Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Benchmarking partner-specific protein representations for phenotype discrimination in human allelic heterogeneity

Yu Zhang[1]
Second Affiliated Hospital & Yuying Children's Hospital of Wenzhou Medical University
Da Lin[1]
Second Affiliated Hospital & Yuying Children's Hospital of Wenzhou Medical University
Ying Chen[2]
Wenzhou Medical University
Yue Liu[2]
Wenzhou Medical University
Yi Zha[1]
Second Affiliated Hospital & Yuying Children's Hospital of Wenzhou Medical University
Wangqiang Feng
healthwz@163.com

Second Affiliated Hospital & Yuying Children's Hospital of Wenzhou Medical University

Research Article

Keywords: allelic heterogeneity, human genetic variation, genotype-phenotype interpretation, protein–protein interactions, protein interfaces, partner-specific representations

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11412735/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Pathogenic variants within the same human gene can cause distinct Mendelian phenotypes. Partner-specific protein representations provide molecular context for studying this allelic heterogeneity. Their technical availability, phenotype-balanced coverage and incremental phenotype information require separate evaluation. We benchmarked structural/interface representations, MutPred-PPI disruption scores and sequence-based SPPIDER-seq and X-PAIR representations in a ClinVar-based cohort. The labeled-versus-sorted control removed named partner assignments by sorting each residue's raw score multiset before independent standardization. Cohort eligibility and statistical procedures were specified before phenotype testing. The combined support criterion required a positive median ΔR^2 with its 95% confidence interval (CI) strictly above zero. Positive ΔR^2 was also required in at least 70% of phenotype-ready genes.

Results

On a common denominator of 132 genes, MutPred-PPI scored 10,917/11,150 submitted structural rows (97.91%). Nine genes reached the representation-coverage threshold, and three met phenotype-balance criteria. Sequence-based representations expanded usable coverage. SPPIDER-seq reached the threshold in 127 genes, with 23 phenotype-balanced genes; X-PAIR reached it in 107 genes, with 15 phenotype-balanced genes. The primary cohort comprised 10 genes and 547 unique wild-type (WT) residue analysis units. X-PAIR median ΔR^2 was 0.00179553 (95% gene-bootstrap CI − 0.00410526 to 0.00495717), with positive ΔR^2 in 8/10 genes. The prespecified combined criterion was not met, and cross-model findings limited generalization. Distinct substitutions at the same WT residue received the same residue × partner vector. The contribution of this design property to the global result was not directly tested.

Conclusions

Sequence-based partner representations substantially expanded phenotype-balanced coverage. Named partner identity provided no robust global incremental phenotype information within the evaluated cohort under the prespecified criterion. The framework distinguishes molecular feature availability from information relevant to phenotype discrimination. Phenotype-resolved interpretation may require combining substitution-specific molecular consequences with partner context and disease-relevant biological context.

Background

Pathogenic variants in the same human gene can cause distinct Mendelian phenotypes, complicating interpretation beyond identifying a damaging allele [1]. Pathogenicity alone does not specify which phenotype a variant may produce. The explanatory value of variant position, type and domain location also varies across genes [1]. Phenotype-specific models have begun to address this problem directly. Variant-to-Phenotype (V2P) conditions variant-effect prediction on disease phenotypes [2]. PheMART integrates protein representations, protein-protein interactions (PPIs), domains, biomedical knowledge and clinical context to connect missense variants with phenotypic consequences [3]. Their performance motivates examining which molecular inputs contribute information within individual genes [2, 3]. Integrating several sources of evidence can leave the contribution of each feature class unresolved. An attribution test can distinguish information carried by named interaction partners from information shared with more generic molecular features.

Experimental and structural studies provide a biological rationale for partner-specific representations. Structural network reconstruction connected disease variants with interaction surfaces [4]. Edgetic experiments showed that pathogenic alleles can selectively perturb interactions without removing the protein from the network [5, 6]. dSysMap linked mutations at different interfaces to different disease associations [7]. PIONEER extended structurally informed, partner-specific interface mapping to a larger interaction network [8]. These studies support evaluating interaction context as one molecular dimension of allelic heterogeneity [6, 8].

The available tools encode different molecular quantities. MutPred-PPI predicts missense-induced, interaction-specific perturbations; SPPIDER-seq predicts partner-conditioned residue interfaces from sequence; X-PAIR extends scalable partner-specific interface prediction [9, 10, 11]. Their usual endpoints include interface localization, interaction disruption or enrichment of disease variants [6, 8, 9]. These endpoints provide evidence distinct from phenotype discrimination among pathogenic alleles within the same gene. Phenotype-specific models ask which consequence a variant may have, whereas partner-aware representations describe potentially relevant molecular contexts [2, 3, 10]. The unresolved attribution question is whether named partner identity provides incremental phenotype information beyond the raw score multiset. Answering this question requires accounting for coverage differences across disease groups. Even with balanced coverage, a score-label association could reflect overall interface propensity rather than the assignment of scores to individual partners.

We addressed this question by evaluating technical scoreability, phenotype-balanced representation coverage and incremental phenotype information in sequence. We assessed structural interfaces, MutPred-PPI and sequence-based partner representations in a ClinVar-based cohort. We then compared labeled partner profiles with an identity-ablation control that sorts scores independently within each residue. Sorting removes named partner assignments while preserving the raw score multiset and dimensionality; the labeled and sorted profiles are then standardized independently. The comparison tests whether disease labels explain more variation when partner identity is retained. Unique wild-type (WT) residues defined the analysis units. Cohort eligibility, partner identities, statistical procedures and the combined cross-gene support criterion were specified before phenotype testing. This framework

distinguishes molecular feature availability from phenotype information attributable to partner assignments among pathogenic alleles.

Methods

Study design

We evaluated technical scoreability, phenotype-balanced coverage and incremental phenotype information in sequence. Structural interfaces and MutPred-PPI disruption scores were assessed before sequence-based representations. Disease labels informed coverage balance before phenotype testing. Model selection depended only on phenotype-blind structural evaluation, which designated X-PAIR as primary and SPPIDER-seq as confirmatory. The selection rule, analysis specification and eligibility criteria were frozen before phenotype-association testing. Table 1 summarizes the criteria used at each evaluation stage.

