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.

Intrinsic sequence evidence cannot reach a pathogenic classification for missense variants under a partitioned ACMG framework

Anees Ahmed Mahaboob Ali
Vellore Institute of Technology
Radhakrishnan Delhibabu
Vellore Institute of Technology
Everette Jacob Remington Nelson
everette.nelson@vit.ac.in

Vellore Institute of Technology

Research Article

Keywords: variant interpretation, ACMG/AMP criteria, variants of uncertain significance, evidence partition, missense classification, probability calibration, differential diagnosis, inherited bleeding and platelet disorders, ClinGen, reproducibility

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10770799/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.

Missense variants in rare-disease genes are predominantly returned as variants of uncertain significance. Under the American College of Medical Genetics and Genomics / Association for Molecular Pathology framework, the criteria a laboratory can apply to a missense variant from its sequence and annotations alone are few and individually weak, and every criterion capable of closing the remaining distance to a pathogenic classification requires functional, segregational, case-control or phenotypic evidence. The criterion encoding phenotypic specificity presupposes a model of disease that a classifier reasoning solely over sequence lacks. How large that shortfall is has not been measured within a framework that refuses to count any item of evidence more than once.

Results.

DISCERN couples a calibrated classification engine to a likelihood-ratio model that discriminates ten clusters of confusable diseases, within a joint posterior in which each criterion is assigned to exactly one factor. Evaluated in inherited bleeding and platelet disorders on the expert-panel surface of the ClinGen Evidence Repository, blinded to ClinVar, it reached an area under the receiver operating characteristic curve of 0.939 (95% confidence interval: 0.906–0.967) on missense variants and 0.927 under a time-split, statistically indistinguishable from REVEL, the predictor supplying its own computational criterion. Out-of-sample expected calibration error was 0.017. The evidence partition achieved complete criterion coverage across 38,802 expert-applied codes and prevented the reuse of evidence that would inflate 33.2% of variants in a coupled model. A treatment-divergence interlock flagged every contraindicated management plan and none of the harmless ones. On intrinsic sequence evidence alone, the framework assigned a pathogenic classification to 0 of 425 missense variants, reaching at most 3.0 of the 6 points required; restoring every criterion that a partitioned pipeline withholds or cannot annotate recovers only 89 variants.

Conclusions.

The ceiling follows from the point arithmetic of the guidelines, not from any implementation, and holds wherever they are applied. Reducing the uncertain fraction requires evidence of a different kind, not a better predictor.

Background

Variant of uncertain significance (VUS) is a characteristic output of diagnostic sequencing in rare diseases [1], which is concentrated among missense variants: on the expert panel surface used throughout this work, 830 of the 974 uncertain classifications sit on missense variants. The reason is arithmetic rather than technological. Under the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) framework [2], the criteria a laboratory can apply to a missense variant,

based solely on the sequence and its annotations, are few and individually weak: absence from population databases and a computational prediction of deleteriousness. Every criterion capable of closing the remaining distance to a pathogenic classification requires evidence that is by construction not in the sequence, namely a functional assay, segregation in a family, case-control counts, or a phenotype specific to one disease. The last of these, PP4, presupposes a model of disease, which a classifier that reasons over sequence alone does not contain. A better predictor, therefore, cannot solve the problem because the shortfall lies in the kind of evidence available rather than in the quality of the prediction. What has been missing is a measurement of how large that shortfall actually is, made under a framework that refuses to count any item of evidence twice. We study this question in inherited bleeding and platelet disorders, a domain chosen because misclassification changes treatment and the discriminating assays are well characterised. These conditions are individually rare but collectively important causes of morbidity, and their diagnosis is made difficult by a structural feature of the phenotype: distinct molecular diseases produce overlapping clinical pictures because they disrupt shared hemostatic pathways. Mucocutaneous bleeding with a normal platelet count, easy bruising, and menorrhagia are common to many of these conditions, so the clinically decisive distinctions frequently rest not on the presenting symptoms but on a specific confirmatory laboratory test. The consequences of confusing look-alike diseases are not academic. Desmopressin, a first-line agent in several bleeding disorders, can precipitate thrombocytopenia in type 2B von Willebrand disease (VWD) by releasing hyperadhesive von Willebrand factor (VWF), and is therefore generally contraindicated for that subtype [3]. Type 2B is, however, readily mistaken for platelet-type VWD, in which the genetic lesion lies in a different gene (GP1BA rather than VWF); platelet-type disease accounts for roughly 15% of cases labelled type 2B [4]. Glanzmann thrombasthenia (GT) and leukocyte adhesion deficiency type 3 (LAD3) share a defect in integrin function but differ completely in management, with the latter requiring hematopoietic stem cell transplantation, the only definitive treatment option [5]. An inherited form of macrothrombocytopenia, such as Bernard-Soulier syndrome (BSS), is often misdiagnosed as immune thrombocytopenia, exposing patients to corticosteroids, immunoglobulin, and, in reported cases, splenectomy [6, 7]. Each of these is a treatment-changing error that turns on disease identity.

