{PDS}(v) = 1 - \frac{\ell_v + \frac{1}{2}(e_v - 1)}{n - 1}.$$

The nearest unique match scores 1, whereas random matching has expected score 0.5; ties receive their mid-rank. Only held-out variants are scored, but each prediction competes against the full same-gene

candidate set. Predictions yielding no finite ranking, including zero-norm or non-finite vectors, and candidate sets containing fewer than two variants are assigned a score of 0.5.

Top-k shortlisting. The same ranking supports a tail statistic. Using ℓ_v and e_v above, the probability that the true allele falls in the top k of its same-gene ranking when ties are broken at random is

$$\mathrm{Hit}_k(v) = \frac{\mathrm{clip}(k - \ell_v,\, 0,\, e_v)}{e_v},$$

evaluated at $k \in \{1, 3, 5, 10\}$, the same shortlist lengths used for the measurement reference. The chance rate is computed per query from that query's own candidate-pool size, $\min(k/N_v, 1)$, because pool size differs by gene (98, 92, 254 and 26). A prediction with no finite distance to its target has no ranking and is assigned its query's chance expectation, matching the convention by which such predictions receive PDS 0.5; scoring them as misses instead is reported as a sensitivity analysis (Supplementary Note 2). Seed, split and query aggregation and the bootstrap follow the PDS estimator exactly, and one bootstrap draw scores all four k, so the recovery curve retains its covariance. Inference is the paired bootstrap of the observed rate minus that method's own chance rate over the same resampled queries.

Cosine distance was used for the primary analysis,

$$d_{\cos}(a, b) = 1 - \frac{a^\top b}{\|a\|_2 \|b\|_2}.$$

Robustness analyses additionally used

$$d_1(a, b) = \|a - b\|_1, \qquad d_2(a, b) = \|a - b\|_2.$$

Under this tie-aware mid-rank definition, uniform assignment of the target label over the complete candidate pool has expected PDS exactly 0.5, including for finite candidate sets and in the presence of ties. We additionally used task-matched permutations that reassigned labels only among held-out variants within a gene and evaluation setting while retaining the full training-plus-held-out candidate columns. This conditional reference can differ from 0.5 because the permuted label set is a restricted subset of the ranking pool, not because of candidate-set size, ties or output dimensionality (Statistical analysis).

Direction recovery and residualization. Agreement with the measured transcriptional response was quantified using Pearson correlation,

$$\text{Pearson-}\delta = \frac{1}{n_{\text{test}}} \sum_{v \in T} \mathrm{corr}\left(\hat{\delta}_v, \delta_v\right),$$

where T denotes the held-out variants.

The descriptive residual Pearson analysis in Fig. 2e (left) centred measured and predicted profiles separately across the same evaluated variants of each gene. Each distinct variant contributed its first available prediction in the fixed evaluation-setting order. Writing T_g for that cohort, the profiles entering the correlation were $\delta_v - |T_g|^{-1} \sum_{w \in T_g} \delta_w$ and $\hat{\delta}_v - |T_g|^{-1} \sum_{w \in T_g} \hat{\delta}_w$. Correlations were averaged across variants with non-constant residual profiles. This analysis describes residual agreement within the evaluated cohort; it does not estimate a residual score using a training-only centre.

For residual PDS in Fig. 2e (right), we instead subtracted the same training-only gene mean from both measured and predicted profiles,

$$\bar{\delta}_g = \frac{1}{|\mathrm{train}(g)|} \sum_{w \in \mathrm{train}(g)} \delta_w,$$

giving

$$r_v = \delta_v - \bar{\delta}_g, \qquad \hat{r}_v = \hat{\delta}_v - \bar{\delta}_g.$$

PDS was then recomputed on these residual profiles. Uniform matching over the complete residual candidate pool still has analytic expectation 0.5. For inference, residual-axis PDS was compared with its own task-matched within-gene permutation distribution, generated by permuting the held-out labels

while retaining the full training-plus-held-out residual candidate pool; this restricted-label reference need not equal 0.5 (Statistical analysis).

Measured-response decomposition. The measurement-only decomposition in Fig. 2d used all eligible benchmark variants of each gene, independently of model-training assignments. Variants required at least five cells; the eligible sets $\mathcal{V}_g$ contained 98 TP53, 92 KRAS, 254 GATA1 and 26 JAK1 variants. For each of $B = 15$ random cell partitions, variant and reference cells were divided into disjoint halves, each capped at 300 cells. Let $\delta_v^{(b)}$ denote the variant-minus-reference mean response from the second half of partition b. The gene-shared mean, mean squared response norm and mean squared residual norm were

$$\mu_g^{(b)} = \frac{1}{|\mathcal{V}_g|} \sum_{v \in \mathcal{V}_g} \delta_v^{(b)},$$
$$T_g^{(b)} = \frac{1}{|\mathcal{V}_g|} \sum_{v \in \mathcal{V}_g} \|\delta_v^{(b)}\|_2^2,$$
$$A_g^{(b)} = \frac{1}{|\mathcal{V}_g|} \sum_{v \in \mathcal{V}_g} \|\delta_v^{(b)} - \mu_g^{(b)}\|_2^2.$$

The displayed allele-residual percentage was

$$100 f_g, \qquad f_g = 1 - \frac{1}{B} \sum_{b=1}^{B} \frac{\|\mu_g^{(b)}\|_2^2}{T_g^{(b)} + \epsilon}, \qquad \epsilon = 10^{-12}. \tag{1}$$

Because residuals sum to zero within each partition, $T_g^{(b)} = \|\mu_g^{(b)}\|_2^2 + A_g^{(b)}$. Thus, apart from the numerical stabilizer, f_g is the mean of the partition-specific squared-residual fractions $A_g^{(b)} / T_g^{(b)}$, with equal weight per variant within each partition. It is not a mean of per-variant norm ratios or a ratio of totals pooled across partitions. The complementary gene-shared percentage is $100(1 - f_g)$. These descriptive fractions include cell-sampling variation and do not quantify the reproducible fraction of allele-specific signal.