Table 1
Evaluation criteria and evidential boundaries

Stage	Scientific question	Prespecified criterion	Observed result	Interpretation
Technical scoreability	Can the representation be generated?	Exact-version valid output; explicit failures	MutPred-PPI: 10,917/11,150 (97.91%)	High scoreability ≠ broad usability
Multi-partner coverage	Is molecular context sufficiently complete?	≥ 3 credible valid partners; C4 ≥ 0.70	MutPred-PPI 9; binary 49; distance 9; SPPIDER-seq 127; X-PAIR 107	Missing dimensions constrain inference
Phenotype balance	Are disease groups comparably represented?	Prespecified disease-group balance; ≥2 groups; ≥10 variants/group; ≥3 dimensions Primary: ≥10 unique residues/group	MutPred-PPI 3; binary 10; distance 3; SPPIDER-seq 23; X-PAIR 15	Balanced availability ≠ phenotype information
Partner identity	Do named partners add information?	A: positive median ΔR^2 with 95% CI entirely above zero B: ≥70% positive genes Require A AND B	Median ΔR^2 0.0018; 95% CI crosses 0; 8/10 positive	Robust global support not established
Robustness	Is evidence representation-invariant?	Confirmatory method + prespecified sensitivities No additional binary cutoff	Spearman − 0.745455; 4/10 positive in both	Method-specific ΔR^2 estimates limit generalization

Coverage and balance outcomes use the common 132-gene denominator. C4 is the fraction of variants with ≥ 3 valid credible partners. Robustness reports the prespecified confirmatory and sensitivity evidence; no additional binary threshold is introduced. The global 95% CI is − 0.00410526 to 0.00495717. Exact results remain in the supplementary workbook.

ClinVar cohort and phenotype harmonization

The ClinVar variant-summary snapshot was dated 3 September 2026 [12]; the monthly reference ClinVar (RCV) XML archive had an internal release date of 29 August 2026. We selected pathogenic/likely pathogenic missense records using prespecified quality, assertion and coordinate-mapping rules. Variation IDs, substitutions and RCV assertions retained separate provenance. The strict cohort required ≥ 2 disease groups with ≥ 10 qualifying variants each. Assertions were harmonized to Mondo Disease

Ontology (MONDO) entities [13] through exact identifiers, synonyms and curated mappings. Parent-child collapse followed prespecified rules; broad assertions were not assigned to specific children. Multilabel and unresolved records remained explicit; groups were not revised after testing.

Protein interaction partner universe

Credible partners were identified from physical interaction records in BioGRID [14] and IntAct [15], together with prespecified high-confidence predictions. We retained evidence tiers and selected partners independently of phenotype-association results. Five confirmed mechanism-incompatible genes were excluded, leaving 132 genes not known to be incompatible. The 128-gene interaction-qualified intersection included 42 structure-ready genes. The common denominator contained 9,978 variant rows and 1,109,989 variant-partner rows. Missing features remained missing; pair-inference success and variant-level coverage used separate denominators.

Partner-specific representations

Experimental biological assemblies were mapped to canonical residues. Observed heavy-atom contacts within 5 Å defined interfaces, with a residue-level union across assemblies for each query-partner pair. Unobserved residues were excluded from contact validation. Binary occupancy and interface distance were separate features. MutPred-PPI v1.0.0 used official neural inference with ProtT5; SPPIDER-seq v2.0-manuscript used the receptor head with ESM2 and windowed full-sequence inference. X-PAIR 0.1.0 used its pretrained checkpoint and Ankh-large. For X-PAIR, both proteins in an interaction pair were required to be 50 – 2,000 residues long. Unsupported inputs were not truncated. Interaction-existence scores were not used to filter interface predictions. WT embeddings and residue tracks were reused across substitutions. Additional file 1: Supplementary Methods and Figures contains exact versions, hashes, implementation checks and Figures S1–S8. Additional file 2: Supplementary Tables and full-precision results contains Tables S1–S9 and the phenotype-analysis and provenance sheets.

Representation coverage and phenotype-balance criteria

Representation required valid features for ≥ 3 distinct credible partners. C4 was the fraction of common-denominator variants meeting this requirement; gene eligibility required C4 ≥ 0.70. Balanced coverage required minor disease-group imbalance, ≥ 2 unambiguous groups with ≥ 10 variants each, and ≥ 3 valid partner dimensions. Additional file 1 defines minor, material and severe imbalance. Phenotype-blind contact validation used macro area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC). Uncertainty was estimated with 1,000 whole-protein bootstrap resamples. Contact-validation results were interpreted as internal estimates because overlap with the models' training data was unknown. Partner-replacement, length and resource checks were also phenotype-blind.

Unique-residue phenotype cohort

One gene × canonical WT residue defined the primary analysis unit. Accession, coordinate and reference-amino-acid checks permitted exact canonical mappings or safe isoform remapping. Registry categories were assigned in the following order of precedence: multilabel variant, ambiguous disease, exclusive cross-disease and single disease. A broader cross-disease-linked flag included any residue connected to ≥ 2 groups. Primary analysis units required safe single-disease labels and valid full tracks from both models, with ≥ 2 groups of ≥ 10 unique residues and ≥ 3 common partners per gene. Multiple substitutions, repeated assertions and review-star weights did not increase the number of analysis units. A secondary SPPIDER-seq extension evaluated genes outside the shared cohort under the same residue and disease-group thresholds, requiring SPPIDER availability alone. Its eligibility, partner sets and correction families were prespecified separately, and it was excluded from the primary global result.

Labeled-versus-sorted phenotype-information test

Within each gene, named partner columns were standardized across primary residues using population means and standard deviations, without disease labels. Constant columns became zero. Euclidean permutational multivariate analysis of variance (PERMANOVA) estimated R^2=SS_between/SS_total and pseudo-F=(SS_between/(g − 1))/(SS_within/(n − g)), with g groups and n residues. Raw control vectors were sorted descending, then independently standardized by rank column. Sorting retained the raw score multiset before standardization while removing named partner assignments. $\Delta R^2 = R^2_{labeled} - R^2_{sorted}$ was an operational information contrast, not a covariate-adjusted causal partner effect. Zero-total-distance matrices were unevaluable.

Labeled and sorted profiles retained identical WT-residue analysis units and dimensionality within each gene. Partner counts varied across genes, so genes were analyzed separately and given equal weight in the global summary. Their profiles were not pooled into a common feature space.

Permutation testing and global support criterion

For each gene, we generated 10,000 unique alternative label assignments that preserved group sizes and excluded the observed assignment. Assignments were shared across models and contrasts. One-sided plus-one *P* values included ties; labeled association used pseudo-F and the paired contrast used ΔR^2. Benjamini-Hochberg false discovery rate (FDR) correction retained four separate ten-gene families: X-PAIR labeled, X-PAIR Δ, SPPIDER labeled and SPPIDER Δ, with $q \leq 0.05$. A 10,000-resample percentile gene bootstrap weighted genes equally. Condition A required positive median ΔR^2 with a 95% confidence interval (CI) strictly above zero; Condition B required positive ΔR^2 in ≥ 70% of all phenotype-ready genes. Both were required. The secondary one-sided sign test did not replace this criterion.