Compounding the diagnostic problem is the burden of genetic uncertainty. A large fraction of the variants discovered in the genes underlying these disorders are classified as VUS, leaving clinicians with a molecular finding they cannot act on. On the ClinGen Evidence Repository (eRepo), the bleeding panel surface used here, 974 of 2,239 expert-classified single-nucleotide variants (43.5%) carry an uncertain classification. The scale of this problem and how much of that the gene-specific criteria can remove are illustrated by the ClinGen Platelet Disorder Variant Curation Expert Panel (PD-VCEP). When gene-specific ACMG/AMP criteria were applied to a pilot set of ITGA2B/ITGB3 variants, the proportion classified as uncertain fell from 29% to 20% [8]. Because loss-of-function alleles are largely resolved by the very strong PVS1 criterion whose strength assignment follows a published decision tree [9], residual uncertainty concentrates in missense variants, which carry the least functional and segregation evidence. The guidelines’ re-expression as a Bayesian point system has made automation both principled and practical [10, 11]. Several automated classifiers implement these rules, among them InterVar [12] and more recently BIAS-2015, benchmarked against the FDA-recognised eRepo that reported improvements over InterVar in both pathogenic and benign sensitivity [13]. In parallel, predictors such as REVEL [14] and AlphaMissense [15] supply the missense pathogenicity signal that the PP3 criterion now draws on; for REVEL, ClinGen recommends calibrated

thresholds [16]. What none of these tools can do is compute PP4 in a principled way: the criterion encoding phenotypic specificity presupposes a disease model that, as noted above, a sequence-only classifier does not contain. We are not alone in this assessment. A recent evaluation of ACMG/AMP tools concluded that the guidelines treat phenotypic evidence as merely supporting through PP4, which prevents optimal variant prioritisation, and found that tools with stronger phenotypic integration outperformed those relying primarily on genomic features [17].

The idea at the centre of DISCERNfollows directly from that gap. If a disease model is what PP4 requires, then a system that already discriminates among diseases can supply calibrated PP4 to the variant engine. The decisive observation that distinguishes two diseases (flow cytometry, ristocetin-induced platelet aggregation, or a family history of leukemia) usually provides functional PS3 evidence as well. The same information that resolves the diagnosis resolves the variant. Mechanically, this is a single joint posterior over disease and variant, $P(D, V \mid E)$, in which disease identity and variant support inform one another. This is distinct from the existing literature on probabilistic variant interpretation. Gene- and syndrome-specific Bayesian models of variant pathogenicity exist and outperform genome-wide approaches [18], but they estimate the probability that a variant is pathogenic; they do not discriminate among look-alike diseases, nor do they couple a disease posterior back onto the variant to resolve a VUS. Phenotype-driven diagnostic systems, AI-MARRVEL [19] and agentic large language model systems such as, DeepRare [20] rank candidate genes or diseases across the rare disease spectrum, but they perform diagnosis not coupled VUS resolution, with the LLM-based systems known to generate unverifiable citations [20]. On the novelty of the partition, that double-counting of evidence is undesirable is well established: the ACMG/AMP guidelines require each criterion to rest on its own evidence [2], and ClinGen Sequence Variant Interpretation guidance recommends against the reputable source criteria, PP5 and BP6, for the related reason that they reuse another laboratory's conclusion [21]. Three things are new here: the formal partition makes exclusive ownership structural rather than advisory, inside a coupled disease-variant model, which is exactly the setting that creates the temptation, because a phenotype-aware classifier is invited to re-add PP4 or PS3 on top of a bottom line that has already absorbed that evidence; the partition is empirically demonstrated against expert panel classifications, criterion by criterion, rather than argued; and its effect is quantified on a genome-wide expert surface. We also distinguish this from the PP5/ClinVar circularity problem, which BIAS-2015 itself flags [13] and which we also independently demonstrate: circularity is re-reading a published conclusion, whereas the partition prevents reusing the evidence that produced it. They are different mechanisms, and they require different guards.