Differential-expression fidelity. For each variant, expression genes were ranked using a Welch t-test comparing variant and reference cells, with at most 300 cells per group. Let $\mathcal{D}_v^{\text{true}}$ denote the $k = 50$ genes with the largest absolute t-statistics and $\mathcal{D}_v^{\text{pred}}$ the k genes with the largest absolute predicted responses. Their overlap was

$$\text{DE overlap}(v) = \frac{|\mathcal{D}_v^{\text{pred}} \cap \mathcal{D}_v^{\text{true}}|}{k}.$$

The measured effect size of gene g in variant v was the difference of group means on the same standardized expression scale used for the t-test,

$$\lambda_{v,g} = \frac{1}{|S_v|} \sum_{i \in S_v} x_{i,g} - \frac{1}{|S_{\text{ref}}|} \sum_{j \in S_{\text{ref}}} x_{j,g},$$

where S_v and S_{ref} are the variant and reference draws entering the test. $\lambda_{v,g}$ is a standardized mean difference in units of the gene's standard deviation, not a log-ratio or log fold change. We therefore call the metric *DE effect-size Spearman*. It is Spearman's rank correlation between $\lambda_{v,g}$ and $\hat{\delta}_{v,g}$ over $\mathcal{D}_v^{\text{true}}$, with ties receiving midranks. Direction agreement is the fraction of those genes for which $\text{sign}(\hat{\delta}_{v,g}) = \text{sign}(\lambda_{v,g})$. The cell draws entering the differential-expression statistics are made independently of those forming the pseudobulk response used for PDS and Pearson-δ, although both take the same reference population.

Statistical unit. The primary statistical unit for model evaluation was the distinct protein-coding variant. Because a variant can be held out under more than one evaluation setting, scores from repeated evaluations of the same variant were first averaged, leaving one value per distinct (gene, variant) pair. The primary model comparison contains 518 held-out evaluations corresponding to 308 distinct variants for predictors defined on every benchmark variant; predictors requiring a sequence embedding are

scored on 511 evaluations and 305 distinct variants, because three JAK1 conditions have no constructible mutant sequence. Compatibility evaluations are nested within the random-holdout setting and do not introduce additional held-out variants.

Measurement resolution

Perturbation detection. We first asked whether each perturbation produced a response that was reproducibly distinguishable from its reference at the available sampling depth. For two cell sets $X = \{x_i\}_{i=1}^{n}$ and $Y = \{y_j\}_{j=1}^{m}$, the energy distance was

$$E(X, Y) = \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|x_i - y_j\|_2 - \frac{1}{n^2}\sum_{i,i'}\|x_i - x_{i'}\|_2 - \frac{1}{m^2}\sum_{j,j'}\|y_j - y_{j'}\|_2.$$

For each perturbation with sufficient cells, cells were randomly divided into two matched halves A and B, and a size-matched sample C was drawn from the corresponding reference population. We calculated the within-perturbation split-half distance

$$D_{\text{self}} = E(A, B)$$

and the perturbation-to-reference distance

$$D_{\text{null}} = E(A, C).$$

The analysis was repeated over 50 random cell splits. The primary analysis used a 50-component principal-component representation at the deepest split-half depth supported by each perturbation. The principal-component basis was fitted once per gene for the allele-resolved datasets and once per dataset for the gene-perturbation screens, on all of that gene's or dataset's cells, perturbed and reference alike. The allele-resolved datasets entered in the native expression spaces described under Expression preprocessing above, not in a common gene set; the three gene-perturbation screens were library-size normalized to 10^4 counts per cell, log transformed and restricted to their 2,000 most highly variable genes before the basis was fitted. It was fitted before any cell split and was not refitted within a depth rung or a split. The basis carries no perturbation label and is shared by both halves of every split, so it fixes the coordinate system in which distances are measured without entering the comparison between D_{self} and D_{null}. The corresponding basis for the broader perturbation panel is described with that criterion below.

Detection eligibility was defined by comparing the separation from the reference with the sampling variation of the perturbation itself. Let

$$S = \text{med}(D_{\text{null}}) - \text{med}(D_{\text{self}})$$

and

$$W = Q_{97.5}(D_{\text{self}}) - Q_{2.5}(D_{\text{self}}).$$

A perturbation was classified as detection-eligible when

$$S > W.$$

This definition was applied identically to the four allele-resolved datasets and to the Replogle, Norman and Adamson gene-perturbation datasets obtained through scPerturb [12,13,33,34]. Detection was read from the energy-distance rows computed on the 50-component principal-component representation. At native depth, each perturbation was evaluated at the deepest rung of the split-half subsample ladder that it supported: 30, 50, 100 and 150 cells per half for the four allele-resolved datasets, and 25, 50, 100, 200, 400 and 800 for the gene-perturbation datasets. A perturbation entered a rung only when it supplied at least twice that rung's cells, which admitted 464 of the 470 allele-resolved variant conditions; the six excluded are JAK1 variants holding fewer than 60 cells. Allele-resolved analysis pools were capped at 300 cells per perturbation, so native depth denotes the deepest evaluated subsample rather than every available cell; 341 of the 464 perturbations were evaluated at the deepest rung of 150 cells per half.

Alternative distance functions, detection criteria, expression representations and reference variability were examined in Supplementary Note 1.

Sibling identification. Detection establishes separation from the reference but does not establish separation among sibling perturbations. We therefore quantified identification using pseudobulk response profiles constructed from disjoint cell samples. For variant v, let δ_v^A and δ_v^B denote response profiles constructed from disjoint cell halves at matched depth. Within-variant sampling variation was measured by

$$D_{\text{self}}(v) = d_{\cos}\left(\delta_v^A, \delta_v^B\right).$$

For two sibling variants v_i and v_j, their between-variant separation was measured in the same response space. A pair was classified as resolvable when its between-variant distance exceeded the average split-half distance of the two variants,

$$d_{\cos}\left(\delta_{v_i}, \delta_{v_j}\right) > \frac{D_{\text{self}}(v_i) + D_{\text{self}}(v_j)}{2}.$$

At the variant level, identification used the more stringent nearest-sibling criterion. Both statistics were averaged over the $S = 15$ split-half repeats before any threshold was applied, giving $\bar{D}_{\text{self}}(v)$ and, for the between-variant distances, $\bar{D}(v_i, v_j)$ computed between the two variants' A-half profiles. Writing $C(v_i)$ for the other scored variants of the same gene, variant v_i was classified as identifiable when

$$\min_{v_j \in C(v_i)} \bar{D}(v_i, v_j) > \bar{D}_{\text{self}}(v_i),$$

with exact ties classified as not identifiable. This variant-level criterion uses the variant's own split-half distance as its threshold and selects the nearest sibling on the same seed-averaged distance matrix; the cross-fitted nearest-competitor statistic defined below is a separate quantity.