Sensitivity analyses

Restricted permutations exchanged labels within fixed position-quintile, domain-proxy and monomer surface/buried strata, preserving group sizes. Informative small assignment spaces were enumerated without merging bins. Prespecified raw-score, poststandardization-sort and correlation-distance sensitivities were retained. Generic position, domain-proxy, three-dimensional (3D) and interface baselines used matched complete-case rows and identical group-size requirements without imputation. Secondary nearest-centroid comparisons used ten repeats of stratified five-fold splits. Scaling was fitted within training folds, and no tuning was performed. The extension was separate from primary bootstrap and multiplicity summaries. Broader-registry same-residue summaries were not tested as explanations of ΔR^2.

Reproducibility and access to data and code

Analysis choices preceded phenotype testing, and all 346 file-integrity checks passed. The specification was prospectively fixed but was not registered in an independent registry. Saved null arrays, assignments, manifests and exact software/model versions support reproduction through the public repository and archive. Additional file 1 describes implementation and failures. Tables S1-S9 document models, coverage, balance, failures, validation, length constraints and resources; Additional file 2 contains full-precision results.

Use of generative AI tools

DeepSeek V4 assisted code development and debugging. ChatGPT 5.6 assisted translation, language editing and formatting. The authors remain responsible for all analytical decisions, results, references, code outputs and the final manuscript. No AI system is listed as an author.

Results

Structural representations show high technical scoreability but limited phenotype-balanced coverage

The strict cohort contained 137 genes; exclusion of five confirmed incompatible cases left 132 genes not known to be incompatible. Structural screening identified 42 structure-ready genes within the 128-gene interaction-qualified intersection. MutPred-PPI scored 10,917/11,150 submitted rows (97.91%) from 10,499 unique scoring units. Failures comprised 182 missing structural residues, 27 reference mismatches and 24 chain problems. This high scoreability applied to submitted structural inputs; gene-level eligibility also depended on representation coverage and phenotype balance. On the common 132-gene denominator, nine genes reached C4 ≥ 0.70 and three satisfied phenotype-balance criteria: *FGFR2*, *MEN1* and *TUBA1A*. Structural binary and distance features reached the coverage threshold in 49 and nine genes, respectively, with ten and three balanced genes (Figs. 1 and 2).

Sequence-based partner representations substantially expand usable coverage

SPPIDER-seq reached C4 ≥ 0.70 in 127 genes, with 23 balanced; X-PAIR reached the threshold in 107 genes, with 15 balanced (Fig. 2). Of 15,942 planned gene-partner pairs per model, 15,471 were scored by SPPIDER-seq and 13,094 by X-PAIR. X-PAIR had 2,377 length-unsupported pairs; each model also had 471 sequence-availability failures. Among 1,109,989 common-denominator variant-partner rows, SPPIDER-seq represented 1,089,797 (98.18%) and X-PAIR 877,919 (79.09%). These counts describe availability under each model's input requirements and representation definition. Sequence-based representations substantially reduced the structural availability bottleneck. Their phenotype information was assessed separately in the subsequent labeled-versus-sorted test.

Internal experimental-contact validation gave X-PAIR macro AUROC 0.739 (95% CI 0.707–0.769) and AUPRC 0.453 (0.394–0.509) across 67 query proteins and 51,928 residue-pair units. Corresponding SPPIDER-seq estimates were 0.560 (0.531–0.589) and 0.304 (0.254–0.359), across 80 proteins and 63,687 units. Overlap between the evaluation data and training data was unknown for both models. Median within-gene partner-track Spearman summaries were approximately 0.745 and 0.926, respectively. These analyses assessed internal contact validity and variation among partner tracks (Figure S1; Table S7).

Named partner identity provides limited global incremental phenotype information

The shared balanced cohort contained 15 genes and 1,034 unique pathogenic residues: 680 single-disease, 39 exclusive cross-disease and 315 multilabel/ambiguous residues. Ten genes met the unique-residue eligibility criteria, contributing 547 WT-residue analysis units and a median of 17 common partners (Fig. 3; Table 2). *FGFR2*, *MFN2*, *PTEN*, *RHO* and *RPGR* did not meet the unique-residue group-size requirement. Multiple substitutions at one WT residue shared a single sequence-interface representation and were not counted as additional analysis units.

Table 2
Primary phenotype-information results

Gene	Primary residues	Group sizes	Partners	X-PAIR labeled R^2	Sorted R^2	ΔR^2	qΔ	SPPIDER ΔR^2
COL3A1	108	97 / 11	29	0.042	0.051	−0.0087	0.978	0.0039
COL4A1	33	17 / 16	21	0.016	0.009	0.0073	0.436	−0.0007
COL4A3	122	112 / 10	6	0.020	0.020	0.0005	0.687	$+7.7\times10^{-7}$
GUCY2D	24	11 / 13	10	0.063	0.085	−0.0223	0.978	0.0012
KCNQ2	77	17 / 60	16	0.001	0.001	$+2.4\times10^{-5}$	0.687	-2.5×10^{-5}
LHCGR	20	10 / 10	6	0.013	0.010	0.0026	0.687	0.0006
MEN1	46	29 / 17	167	0.009	0.008	0.0016	0.687	$+3.1\times10^{-5}$
MMUT	54	11 / 43	18	0.008	0.006	0.0020	0.687	0.0008
RS1	31	14 / 17	4	0.013	0.001	0.0125	0.039	−0.0180
TUBA1A	32	11 / 21	554	0.010	0.006	0.0045	0.687	−0.0033

Disease groups follow prespecified lexicographic MONDO order within each gene and do not denote shared diseases across genes. ΔR^2=labeled−sorted. qΔ denotes the Benjamini-Hochberg false discovery rate-adjusted P value for the ΔR^2 contrast. R^2 is shown to three decimals, ΔR^2 to four and qΔ to three significant digits. Signed scientific notation preserves tiny nonzero ΔR^2 values. Exact values and MONDO mappings are provided in the supplementary workbook. Partner dimensionality varies across genes; labeled and sorted profiles are dimension-matched within each gene. *RS1* was the only X-PAIR gene with qΔ < 0.05. Its labeled association was not significant, and its direction was not reproduced by SPPIDER-seq.

Median X-PAIR labeled R^2 was 0.0130652 and median sorted R^2 was 0.0084828. The median paired ΔR^2 was 0.00179553 (interquartile range 0.00014122–0.00403078; 95% gene-bootstrap CI − 0.00410526 to 0.00495717). ΔR^2 was positive in eight of ten genes; the secondary one-sided sign test *P* was 0.0546875. The confidence-interval component of the prespecified criterion was not met, whereas the positive-gene fraction component was met (80%); the combined criterion was not satisfied (Fig. 4).