This study is deliberate about the boundary between what it proves and what it pre-registers. We formalise the coupled disease-variant model and the evidence partition that guarantees each ACMG criterion enters the joint posterior exactly once. We validate the variant classification engine at scale on the FDA-recognised eRepo expert panel surface, under a ClinVar-blinded and time-split protocol, with confidence intervals (CI) and paired tests throughout. We characterise a trustworthiness layer (calibration, abstention, and a treatment-divergence safety interlock) and quantify the ceiling that intrinsic evidence imposes on missense resolution. Differential diagnosis is presented as a scoped proof-of-concept, reported against a phenotype-blind baseline and against an external phenotype-driven ranker. Finally, we make it explicit that the disease-variant coupling, the system's defining novelty, has not yet been demonstrated in real data; we explain why

public data cannot prove it, and we pre-register its definitive cohort evaluation with an explicit falsification condition.

Methods

Framework overview

DISCERN comprises three coupled components and a trustworthiness layer. The variant classification engine assigns each variant a calibrated posterior probability of pathogenicity. The disease discrimination engine assigns, within a cluster of confusable diseases, a posterior over disease identity. The coupling combines the two into a joint posterior, P(D, V | E). Surrounding these, a calibration-and-abstention module and a safety interlock convert raw model output into decisions a clinician can audit (Fig. 1). Figure S1 (Additional file 1) shows the entire pipeline for a single case, from the evidence supplied through the criterion trail to the safety stop and the recommended next test. The system is implemented in Python with a deterministic core, covered by 240 unit tests in the public repository, all of which pass in continuous integration; a further 6 tests cover the clinician-facing web application, which is kept out of the public repository. All reported numbers were produced under a single locked configuration.

Variant classification

The variant engine implements the ACMG/AMP criteria [2] within the ClinGen point-based Bayesian formulation, in which each criterion contributes signed, strength-weighted points that are mapped through an exponentially scaled odds-of-pathogenicity function to a posterior probability with community-defined classification boundaries [10, 11]. Disease-specific rule sets are encoded as versioned specifications for the genes within scope, including the PD-VCEP specification for ITGA2B/ITGB3 [8], the Coagulation Factor Deficiency VCEP specification for F8/F9, and the Myeloid Malignancy VCEP specification for RUNX1, each specified by a ClinGen [22] VCEP and drawn from the Criteria Specification Registry [23]; PVS1 strength is assigned by the published loss-of-function decision tree [9]. The PP3/BP4 missense signal is taken from REVEL at ClinGen-calibrated thresholds [14, 16], with AlphaMissense [15] available as an alternative predictor. Population frequency evidence is drawn from gnomAD [24, 25].

The evidence partition

The methodological core of DISCERN is a partition of the ACMG evidence space across the factors of the joint model. Each criterion is owned by exactly one factor, and we set out the membership explicitly because it matters for interpreting the ceiling result. The variant factor owns PVS1, PS1, PS4, PM1, PM2, PM4, PM5, PP2, PP3, PP5, BA1, BS1, BS2, BP1, BP3, BP4, BP5, BP6 and BP7; PP5 and BP6 are owned for completeness but are never applied, consistent with the recommendation against them [21] and also the ClinVar-blinded protocol described below. The disease factor owns PP4; the functional factor owns PS3 and BS3; the segregation factor owns PP1 and BS4; the phasing factor owns PM3 and BP2; and the *de novo* factor owns PS2 and PM6. PM1 (hotspot or critical domain) and PM5 (same residue, different change) are variant-

intrinsic and are not routed away. Because ownership is exclusive, a criterion cannot contribute to the joint posterior more than once. The guarantee is structural, not a tuning choice. The partition is enforced in code and verified by a unit test that adds PP4, PS3, PP1, and PM3 as explicit codes and asserts the variant marginal is unchanged to within 10^{-12}.