Split-half reproducibility reference. To quantify the full-pool identification supported by the measurement itself, we constructed a split-half reproducibility reference using the same PDS evaluation applied to prediction models. For each held-out variant, its cells and the corresponding reference cells were divided into two disjoint groups. Pseudobulk responses were constructed separately from the two disjoint halves, and one response was treated as the prediction while the other was treated as the evaluation target. PDS was then calculated using the same candidate variants, distance function, tie handling and generalization setting as in the model evaluation.

The procedure was repeated over 15 random split-half subsamples and averaged for each variant before aggregation. This quantity measures within-experiment cell-sampling reproducibility at the specified sampling depth and protocol. It is an empirical reference rather than a mathematical bound: predictors that denoise responses or borrow information across variants can in principle exceed it.

The same ranking also supports a shortlist reading. For each held-out variant and each split-half repeat we retained the rank of the true allele among the same-gene candidates instead of only its percentile, and recorded whether that rank fell within $k \in \{1, 3, 5, 10\}$. Ties were resolved as the expected hit under uniform tie-breaking, the indicator analogue of the mid-rank convention PDS uses; one variant, GATA1 S274G, produced them. Hit rates were averaged over the 15 repeats and then over distinct variants, with 2,000-resample percentile bootstraps over distinct variants, matching the aggregation of the reference itself. Each split's candidate set is the union of its training and test variants, which is that gene's full scored pool for every split, so a variant held out under several generalization settings receives the same score under each and the effective sample is the 308 distinct variants. The chance rate is $\min(k/N, 1)$, the probability that a uniformly random ordering of a pool of N candidates places the true allele in the top k; N is 98, 92, 254 and 26 for TP53, KRAS, GATA1 and JAK1, so the chance rate is dataset-specific.

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, the same reference was recomputed with the two groups drawn from disjoint batch groups through the identical harness; this batch-disjoint sensitivity analysis, its implementation and its results are described in Supplementary Note 1. No dataset

analysed here contains independently repeated experiments, so this analysis is a within-experiment sensitivity check rather than cross-experiment validation.

Pairwise sibling discrimination. Full-pool identification requires a prediction to outrank an entire candidate set, so a weak PDS does not by itself establish that a prediction carries no allele-specific information. We therefore evaluated each prediction on individual sibling pairs, where the candidate set is two by construction and chance is 0.5 irrespective of the gene's pool size.

For a pair of variants v_i and v_j, a predictor resolves the pair when the correct assignment of its two predictions to the two measured responses is closer than the swap,

$$d(\hat{\delta}_i, \delta_i) + d(\hat{\delta}_j, \delta_j) < d(\hat{\delta}_i, \delta_j) + d(\hat{\delta}_j, \delta_i),$$

with d the cosine distance used throughout; exact ties score 0.5. The measurement arm applies the identical statistic with one disjoint cell half in place of the prediction, so both arms are scored against the same half-depth target rather than giving the models a deeper one.

A pair was called measurement-resolvable under the criterion already used for allele identification, namely a between-variant distance exceeding the mean of the two variants' split-half distances. Pair selection and pair evaluation used disjoint blocks of subsample seeds, so that the same sampling fluctuation cannot both determine eligibility and grade discrimination accuracy. Only held-out variants with model predictions were scored, and a pair held out under several generalization settings was averaged across them before entering the bootstrap, so that no pair contributed more than once. The displayed confidence intervals are 2,000-resample percentile bootstraps over distinct pairs. These intervals are descriptive because distinct pairs can share a variant. We therefore performed a variant-node bootstrap sensitivity analysis: the variants incident to the retained pair graph were sampled with replacement, each original edge was weighted by the product of the sampled multiplicities of its two endpoints, and the same draw was used for the measurement, every model and their paired difference. The sensitivity analysis used 1,000,000 draws and is reported in Supplementary Note 2. Predictions containing non-finite values were excluded from pairwise scoring and their counts are reported alongside the accuracies.

Nearest-competitor separation. We next examined which properties of the measured response geometry determine attainable full-pool allele identification. Within-condition sampling variation was estimated from disjoint split-half response profiles. Mean between-perturbation separation was corrected for this sampling variation to estimate the response separation attributable to perturbation identity rather than finite-cell noise.

Because PDS is a ranking statistic, we separately quantified separation from each perturbation's nearest competitor. Nearest-competitor separation was estimated by cross-fitting: the nearest sibling was selected using one measurement of the responses and its separation was evaluated using a second, disjoint measurement. This prevents the same sampling fluctuation from both selecting a competitor and making that competitor appear artificially close.

Normalized separation statistics. At depth m, each variant v contributes two profiles $S_{v,1}, S_{v,2} \in \mathbb{R}^K$ from disjoint groups of m cells, each a cell mean minus a shared reference half-mean. With $H = 2$ blocks per variant and n variants, the per-profile sampling noise is

$$\hat{\eta}^2 = \frac{1}{n}\sum_{v=1}^{n} \frac{1}{H-1}\sum_{h=1}^{H} \left\| S_{v,h} - \bar{S}_v \right\|^2, \qquad \bar{S}_v = \frac{1}{H}\sum_h S_{v,h}.$$

Mean pairwise separation relative to noise subtracts the noise carried by a distance between two block means,

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

Nearest-competitor separation replaces the mean over pairs with each variant's closest competitor, selected on one block and evaluated on the other, symmetrized over the two directions:

$$j_v^{(1)} = \arg\min_{u \neq v} \left\| S_{v,1} - S_{u,1} \right\|^2, \qquad j_v^{(2)} = \arg\min_{u \neq v} \left\| S_{v,2} - S_{u,2} \right\|^2,$$