No X-PAIR labeled association met its FDR threshold. *RS1* had Δ-contrast $q = 0.0389961$, but its overall labeled association was not significant and its ΔR^2 direction was not reproduced by SPPIDER-seq. This

gene-specific contrast was evaluated separately from the global criterion and lacked allele-specific biological validation. Table 2 includes all ten genes, irrespective of ΔR^2 direction or significance. The paired-null example in Figure S2 displays the alphabetically first gene.

SPPIDER-seq median ΔR^2 was 0.0000158735, with six positive genes. Across models, Spearman correlation was − 0.745455 and Pearson correlation − 0.582; four of ten genes were positive in both. For *COL3A1*, SPPIDER yielded labeled q = 0.0469953 and Δ q = 0.0409959, while X-PAIR yielded ΔR^2=−0.00870360. The gene-specific findings therefore differed between representations. Restricted and unrestricted nominal Δ statuses agreed in all ten genes; *RS1* restricted *P* was 0.00239976 (Figure S3). All three evaluable correlation-distance contrasts, for *KCNQ2*, *MEN1* and *TUBA1A*, were negative; the other seven were unevaluable. Raw-score and alternative-sort sensitivities remained secondary. The eight-gene SPPIDER extension had seven evaluable genes and 702 residues; five ΔR^2 values were positive, with no corrected labeled or Δ association. It remained separate from the primary analysis. Length, full coverage, partner-replacement and resource summaries are retained in Figures S4-S7.

Complete-case comparisons evaluated position in ten genes, generic interface features in five, generic 3D context in four and domain proxies in two. X-PAIR showed no consistent superiority over position. The domain proxy used prespecified UniProt annotations rather than a complete Pfam/InterPro resource. Incomplete annotations and the small domain-proxy subset prevented an overall model ranking. Each descriptive marginal contrast used the rows shared with its comparator (Figure S8). These comparisons did not estimate conditional independence or clinical prediction performance.

WT residue-level representation granularity limits substitution-specific information

Distinct substitutions at the same WT residue received the same residue × partner vector (Fig. 3). Such vectors cannot intrinsically encode substitution-specific phenotype differences at that residue. Broader-registry cross-disease-linked proportions were 69/146 (47.26%) for *KCNQ2*, 38/84 (45.24%) for *MEN1*, 23/54 (42.59%) for *RS1* and 32/64 (50.00%) for *TUBA1A*. These denominators include annotated residues beyond the primary single-disease cohort. This design limits the information available to distinguish substitutions that share a WT residue. Its contribution to the observed global result was not directly tested, and these summaries were not used to normalize ΔR^2.

Discussion

Sequence-based partner representations substantially expanded phenotype-balanced coverage and enabled a within-gene test of phenotype information. SPPIDER-seq and X-PAIR supported balanced representation in 23 and 15 genes, respectively, compared with three for MutPred-PPI on the common denominator. The primary analysis of ten genes and 547 WT-residue analysis units yielded a small positive median X-PAIR ΔR^2, with a 95% CI crossing zero. Eight of ten genes had positive ΔR^2 values, but the combined global criterion was not met, and cross-model findings limited generalization. The results distinguish improved access to molecular representations from evidence that named partner

assignments help distinguish phenotypes. Small or heterogeneous contributions remain possible within the evaluated cohort and representation definitions. Experimental edgetic studies support the biological importance of selectively altered interactions [5, 6]. Our evaluation asks how much phenotype information the tested computational features retain. Expanded coverage and limited global incremental information can coexist, and their interpretation depends on the representation being evaluated. Gene-specific findings remained secondary to the global criterion, with cross-model direction and multiplicity correction retained in their interpretation.

The evaluation follows an evidence hierarchy from technical scoreability through interface validity and partner specificity to phenotype information. Phenotype-aware models such as V2P and PheMART assess variant consequences using phenotypic and molecular inputs [2, 3]. Structural network reconstruction, dSysMap and PIONEER connect variants with interfaces and interaction context [4, 7, 8]. MutPred-PPI, SPPIDER-seq and X-PAIR encode different aspects of interaction perturbation or partner-conditioned interfaces [9, 10, 11]. Accurate interface localization answers a different question from identifying the phenotype produced by a pathogenic allele. Our labeled-versus-sorted control evaluates the information contributed by assigning interface scores to named partners. The sorted control was constructed from the same raw score multiset while removing named partner assignments, before independent standardization. Both profiles retain the same WT-residue analysis units and dimensionality. We analyzed genes separately and gave them equal weight. Counting WT residues as analysis units avoided counting substitutions with identical representations at the same residue as separate observations. Independent standardization makes ΔR^2 an operational contrast between feature geometries, not a covariate-adjusted causal partner effect. Separate denominators for submitted structural rows, planned interaction pairs and WT-residue analysis units distinguish technical success from cohort usability. Prespecified eligibility and support criteria allow partner-identity information to be evaluated alongside conventional interface metrics. Internal contact validation and partner-track variation address intermediate levels of the hierarchy, while phenotype information remains a separate endpoint.

A central representation issue is the mismatch between WT residue × partner features and substitution-specific molecular consequences. The evaluated sequence models assign identical partner vectors to different substitutions at one position [10, 11]. Such features cannot intrinsically distinguish those substitutions, even when their biological effects differ. Relevant differences could include affinity changes, interaction gain or loss, dominant-negative effects, altered stability, conformational changes or biochemical activity. Their contribution to the current result was not tested. MutPred-PPI and studies of mutation mechanisms address variant-specific structural and interaction consequences [9, 16]. Livesey and Marsh show that agreement between variant-effect predictors and multiplexed assays depends on the endpoints and mechanisms captured [17]. FuncVEP illustrates how functional evidence can improve missense variant-effect prediction, while saturation genome editing provides allele-specific evidence for interpreting *BRCA1* variants [18, 19]. Reviews of multiplexed assays describe their opportunities and the calibration needed for variant interpretation [20]. These studies support representations that describe both the substitution and its molecular context. Partner-aware features can help prioritize interaction

experiments and functional assays. A predicted interface alone, however, leaves the consequence of a particular amino-acid change unresolved. Future phenotype-resolved models should test the incremental information from combining perturbation features with partner identity against matched generic features. Independent cohorts are needed to assess whether that contribution generalizes. This approach evaluates substitution-specific information while retaining partner context as molecular evidence.