The joint model and its factorisation

The joint is

$$P(D, V \mid E) \propto P(E_{pheno} \mid D) \cdot P(E_{geno} \mid V) \cdot P(E_{func} \mid D, V) \cdot P(V \mid G, D) \cdot P(G \mid D) \cdot P(D)$$

where G is the gene in which the sequenced variant lies. The two gene-dependent terms, P(V | G, D) and P(G | D), are distinct with both required. P(V | G, D) is the coupling: given the disease and the variant's location, how likely the variant is to be pathogenic. P(G | D) is the location itself: a variant found in F8 is evidence for hemophilia A over hemophilia B. Because P(V | G, D) normalises over variant states, it cancels in the marginalisation to a disease posterior, so without the explicit P(G | D) term, the gene would carry no information about disease identity at all. Splitting the two completes the factorisation without duplicating evidence, with the partition test mentioned above remaining unaffected. Two constraints bound the gene term, both derived from the architecture rather than fitted to a benchmark. It must remain weaker than a cluster's deciding observation, or no laboratory result could overturn the gene, and the abstention rule's next-test recommendation would be vacuous; the committed likelihood ratio of 4 is the largest value at which the sharpest assay in the knowledge base (RIPA mixing, ratio ≈ 11) still overturns it. It must also not veto a treatment-divergent competitor: finding a variant in one gene does not exclude a disease of another gene, which may never have been sequenced, so the safety interlock is evaluated on a gene-blind posterior against the engine's own leading call. A sensitivity sweep across the full range of this parameter is shown in Figure S2 (Additional file 1); safety behaviour is invariant over this range, and the diagnostic result is flat above the committed value.

Disease discrimination

The disease discrimination engine operates over ten clusters of confusable diseases spanning 31 diseases and 29 causal genes, drawn from the International Society on Thrombosis and Haemostasis Scientific and Standardization Committee's (ISTH SSC) curated gene list for bleeding, thrombotic, and platelet disorders [26] and its ClinGen gene-disease validity assessment [27], with ITGB2 added because LAD1 belongs in the integrin-disorder differential (Table 1). The clusters are grouped here as integrin disorders (GT, LAD1 and 3, RASGRP2-related), VWD and GPIb-related (type 2B vs platelet-type; type 2N vs mild hemophilia A; mild/low VWF), macrothrombocytopenias, alpha granule and dense granule disorders, Scott syndrome, thrombocytopenia-with-leukemia-predisposition genes (RUNX1, ANKRD26, ETV6), and coagulation factor deficiencies. Within a cluster, discrimination is computed as a product of likelihood ratios over the features that separate the diseases in that cluster, each sourced to a specific primary reference (PMID) and guarded in continuous integration so that an unsupported value cannot be committed. All likelihood ratios, with their sample sizes and sources, are given in Table S1 (Additional file 2).

Table 1

The ten clusters of confusable conditions on which the framework was evaluated.

Cluster	Confusable conditions	Genes	Deciding observation	Consequence of confusion	LRs
C1	Glanzmann thrombasthenia; Leukocyte adhesion deficiency type III; RASGRP2 (CalDAG-GEFI) platelet dysfunction; Leukocyte adhesion deficiency type I	FERMT3, ITGA2B, ITGB2, ITGB3, RASGRP2	Platelet aggregometry with αIIbβ3 expression and activation; leukocytosis and delayed umbilical separation	Leukocyte adhesion deficiency types I and III need haematopoietic stem cell transplantation; Glanzmann thrombasthenia does not [5, 28]	20
C2	Type 2B von Willebrand disease; Platelet-type von Willebrand disease; Type 2A von Willebrand disease	GP1BA, VWF	Ristocetin-induced platelet aggregation (RIPA) mixing study, distinguishing plasma from platelet origin	Desmopressin is contraindicated in type 2B von Willebrand disease (VWD) and can worsen thrombocytopenia [3]	11
C3	Bernard–Soulier syndrome; MYH9-related disorder; Immune thrombocytopenia*	GP1BA, GP1BB, GP9, MYH9	Blood smear and flow cytometry for CD42 (GPIb-IX-V)	Corticosteroids and splenectomy are inappropriate for an inherited macrothrombocytopenia misread as immune thrombocytopenia (ITP) [6, 7]	12
C4	RUNX1 familial platelet disorder with myeloid malignancy; ETV6 thrombocytopenia with malignancy predisposition; ANKRD26-related thrombocytopenia; Immune (idiopathic) thrombocytopenia*	ANKRD26, ETV6, RUNX1	Germline panel, pedigree, platelet size and dysmegakaryopoiesis	Myeloid malignancy surveillance is required, and an affected relative must not be used as a transplant donor [29]	23