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

The subtracted term is $2\hat{\eta}^2$ rather than $2\hat{\eta}^2/H$, because both distances are taken between single measurements rather than between block means. Cross-fitting prevents reuse of the same cell-sampling fluctuation for selecting the competitor and reading its distance; it does not by itself make the two estimates statistically independent. Under a working model of independent, mean-zero and exchangeable per-profile sampling errors with a common variance across variants, the global subtraction by $2\hat{\eta}^2$ gives s_v expectation zero when no between-variant signal is present. Cross-fitting alone does not establish this centring under variant-specific heteroskedasticity, because one pooled noise term need not equal the expected distance noise for every selected pair. Separations were computed within each of $S = 10$ independent cell partitions against the pooled $\hat{\eta}^2$, averaged across partitions, and only then summarized,

$$\rho^2_{\mathrm{nn}} = \frac{\mathrm{median}_v\left(\frac{1}{S}\sum_{s=1}^{S} s_v^{(s)}\right)}{2\hat{\eta}^2},$$

in that order, because separations formed from partition-averaged profiles would carry only $1/S$ of the noise the correction removes and would therefore be over-corrected. Zero rather than one is the nominal reference for both statistics under the exchangeable, common-variance noise conditions above, since the noise expectation has already been subtracted. Under those conditions, $\rho^2_{\mathrm{nn}} = 0$ means that nearest-competitor separation is what sampling noise alone produces; with heteroskedastic errors, departures from zero can also reflect imperfect centring by the pooled correction. Estimates are reported unclipped, so a negative value indicates a draw at or below the pooled noise floor rather than a worse separation. Intervals are 95% percentile bootstraps of the median over the n variants, with 2,000 resamples and a fixed seed, holding $\hat{\eta}^2$ and the cell partitions at their point estimates.

To distinguish effects of average signal magnitude from response geometry, we additionally generated controlled synthetic response configurations in which separation rank, amplitude distribution, candidate-set size and sampling depth were varied while overall signal relative to within-condition variation was controlled. Each configuration was passed through the same split-half PDS procedure. Mean pairwise separation and cross-fitted nearest-competitor separation were then compared with the attainable reproducibility reference across measured and controlled-resolution configurations. Simulation construction and the normalized separation metrics are described in Supplementary Note 4.

Benchmark-design analyses

Controlled predictor family. Measurement reproducibility at the level of individual perturbations does not by itself determine whether a benchmark can reliably distinguish predictors. We therefore constructed a family of predictors whose relative quality was fixed before evaluation. For variant v,

$$\hat{\delta}(v, \alpha) = (1 - \alpha)\bar{\delta}_g + \alpha \delta_v^{\mathrm{build}},$$

where $\bar{\delta}_g$ is the gene-shared mean response and $\delta_v^{\mathrm{build}}$ is a variant-specific response estimated from cells disjoint from those used for evaluation. Increasing α increases the contribution of variant-specific information and therefore defines an ordered predictor family independently of the evaluation data. The weights were $\alpha \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and the depths $m \in \{25, 50, 100, 150, 250\}$ cells per half. For each gene, depth and draw, a variant's cells were shuffled once and cut into a build half and a disjoint evaluation half of m cells, and the reference cells were cut by the same shuffle into two disjoint halves, so a build profile and its evaluation target share neither perturbed cells nor a reference draw. A variant entered a depth only if it held at least $2m$ cells, and only cohorts identical across draws were used. The gene-shared anchor is the mean of the build profiles over that gene-by-depth cohort, so both anchors are estimated in build cells alone. Because every variant receives the same prediction at $\alpha = 0$, the cohort-mean PDS there is exactly 0.5 by construction rather than an estimate.

Each predictor was scored by PDS against the held-out evaluation cells. Bootstrap resampling of held-out perturbations was then used to estimate two benchmark-level quantities: the probability of recovering the complete design-defined ordering of the predictor family and the probability of selecting the highest-quality predictor. These quantities were evaluated as functions of the split-half reproducibility reference and the number of perturbations available for scoring.

Each predictor's per-variant PDS was averaged over 10 subsample draws before aggregation, and the ordering and winner probabilities were estimated from 1,000 bootstrap resamples of the cohort's variants.

The same predictor construction was applied to Norman, Replogle, Adamson, the Virtual Cell Challenge training dataset[35] and sci-Plex[36]. For this cross-dataset comparison, perturbations with at least 100 cells were evaluated at 50 cells per half in 50 principal components, over the seven weights $\alpha \in \{0, 0.5, 0.7, 0.85, 0.925, 0.96, 1.0\}$, whose spacing narrows towards $\alpha = 1$, averaging 8 draws and bootstrapping 500 times. This arm differs from the allele arm in its dimension reduction and in its draw and resample counts, and is therefore compared within itself rather than against the allele-resolved numbers.

Candidate-set composition. Candidate-set size and candidate-set composition were varied separately, recomputing the split-half reproducibility reference at each candidate-set definition through the same harness used for model scoring. Here K is the total candidate-set size, including the target. In the uniform arm, $K-1$ competitors were drawn without replacement from the same-gene pool, always retaining the true target, at $K \in \{5, 10, 20, 25, 45, 90\}$ and the full pool, with 200 draws per seed. In the nearest-competitor arm, the $K-1$ competitors closest to the target were retained, selected on the cell half disjoint from the one used for scoring so that the same sampling fluctuation cannot both choose a competitor and make it appear close. For comparison across genes, we also evaluated the common subset size $K = 25$; the complete JAK1 pool contained $K = 26$ candidates.

Uniform subsampling leaves the expected score unchanged by construction. If the full pool contains $n-1$ competitors of which l are strictly closer to the prediction than the true target and $e-1$ are tied with it, then drawing $K-1$ competitors uniformly gives

$$E[L_K] = (K-1)\frac{l}{n-1}, \qquad E[E_K] = (K-1)\frac{e-1}{n-1},$$

and since $\mathrm{PDS}_K = 1 - (L_K + E_K/2)/(K-1)$,

$$E[\mathrm{PDS}_K] = \mathrm{PDS}_n.$$

Reducing the candidate set uniformly therefore changes the resolution of the percentile, not its expectation. Counts were sampled directly from the corresponding multivariate hypergeometric distribution rather than by enumerating candidate lists.