The precision and scope of this benchmark depend on cohort size, annotation quality and the available representations. Only ten genes met the phenotype-ready criteria, and unequal disease-group sizes produced heterogeneous within-gene precision. Aggregated ClinVar assertions and MONDO harmonization may yield coarser disease groups than expert-curated phenotype definitions [12, 13]. Larger, independently curated cohorts and more detailed phenotype descriptions could improve estimation of small or heterogeneous contrasts [2, 3]. The largely static partner catalog and canonical protein coordinates may omit interactions or isoforms relevant to particular tissues and cellular states. Cell-specific interactome remodeling, tissue-specific isoform effects and tissue-resolved protein associations provide complementary evidence for evaluating these contexts [21, 22, 23]. Interaction resources also incompletely cover disease-relevant partners [15, 24]. The effects of these omissions on the global result remain untested. Unknown overlap with model training data could make internal interface-validation estimates optimistic; transparent training identities and strictly held-out evaluations remain necessary [25, 26]. Sequence redundancy can further inflate internal benchmark performance, as illustrated by Handshake [27]. These concerns do not establish phenotype-label leakage in the current study. Incomplete generic annotations prevented uniform comparisons across genes, and the matched comparisons showed no consistent superiority over position. Sensitivity analyses and generic-feature comparisons retained their prespecified secondary role. Complete annotations and matched denominators would support more uniform comparisons. Gene-specific contrasts also lack allele-specific experimental evidence for a molecular mechanism. Interaction assays, deep mutational scanning and calibrated multiplexed assays provide appropriate routes for follow-up [6, 20, 28]. Larger samples can improve precision, while substitution-aware features and disease-relevant context can evaluate information absent from the current WT representation. Allele-specific validation would address biological consequences beyond those captured by this interface-based benchmark.

Conclusions

This benchmark evaluates technical availability, phenotype-balanced coverage and incremental phenotype information as distinct aspects of human allelic heterogeneity. Sequence-based partner representations substantially expanded coverage. Named partner identity provided no robust global incremental phenotype information in the evaluated cohort under the prespecified combined criterion. WT residue-level partner vectors cannot intrinsically distinguish substitutions at the same residue. The contribution of this limitation to the global result remains untested. Future phenotype-resolved models should combine substitution-specific perturbation, partner context and disease-relevant biological

context, then assess their incremental contribution in independently curated cohorts. The framework distinguishes access to molecular features from evidence relevant to phenotype discrimination.

Abbreviations

AUROC
area under the receiver operating characteristic curve
AUPRC
area under the precision-recall curve
BH
Benjamini–Hochberg
CI
confidence interval
FDR
false discovery rate
MONDO
Mondo Disease Ontology
PERMANOVA
permutational multivariate analysis of variance
PPI
protein–protein interaction
RCV
reference ClinVar
WT
wild type
XML
extensible markup language
ΔR^2
labeled-minus-sorted R^2 contrast.

Declarations

Ethics approval and consent to participate

This study used secondary, publicly available, non-identifiable data. The ethics committee of the Second Affiliated Hospital of Wenzhou Medical University determined that the study was exempt from ethics review; no ethics approval number was assigned. No new participant recruitment or collection of identifiable personal data occurred, and individual consent was not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Funding

This research received no specific funding.

Author Contribution

Yu Zhang: Conceptualization, methodology, software, code management, data curation, formal analysis, visualization, writing – original draft. Da Lin: Data curation, investigation, validation, writing – review and editing. Ying Chen: Investigation, validation, visualization, writing – review and editing. Yue Liu: Data curation, investigation, validation, writing – review and editing. Yi Zha: Validation, visualization, writing – review and editing. Wangqiang Feng: Conceptualization, supervision, project administration, resources, writing – review and editing.

Acknowledgements

Not applicable.

Data Availability

he public datasets and annotation resources used in this study are described in the Methods and Additional file 1. Analysis code is available on GitHub (https://github.com/seefreewind/allelic-heterogeneity-partner-representations). Derived datasets, frozen analysis outputs, source manifests, permutation and bootstrap results, and figure source data are deposited in Zenodo (version 1.0.0; https://doi.org/10.5281/zenodo.23435673) [29]. Additional files 1 and 2 provide supplementary methods, figures and full-precision tables. Data release dates, software and model versions, and file hashes are documented in Additional file 1 and the archive manifests. Third-party model weights and datasets must be obtained from their original providers under the applicable licences; full encoder caches are not included in the archive.

References

1. Lazareva TE, Barbitoff YA, Nasykhova YA, et al. Statistical Dissection of the Genetic Determinants of Phenotypic Heterogeneity in Genes with Multiple Associated Rare Diseases. Genes. 2023;14:2100. https://doi.org/10.3390/genes14112100.

2. Stein D, Kars ME, Milisavljevic B, et al. Expanding the utility of variant effect predictions with phenotype-specific models. Nat Commun. 2025;16:11113. https://doi.org/10.1038/s41467-025-66607-w.
3. Wen J, Zeng S, Bonzel CL et al. April. Phenotypic prediction of missense variants via deep contrastive learning. Nature Biomedical Engineering. Published online 14 2026. https://doi.org/10.1038/s41551-026-01636-4
4. Wang X, Wei X, Thijssen B, et al. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol. 2012;30:159–64. https://doi.org/10.1038/nbt.2106.
5. Zhong Q, Simonis N, Li QR, et al. Edgetic perturbation models of human inherited disorders. Mol Syst Biol. 2009;5:321. https://doi.org/10.1038/msb.2009.80.
6. Sahni N, Yi S, Taipale M, et al. Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. Cell. 2015;161:647–60. https://doi.org/10.1016/j.cell.2015.04.013.
7. Mosca R, Tenorio-Laranga J, Olivella R, et al. dSysMap: exploring the edgetic role of disease mutations. Nat Methods. 2015;12:167–8. https://doi.org/10.1038/nmeth.3289.
8. Xiong D, Qiu Y, Zhao J, et al. A structurally informed human protein–protein interactome reveals proteome-wide perturbations caused by disease mutations. Nat Biotechnol. 2025;43:1510–24. https://doi.org/10.1038/s41587-024-02428-4.
9. Stewart R, Laval F, Coppin G et al. Predicting interaction-specific protein–protein interaction perturbations by missense variants with MutPred-PPI. In: Proceedings of the 30th Annual International Conference on Research in Computational Molecular Biology (RECOMB); 2026. https://doi.org/10.64898/2025.12.20.695738
10. Porollo A, Jadhav O, Alvarez A, et al. SPPIDER-seq: sequence-based partner-aware predictor of protein–protein interaction sites. Bioinformatics. 2026;42:btag388. https://doi.org/10.1093/bioinformatics/btag388.
11. Rescalli S, Carbone A. X-PAIR: an ultrafast multitask framework for proteome-scale reconstruction of PPI networks and partner-specific interfaces from sequence. bioRxiv (preprint). 2026. https://doi.org/10.64898/2026.07.20.739596.
12. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46:D1062–7. https://doi.org/10.1093/nar/gkx1153.
13. Vasilevsky NA, Toro S, Matentzoglu N, et al. Mondo: integrating disease terminology across communities. Genetics. 2026;232:iyaf215. https://doi.org/10.1093/genetics/iyaf215.
14. Stark C, Breitkreutz BJ, Reguly T, et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34:D535–9. https://doi.org/10.1093/nar/gkj109.
15. Orchard S, Ammari M, Aranda B, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014;42:D358–63. https://doi.org/10.1093/nar/gkt1115.