Sampling-depth analysis. To quantify how detection changes with measurement depth while holding perturbation identity fixed, the split-half analysis was repeated at 30, 50, 100 and 150 cells per half. A perturbation entered a depth condition only when at least twice the required number of cells was available. Within each gene, comparisons across depths used the cohort of perturbations evaluable at the deepest depth being compared, so changes along a depth curve reflect sampling depth rather than changing perturbation composition.

For the broader gene-perturbation datasets, native-depth detection was additionally compared with detection after subsampling to 50 cells per half using only perturbations evaluable under both conditions. The same $S > W$ criterion was used throughout.

Repeated-pilot aggregate for detection association. We assessed whether pilot-derived quantities were associated with a separately estimated detection label. Cells were reduced to 50 components by principal-component analysis, or by truncated singular value decomposition where the feature matrix was too wide for an exact decomposition. The component basis was fitted once to the stored cells selected for each screen before the pilot–evaluation partitions were drawn; evaluation cells therefore contributed to this unsupervised representation fit. Perturbed cells were randomly permuted ten times. Within each permutation, the first 50 cells formed the pilot pool and the next 200 cells formed the evaluation pool, which was divided into two 100-cell halves for the target depth, so the perturbed pilot and

evaluation cells were disjoint within that permutation. The same cells could recur in different roles after re-permutation. From the first $P \in \{25, 50\}$ pilot cells we computed a pilot effect magnitude and a training-free pilot signal-to-variation score,

$$\mathrm{eff}_P = \left\| \bar{x}_{\mathrm{pilot},P} - \bar{x}_{\mathrm{ref},P} \right\|_2, \qquad \mathrm{SNR}_P = \frac{\mathrm{eff}_P}{\left\| \bar{x}_{1:h} - \bar{x}_{h+1:2h} \right\|_2}, \qquad h = \lfloor P/2 \rfloor,$$

where $\bar{x}_{\mathrm{ref},P}$ is the centroid of P reference cells drawn without replacement from the control population, and the denominator of SNR_P is the distance between the centroids of two disjoint $h = \lfloor P/2 \rfloor$-cell subsets of the same pilot pool, so that it measures within-perturbation cell-sampling variation alone. For odd P, one pilot cell is unused in this denominator. The 25-cell pilot is nested within the 50-cell pilot within each permutation. Reference cells for the pilot and evaluation quantities were sampled independently without replacement within each draw but were not constrained to be mutually exclusive either within or across permutations.

For each perturbation, eff_P and SNR_P were first averaged over the ten permutations. Detection eligibility at a target depth of 100 cells per half was also defined only after aggregating the ten evaluation draws. Specifically, S was the median evaluation perturbation-to-reference distance minus the median evaluation split-half distance, W was the difference between the 97.5th and 2.5th percentiles of the ten split-half distances, and the aggregate label was $\mathbf{1}(S > W)$. The analysis therefore does not produce ten repeat-level AUROCs. It relates a ten-repeat mean pilot score to a ten-repeat aggregate evaluation label and estimates one AUROC per dataset.

The learned score was a logistic regression on the single standardized feature eff_P, fitted leave-one-dataset-out; because the model carries one feature with a positive fitted coefficient, its predicted probability is a monotone function of eff_P, so the fit affects calibration rather than ranking. The training-free score used SNR_P directly, with no fitting and no threshold. AUROC was calculated within each held-out dataset rather than after pooling datasets with different class prevalences. It is reported for Replogle, Norman and Adamson, the three external gene-perturbation screens named in advance for this analysis; the allele-resolved datasets are not used to estimate it, because TP53 and KRAS carry a single eligibility class. Intervals are 95% percentile bootstraps over perturbations with 2,000 resamples, discarding any resample that does not contain both classes; the AUROC is tie-aware, so tied aggregate pilot scores contribute one half. Because the analysis aggregates across re-partitions and permits cross-repetition cell reuse, the reported AUROCs quantify a repeated-pilot aggregate rather than the predictive performance of one fixed 25- or 50-cell pilot.

Sampling-depth projection. The controlled-geometry sweep was inverted into a sampling-depth projection in four steps: choosing a measurement axis by evidence, screening candidate outcomes for monotonicity, calibrating the retained outcome against the axis, and inverting that calibration under a noise-scaling assumption.

Axis and outcome. The axis is the median cross-fitted nearest-competitor separation ρ^2_{nn} defined above, and the calibrated outcome is the split-half reproducibility reference. Both were selected from candidates by prespecified screens on the sweep, one ranking candidate axes by their association with the reference and one requiring the calibrated outcome to be monotone in the axis over its upper range; the two ordering-recovery probabilities fail that second screen and are not inverted. The screens and their values are reported in Supplementary Note 5.

Calibration. Configurations were stratified by candidate-set size into $n \leq 30$, $30 < n \leq 70$ and $n > 70$ (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 outcome of that bin; and the knot ordinates were then replaced by their running maximum, which imposes monotonicity and nothing further. All three strata yielded 24 knots. The calibration frames 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. For a target τ, let $x^*(n, \tau)$ be the smallest knot abscissa whose monotonized ordinate reaches τ; values between knots were not interpolated. The nearest-competitor separation is a property of the perturbations and is taken to be independent of sampling depth, whereas the per-profile

noise scales as $\hat{\eta}^2 \propto 1/m$, so that $\rho^2_{\mathrm{nn}}(m) = \rho^2_{\mathrm{nn}}(m_0)\, m/m_0$ and the depth projected from a pilot at m_0 cells per group is

$$\hat{m}(\tau) = m_0 \, \frac{x^*(n,\tau)}{\rho^2_{\mathrm{nn}}(m_0)}.$$

Targets were $\tau \in \{0.7, 0.9\}$ on the split-half reproducibility reference, whose chance value is 0.50; they are not thresholds on any model score. Candidate-set size enters the projection only by selecting one of the three strata and does not otherwise appear in the returned number. The above-noise fraction was deliberately not inverted, because a threshold-crossing fraction does not scale with depth.