16. Gerasimavicius L, Livesey BJ, Marsh JA. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. Nat Commun. 2022;13:3895. https://doi.org/10.1038/s41467-022-31686-6.
17. Livesey BJ, Marsh JA. Why variant effect predictors and multiplexed assays agree and disagree. Nat Commun. 2026;17:10295. https://doi.org/10.1038/s41467-026-77211-x.
18. Kayaalp B, Çil K, Conil C, et al. Prediction of human missense variant effects from functional evidence. Nat Genet. 2026;58:2350–60. https://doi.org/10.1038/s41588-026-02727-3.
19. Findlay GM, Daza RM, Martin B, et al. Accurate classification of BRCA1 variants with saturation genome editing. Nature. 2018;562:217–22. https://doi.org/10.1038/s41586-018-0461-z.
20. McEwen AE, Tejura M, Fayer S, et al. Multiplexed assays of variant effect for clinical variant interpretation. Nat Rev Genet. 2026;27:137–54. https://doi.org/10.1038/s41576-025-00870-x.
21. Huttlin EL, Bruckner RJ, Navarrete-Perea J, et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell. 2021;184:3022–e304028. https://doi.org/10.1016/j.cell.2021.04.011.
22. Weykopf G, Badonyi M, Friman ET, et al. Disease-associated genetic variants can cause missense effects in tissue-specific protein isoforms. Nat Commun. 2026;17:7627. https://doi.org/10.1038/s41467-026-74280-w.
23. Laman Trip DS, van Oostrum M, Memon D, et al. A tissue-specific atlas of protein–protein associations enables prioritization of candidate disease genes. Nat Biotechnol. 2026;44:654–67. https://doi.org/10.1038/s41587-025-02659-z.
24. Luck K, Kim DK, Lambourne L, et al. A reference map of the human binary protein interactome. Nature. 2020;580:402–8. https://doi.org/10.1038/s41586-020-2188-x.
25. Walsh I, Fishman D, Garcia-Gasulla D, et al. DOME: recommendations for supervised machine learning validation in biology. Nat Methods. 2021;18:1122–7. https://doi.org/10.1038/s41592-021-01205-4.
26. Grimm DG, Azencott CA, Aicheler F, et al. The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. Hum Mutat. 2015;36:513–23. https://doi.org/10.1002/humu.22768.
27. Haspel N, Handshake. preprint) : Partner-Specific Protein–Protein Binding Site Prediction at Scale Using ProstT5 and Cross-Chain Attention. bioRxiv (2026. https://doi.org/10.64898/2026.06.04.730144
28. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014;11:801–7. https://doi.org/10.1038/nmeth.3027.
29. Partner representation audit. for allelic heterogeneity [software and derived-data archive]. Version 1.0.0. Zenodo; 2026. https://doi.org/10.5281/zenodo.23435673.

Figures

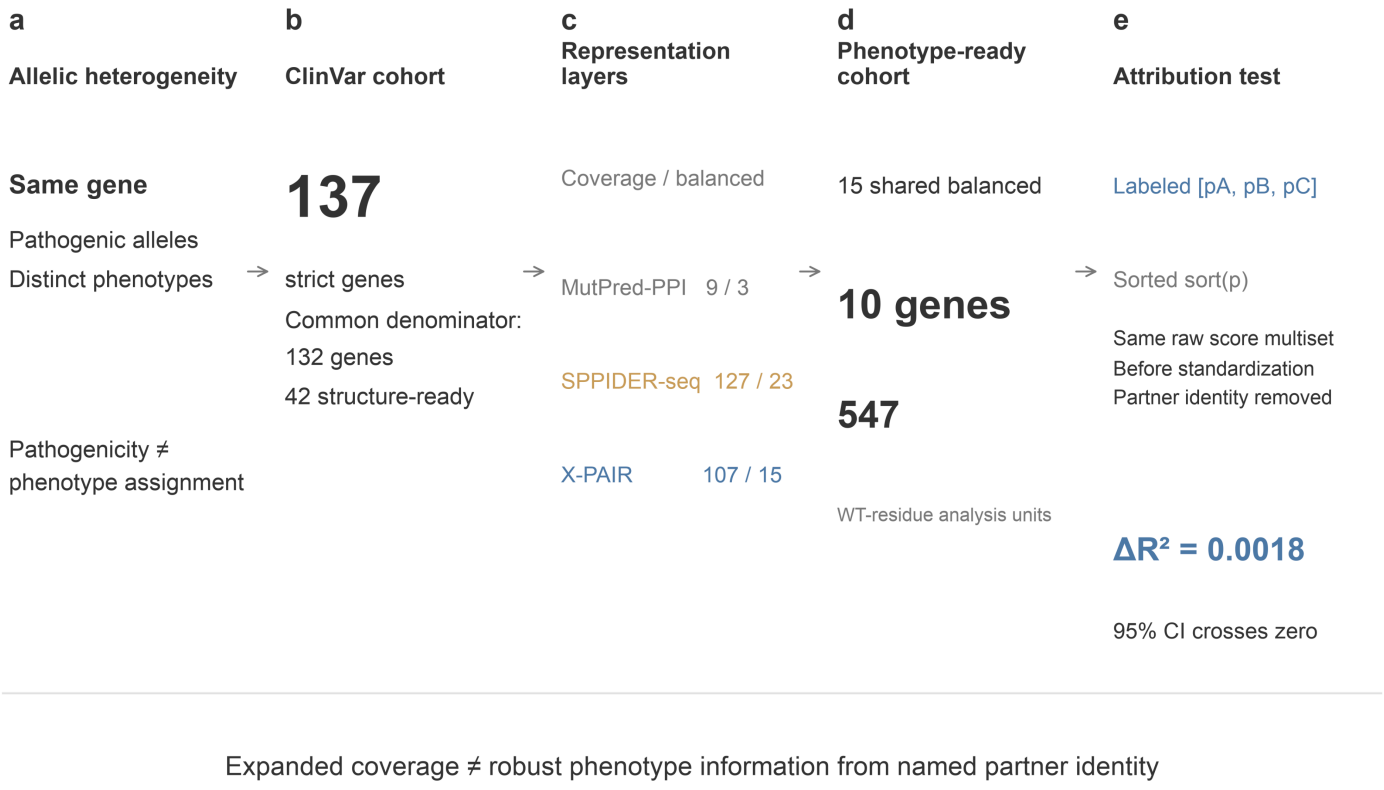


Figure 1

Evaluation proceeds from representation availability to phenotype information. a, Pathogenic alleles in one gene can produce distinct phenotypes. b, The strict cohort comprised 137 genes; exclusions left a common denominator of 132 genes and 42 structure-ready genes within the 128-gene interaction-qualified intersection. c, Coverage / phenotype-balanced counts are 9/3 for MutPred-PPI, 127/23 for SPPIDER-seq and 107/15 for X-PAIR on the common 132-gene denominator. d, Fifteen shared balanced genes yielded ten eligible genes and 547 WT-residue analysis units. e, The sorted control is constructed from the same raw score multiset while removing named partner assignments, before independent standardization. Dimensionality is matched within each gene. Median ΔR^2 was 0.00179553, with a 95% CI crossing zero. Arrows indicate evaluation order; expanded availability alone does not establish phenotype information.