Refusal and scope. No depth is returned when the stratum held fewer than 20 configurations, when the calibration never reaches τ at that candidate-set size, when the pilot estimate satisfies $\rho^2_{\mathrm{nn}}(m_0) \leq 0$, or when the pilot's lower confidence bound does not exceed zero. The last gate prevents separations whose intervals cover zero from returning depths of thousands of cells per perturbation that would be extrapolations from noise. No interval is placed on a returned depth; the pilot lower bound acts as a gate and is not propagated. The constructed configurations span candidate-set sizes of 20, 45 and 90, whereas the finite projections are read off at 247 and at 14, 9 and 5, so every reported projection extrapolates in candidate-set size and is a model-based projection rather than a validated depth requirement. The full calibration, including the axis-selection and monotonicity screens and the per-dataset projections, is given in Supplementary Note 5.

Benchmark-size resolution. The number of variants required to compare models was measured on the design-ordered predictor family, whose relative quality is fixed before evaluation. Difficulty was indexed by the realized difference in PDS between the two predictors on the full cohort, not by the interpolation weight separating them: the map from weight to score saturates, so equal weight gaps correspond to very unequal score gaps. All 15 weight pairs were used, giving a range of realized differences.

For each gene and depth, held-out variants were subsampled *without* replacement at $N_{\mathrm{test}} \in \{5, 10, 20, 40, 80, 160\}$, 200 draws each, and both predictors were scored on the same draw. The reported quantity is the fraction of draws in which the better predictor received the higher mean score. A full-cohort point was also computed, but it is a single deterministic evaluation rather than a 200-draw estimate and is not included in the grouped curves.

Results are grouped by two prespecified sets of edges. The split-half reproducibility reference of a gene-by-depth condition, taken as the score of the fully variant-specific member ($\alpha = 1$) on the disjoint evaluation half, is binned as $r \leq 0.52$, $0.52 < r \leq 0.65$ and $r > 0.65$, where 0.65 is the prespecified high-reproducibility operating point used for the broader perturbation panel and 0.52 is the upper edge of the near-chance band. This operating point describes measurement reproducibility; it does not by itself imply that a benchmark can reliably order models. The realized PDS difference is binned as $\Delta \leq 0.01$, $0.01 < \Delta \leq 0.05$ and $\Delta > 0.05$, where 0.05 is the equivalence margin used elsewhere in this work. Both sets of edges were fixed before this analysis. Only weight pairs with a positive realized difference entered the grouping, so the reported probabilities are conditional on the design ordering being realized in the measured scores. Across the released table, including deterministic full-cohort evaluations, 318 of 1,185 gene-by-depth-by-pair-by-size rows had a non-positive difference, all in TP53 and KRAS. Among the 945 subsampled-benchmark rows eligible for the curves, this criterion excluded 265 rows and retained 680. Within a bin, recovery probabilities were pooled across the contributing gene-by-depth conditions and weight pairs at each N_{test}, and we report the smallest tested N_{test} at which that mean reaches 0.9; no monotonicity was imposed, although every reported curve was monotone. Because a condition contributes only at sizes its cohort can supply, the set of conditions averaged at a given N_{test} is not fixed across the ladder. Where that smallest value is also the smallest size tested, it is written as ≤ 5 and denotes a bound rather than an estimated threshold. Combinations of reference and difference that do not occur are structurally unreachable, because a dataset near the measurement floor cannot generate a large realized difference. Cohort size falls as depth rises because a variant must contribute cells to both halves, so the table is read within a gene–depth condition rather than across depths.

Broader perturbation panel

Prespecified criterion.

The broader arm of PertResolve-Bench evaluates measurement resolution beyond allele-resolved perturbation data using a prespecified three-part criterion for perturbation detection, perturbation identification and benchmark-level ranking resolution. Eligible screens required a defined perturbation label, an identifiable reference condition and at least five perturbations represented by at least 200 cells.

The analysis was initially specified for six screens: Norman, Replogle, Adamson, the Virtual Cell Challenge training dataset[35], the multiplexed genetic and chemical perturbation screen of McFarland *et al.*[37], and a subset of Tahoe-100M[38]. The same criterion and decision thresholds were subsequently applied to an expanded panel comprising 31 configurations across 14 public resources. A configuration denotes one resource evaluated within one cellular context and one readout modality; configurations are therefore nested within resources rather than statistically independent resources. The panel represents a selected evaluation set under the stated eligibility criteria rather than an exhaustive census of eligible public datasets. McFarland and Tahoe-100M were evaluated using predefined subsamples of their larger resources.

For each configuration, the primary analysis used 50 cells per group, eight random cell partitions, 1,000 bootstrap resamples and a 50-component principal-component representation. RNA measurements were analysed from the stored count matrix without library-size normalization, log transformation or gene selection, with the principal-component basis fitted separately within each resource and context before cell partitioning. For readouts with at most 50 measured features, including the surface-protein datasets, distances were computed directly in the measured feature space. A screen was classified as detectable when more than 50% of perturbations met the perturbation-versus-reference criterion and identifiable when more than 50% met the nearest-competitor criterion; a split-half reproducibility reference above 0.65 defined the prespecified high-reproducibility operating point. This last label does not assert reliable model ordering, which additionally depends on model separation and benchmark size. Sensitivity to preprocessing and implementation choices, including expression representation, is reported in Supplementary Note 5.

Count-space panel pipeline and depth-ladder pipeline. Two pipelines produce the detection fractions reported in this work. The *panel criterion* computes distances in 50 principal components of the unnormalized count matrix, with no library-size normalization, no log transformation and no gene selection, over eight random cell partitions, including a perturbation when it holds at least 200 cells. It is the prespecified criterion, and it produces every detection fraction in Fig. 5 and the reference arm of Supplementary Fig. 2. The *split-half depth ladder* computes distances in 50 principal components of each dataset's own analysis space, which for the three gene-perturbation screens is library-size-normalized, log-transformed expression restricted to 2,000 highly variable genes and for the allele-resolved datasets is the native space of that dataset, over 50 random cell splits, including a perturbation when it holds at least twice the depth being tested; it is the pipeline behind Fig. 4c,d and Fig. 6f. The two pipelines and the two cell-inclusion thresholds each shift the absolute detection fraction, in opposite directions, so detection fractions are compared only within a pipeline and inclusion threshold.

Statistical analysis

The primary statistical unit for model evaluation was the distinct protein-coding variant. For PDS, scores were first averaged over 15 random cell subsamples within each gene–setting–variant unit. When a variant was evaluated under more than one generalization setting, these setting-level means were then averaged before aggregation over distinct $(\text{gene}, \text{variant})$ pairs. Point estimates and 95% confidence intervals were obtained using a 2,000-resample nonparametric percentile bootstrap over these pairs. Confidence intervals excluding 0.50 were interpreted against the exact full-candidate uniform-matching expectation. The task-matched restricted-label permutation tests described below address a different conditional reference and are reported separately.

Pearson-δ was aggregated over the same distinct-variant unit. When a variant was evaluated under more than one generalization setting, its scores were averaged across settings before aggregation. Non-finite prediction vectors were excluded from Pearson-δ rather than represented by zero, and their coverage is reported separately. Point estimates and 95% confidence intervals were calculated using the

same 2,000-resample percentile bootstrap over distinct (gene, variant) pairs. Positive direction recovery was defined by a lower 95% confidence bound above zero.

The tie-aware mid-rank definition gives an exact expectation of 0.50 when the target label is assigned uniformly over every column of the complete candidate pool. As an implementation check for the primary feature-based predictions, we also constructed held-out-only square distance matrices within each gene and evaluation setting and permuted the prediction rows 1,000 times; both the permuted labels and candidate columns in this check were restricted to the held-out set, so its expectation is likewise 0.50. This check produced the grey range in Fig. 2c but is not a calibration that permits finite candidate sets or ties to move the theoretical chance level.

Task-matched permutations for the residual analysis and PerturbNet instead retained all training and held-out candidate columns used by the reported PDS while reassigning prediction rows only among the available held-out target labels within each gene and evaluation setting. The restricted held-out label set is not generally uniform over the complete ranking pool, so this conditional permutation distribution need not be centred at 0.50. The setting-specific distance matrix and, for PerturbNet, its 50-dimensional output representation were held fixed. PerturbNet used the same label permutation across its 15 cell-subsample seeds; scores were averaged first over seeds within gene–setting–variant, next across settings within gene–variant, and finally over 303 distinct variants, matching the reported estimator. With 1,000,000 permutations and random seed 0, its observed PDS was 0.530143, the exact restricted-label null mean was 0.515500, the central 95% permutation interval was 0.496171–0.534942 and the one-sided plus-one permutation value was 0.069768. Residual-axis permutations used the same restricted-label/full-candidate principle. They were run in two stages, with 50 permutations for every method and 1,000 for methods whose preliminary value was below 0.05; residual-axis P values were adjusted across the 25 scored methods. Additional details are provided in Supplementary Note 2.

Detection and identification fractions were treated as binomial proportions and are reported with 95% Wilson score intervals. Native-versus-depth-matched detection comparisons used a paired 2,000-resample bootstrap over perturbations, with each resample applied to both depth conditions. Confidence intervals for continuous per-gene measurement quantities were obtained by nonparametric bootstrap over perturbations unless otherwise stated.

All stochastic analyses used fixed random seeds. No statistical method was used to predetermine the number of biological conditions. For the primary allele-resolved benchmark, all identified public datasets and variant conditions meeting the predefined eligibility criteria were retained. A residue-clustered bootstrap assessing dependence among substitutions at the same residue is reported in Supplementary Note 2.

Software

Primary analyses were performed in Python 3.11.14 using NumPy 2.2.6, pandas 2.3.3, SciPy 1.14.1 and scikit-learn 1.7.2. Single-cell data processing used scanpy 1.10.3 in Python and `scran` with `SingleCellExperiment` in R/Bioconductor. Protein-language-model embeddings were generated using fair-esm 2.0.0 and PyTorch 2.10.0 with CUDA 12.6.

Released-model implementations were run in package-specific environments. Biolord 0.0.3 and the scGen and scVIDR adaptations used scvi-tools 1.4.3; CellFlow used `cellflow-tools` 0.1.1 and JAX; and the CPA adaptation used `cpa-tools` 0.8.8 with scvi-tools 0.20.3. Random seeds were fixed throughout the analysis.

Acknowledgements

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

Competing interests

The authors declare no competing interests.

Data availability

All single-cell datasets analysed in this study are publicly available. TP53 and KRAS Perturb-seq data are available from the Gene Expression Omnibus under accession GSE161824[17]. GATA1 Perturb(BE)-seq data are available under accession GSE215253[18]; the processed expression matrix and predefined partitions used here were obtained from the PerturbNet repository (`cyclopeta/PerturbNet_reproduce`)[27]. JAK1 scSNV-seq data are available from the European Nucleotide Archive under accession PRJEB48915, with processed objects deposited at Zenodo (10.5281/zenodo.10418435)[19]. The Replogle, Norman and Adamson datasets used in the measurement-resolution analyses were obtained through scPerturb[34]. Sources and accession information for the additional perturbation datasets are provided in Supplementary Data 1.

Derived PertResolve-Bench tables, including allele-resolved variant annotations, cell counts, evaluation partitions, per-variant prediction results and measurement-resolution results for the broader perturbation panel, are available at `https://github.com/Boom5426/PertResolve`. Expression matrices and protein-language-model embeddings are not redistributed where restricted by size or source distribution and can be reconstructed from the public datasets using the preprocessing procedures described in Methods.

Code availability

Code for constructing PertResolve-Bench, fitting predictors, computing PertResolve-Eval metrics, performing measurement-resolution analyses and reproducing the main figures is available at `https://github.com/Boom5426/PertResolve`.

PertResolve-Eval is distributed as the `pertresolve` Python package with both Python API and command-line interfaces for applying the PertResolve workflow directly to new single-cell perturbation datasets, including AnnData (`.h5ad`) inputs. Given perturbation labels and a reference condition, the framework reports perturbation detectability, nearest-competitor identification, split-half reproducibility, local response geometry and benchmark-level model-ranking resolution.

References

[1] Charlotte Bunne, Yusuf Roohani, Yanay Rosen, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187:7045–7063, 2024. doi: 10.1016/j.cell.2024.11.015.