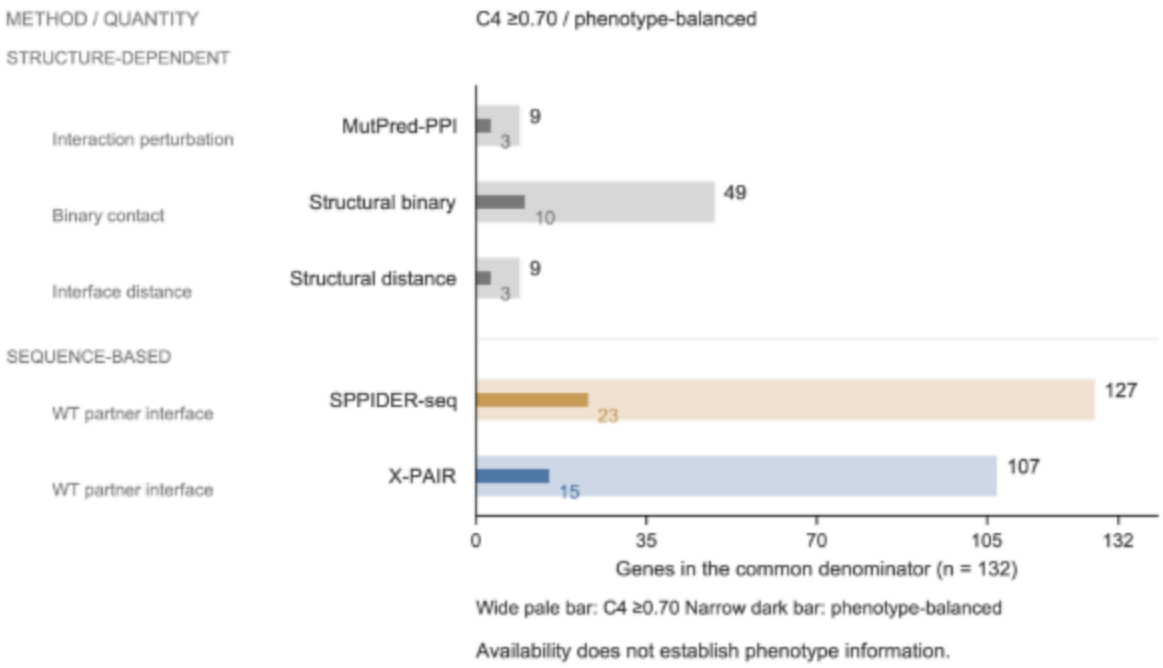


Figure 2

Common-denominator coverage separates broad availability from phenotype balance. Wide pale bars show genes with C4≥0.70; narrow dark bars show phenotype-balanced genes among the same 132 genes. Coverage / balanced counts are 9/3, 49/10, 9/3, 127/23 and 107/15 for MutPred-PPI, structural binary, structural distance, SPPIDER-seq and X-PAIR, respectively. C4 is the fraction of variants with ≥3 valid credible partners. Counts are descriptive and have no inferential intervals. The methods encode different molecular quantities; availability does not establish phenotype information or a ranking of phenotype discrimination.

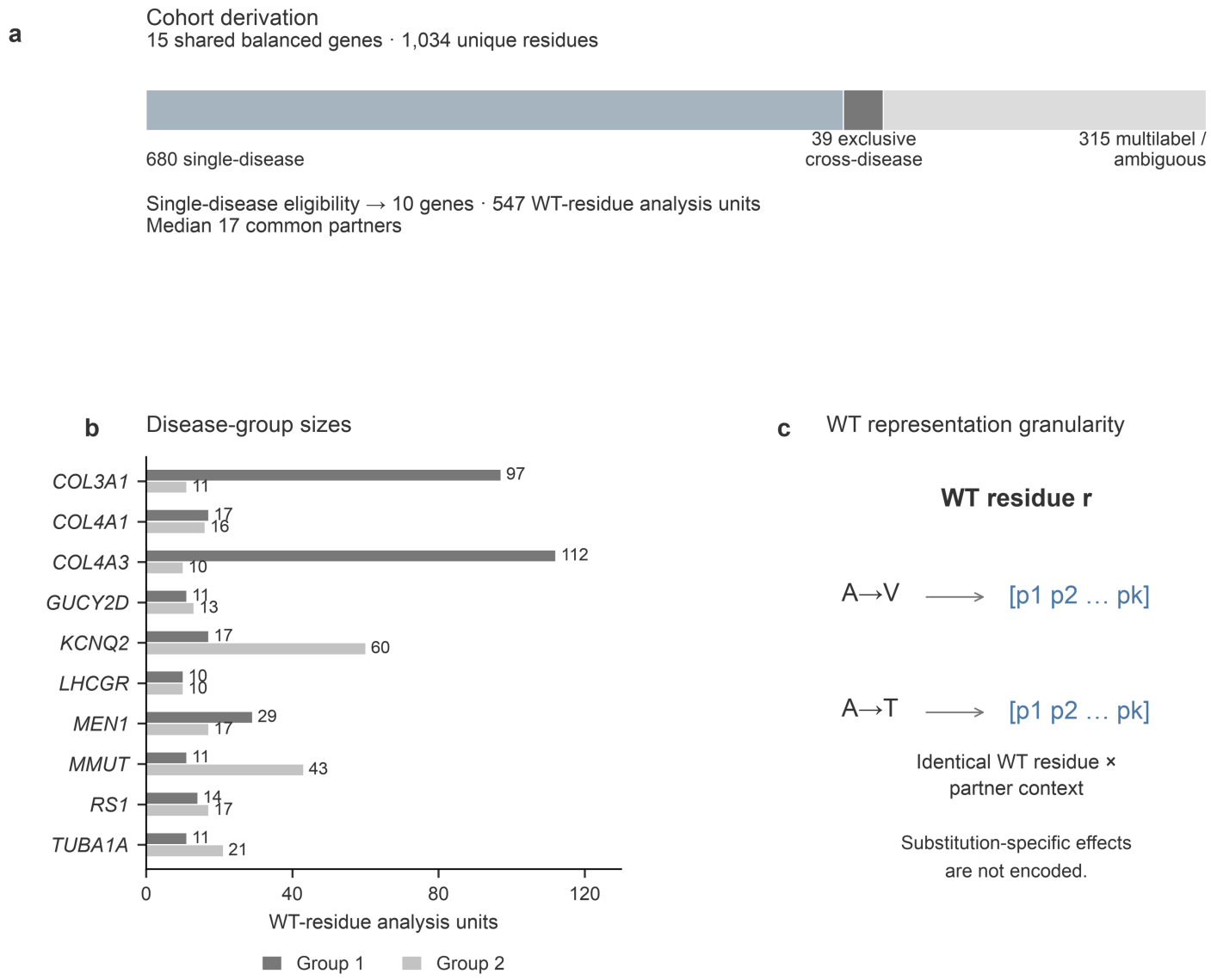


Figure 3

Cohort eligibility and representation granularity define the phenotype-information test. a, Fifteen shared balanced genes contained 1,034 unique residues: 680 single-disease, 39 exclusive cross-disease and 315 multilabel/ambiguous. Single-disease unique-residue eligibility yielded ten genes, 547 WT-residue analysis units and a median of 17 common partners. b, Paired disease-group sizes retain the prespecified gene order; groups follow lexicographic MONDO order within each gene and do not denote the same diseases across genes. c, Illustrative A→V and A→T substitutions at one WT residue receive the same residue × partner vector; $p_1,\ldots,p_k$ denote WT partner scores. Counts are descriptive. This representation property was not tested as the cause of the global result.

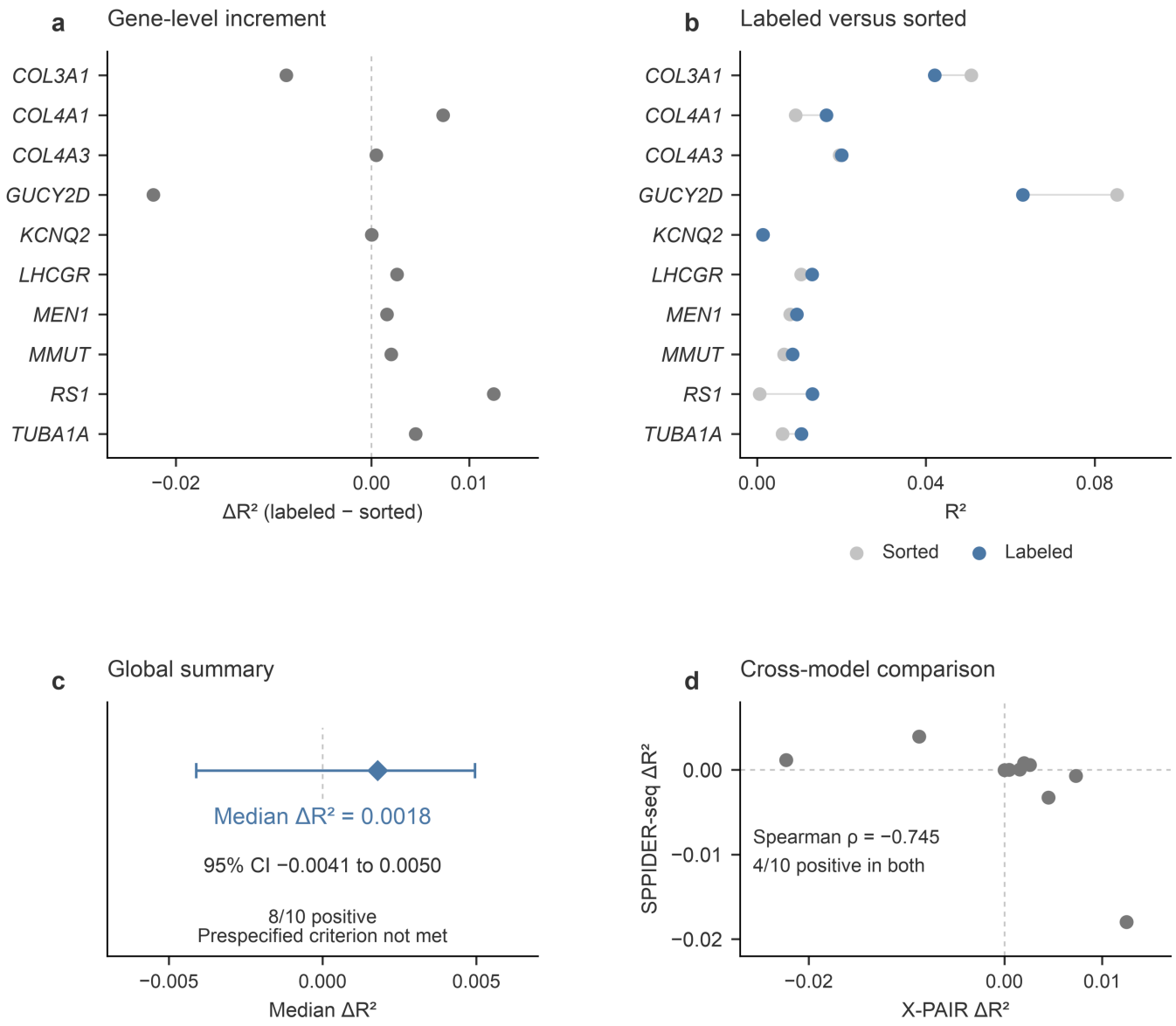


Figure 4

Named partner identity did not show robust global incremental phenotype information. a, X-PAIR gene-level ΔR² estimates retain all ten primary genes. b, Labeled and sorted R² are paired on the same WT-residue analysis units, with dimensionality matched within each gene. c, The equal-weight gene median ΔR²=labeled−sorted was 0.00179553 (95% percentile CI −0.00410526 to 0.00495717 from 10,000 saved gene-bootstrap resamples); 8/10 genes were positive. The prespecified combined criterion required a positive median ΔR² with its 95% CI strictly above zero and positive ΔR² in ≥70% of phenotype-ready genes and was not met. d, Paired model-specific ΔR² estimates had Spearman correlation −0.745455, Pearson correlation −0.582 and 4/10 positive ΔR² values in both models. Dashed lines mark zero. Gene-level panels show estimates without gene-specific intervals. Cross-model comparison is descriptive and confirmatory; the operational contrast does not identify a causal partner.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile2.xlsx