[2] Haotian Cui, Alejandro Tejada-Lapuerta, Maria Brbić, et al. Towards multimodal foundation models in molecular cell biology. *Nature*, 640:623–633, 2025. doi: 10.1038/s41586-025-08710-y.

[3] Ewen Callaway. Can AI build a virtual cell? Scientists race to model life's smallest unit. *Nature*, 643: 13–14, 2025. doi: 10.1038/d41586-025-02011-0. News feature.

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

[5] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19:e11517, 2023. doi: 10.15252/msb.202211517.

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

[7] Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, et al. CellFlow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025. doi: 10.1101/2025.04.11.648220.

[8] Haotian Cui, Chloe Wang, Hassaan Maan, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

[9] Minsheng Hao, Jing Gong, Xin Zeng, et al. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21:1481–1491, 2024. doi: 10.1038/s41592-024-02305-7.

[10] Yanay Rosen, Yusuf Roohani, Ayush Agrawal, et al. Universal cell embedding provides a foundation model for cell biology. *Nature*, 656:183–191, 2026. doi: 10.1038/s41586-026-10689-z.

[11] Yiming Li, Min Zeng, Jun Zhu, et al. UniPert-G2CP bridges genetic and chemical screens from molecular representation to phenotype modeling. *Cell*, 189:4946–4963.e8, 2026. doi: 10.1016/j.cell.2026.06.005.

[12] Britt Adamson, Thomas M. Norman, Marco Jost, et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167:1867–1882.e21, 2016. doi: 10.1016/j.cell.2016.11.048.

[13] Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185:2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013.

[14] Evan H. Baugh, Hua Ke, Arnold J. Levine, Richard A. Bonneau, and Chang S. Chan. Why are there hotspot mutations in the TP53 gene in human cancers? *Cell Death & Differentiation*, 25:154–160, 2018. doi: 10.1038/cdd.2017.180.

[15] John C. Hunter, Anuj Manandhar, Martin A. Carrasco, Deepak Gurbani, Sudershan Gondi, and Kenneth D. Westover. Biochemical and structural analysis of common cancer-associated KRAS mutations. *Molecular Cancer Research*, 13:1325–1335, 2015. doi: 10.1158/1541-7786.MCR-15-0203.

[16] Nidhi Sahni, Song Yi, Mikko Taipale, et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, 161:647–660, 2015. doi: 10.1016/j.cell.2015.04.013.

[17] Oana Ursu, James T. Neal, Emily Shea, et al. Massively parallel phenotyping of coding variants in cancer with Perturb-seq. *Nature Biotechnology*, 40:896–905, 2022. doi: 10.1038/s41587-021-01160-7.

[18] Jorge D. Martin-Rufino, Nicole Castano, Michael Pang, et al. Massively parallel base editing to map variant effects in human hematopoiesis. *Cell*, 186:2456–2474.e24, 2023. doi: 10.1016/j.cell.2023.03.035.

[19] Sarah E. Cooper, Matthew A. Coelho, Magdalena E. Strauss, et al. scSNV-seq: high-throughput phenotyping of single nucleotide variants by coupled single-cell genotyping and transcriptomics. *Genome Biology*, 25:20, 2024. doi: 10.1186/s13059-024-03169-y.

[20] Heon Seok Kim, Susan M. Grimes, Tianqi Chen, Anuja Sathe, Billy T. Lau, Gue-Ho Hwang, Sangsu Bae, and Hanlee P. Ji. Direct measurement of engineered cancer mutations and their transcriptional phenotypes in single cells. *Nature Biotechnology*, 42:1254–1262, 2024. doi: 10.1038/s41587-023-01949-8.

[21] Douglas M. Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11:801–807, 2014. doi: 10.1038/nmeth.3027.

[22] Pascal Notin, Aaron Kollasch, Daniel Ritter, et al. ProteinGym: large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems*, volume 36, pages 64331–64379, 2023. doi: 10.52202/075280-2810.

[23] Daniel Esposito, Jochen Weile, Jay Shendure, et al. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology*, 20:223, 2019. doi: 10.1186/s13059-019-1845-6.

[24] 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: 1657–1661, 2025. doi: 10.1038/s41592-025-02772-6.

[25] Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, et al. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, 44:1050–1059, 2026. doi: 10.1038/s41587-025-02777-8.

[26] Jordan W. Squair, Matthieu Gautier, Claudia Kathe, et al. Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12:5692, 2021. doi: 10.1038/s41467-021-25960-2.

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

[28] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*, volume 34, pages 29287–29303, 2021.

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

[30] Omar Kana, Rance Nault, David Filipovic, Daniel Marri, Tim Zacharewski, and Sudin Bhattacharya. Generative modeling of single-cell gene expression for dose-dependent chemical perturbations. *Patterns*, 4(8):100817, 2023. doi: 10.1016/j.patter.2023.100817.

[31] Jun Cheng, Guido Novati, Joshua Pan, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381:eadg7492, 2023. doi: 10.1126/science.adg7492.

[32] Abhinav K. Adduri, Dhruv Gautam, et al. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 2026. doi: 10.1016/j.cell.2026.07.052. In press, corrected proof.

[33] Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365:786–793, 2019. doi: 10.1126/science.aax4438.

[34] Stefan Peidli, Tessa D. Green, Ciyue Shen, et al. scPerturb: harmonized single-cell perturbation data. *Nature Methods*, 21:531–540, 2024. doi: 10.1038/s41592-023-02144-y.

[35] Yusuf H. Roohani, Tony J. Hua, Po-Yuan Tung, et al. Virtual Cell Challenge: toward a Turing test for the virtual cell. *Cell*, 188:3370–3374, 2025. doi: 10.1016/j.cell.2025.06.008.

[36] Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367:45–51, 2020. doi: 10.1126/science.aax6234.

[37] James M. McFarland, Brenton R. Paolella, Allison Warren, et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nature Communications*, 11:4296, 2020. doi: 10.1038/s41467-020-17440-w.

[38] Jesse Zhang, Airol A Ubas, Richard de Borja, et al. Tahoe-100M: a giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025. doi: 10.1101/2025.02.20.639398. Preprint.