Each cluster groups conditions that present similarly but differ in management, together with the observation that discriminates between them. Most of the deciding observations are laboratory assays whose discriminating results are not represented in the HPO, which bears on the phenotype-driven comparison in the Results. LRs give the number of likelihood ratios the discrimination model holds for that cluster; each one carries a PubMed source identifier and a sample size, provided in full in Table S1 (Additional file 2), and a continuous integration check fails if any entry lacks either. Gene lists, condition names and likelihood-ratio counts are read from the archived cluster definitions. *Acquired condition, included because inherited thrombocytopenias are frequently misdiagnosed as it.

Cluster	Confusable conditions	Genes	Deciding observation	Consequence of confusion	LRs
C5	Von Willebrand disease type 2N; Mild/moderate haemophilia A	F8, VWF	Von Willebrand factor to factor VIII binding assay	Recombinant factor VIII monotherapy fails in type 2N VWD, which needs a VWF-containing concentrate [30]	10
C6	Chediak-Higashi syndrome; Hermansky-Pudlak syndrome; Grey platelet syndrome	AP3B1, HPS1, HPS3, HPS4, HPS5, HPS6, LYST, NBEAL2	Electron microscopy, blood smear for giant granules, and workup for haemophagocytic lymphohistiocytosis (HLH)	Chediak-Higashi syndrome carries HLH risk and is treated by transplantation [31]	10
C7	Grey platelet syndrome; GFI1B-related platelet disorder; Quebec platelet disorder; ARC syndrome	GFI1B, NBEAL2, PLAU, VPS33B	Blood smear, alpha-granule electron microscopy, and the urokinase overexpression assay	Quebec platelet disorder requires antifibrinolytics; platelet transfusion is ineffective and is contraindicated [32]	12
C8	Factor XIII deficiency; Haemophilia A; Haemophilia B	F13A1, F13B, F8, F9	Individual factor activity assays, with a normal prothrombin time and activated partial thromboplastin time in factor XIII deficiency	Recombinant factor XIII-A2 treats F13A1 deficiency but not F13B deficiency [33]	10
C9	Type 1 von Willebrand disease / low von Willebrand factor; Mild platelet function defect; no identifiable disorder	VWF	Repeat von Willebrand factor measurement, light transmission aggregometry, and a standardised bleeding assessment score	Risk of both over-treatment and under-treatment	10
C10	Scott syndrome; a normal routine haemostatic workup	ANO6	Platelet phosphatidylserine exposure and a prothrombinase-based assay	Routine coagulation and platelet studies are normal; the diagnosis is missed without a specific assay	10

Each cluster groups conditions that present similarly but differ in management, together with the observation that discriminates between them. Most of the deciding observations are laboratory assays whose discriminating results are not represented in the HPO, which bears on the phenotype-driven comparison in the Results. LRs give the number of likelihood ratios the discrimination model holds for that cluster; each one carries a PubMed source identifier and a sample size, provided in full in Table S1 (Additional file 2), and a continuous integration check fails if any entry lacks either. Gene lists, condition names and likelihood-ratio counts are read from the archived cluster definitions. *Acquired condition, included because inherited thrombocytopenias are frequently misdiagnosed as it.

The disease-variant coupling

A confirmed or strongly supported disease identity assigns calibrated PP4 to the variant; the decisive functional test that establishes the disease assigns PS3. Conversely, variant-level support sharpens the posterior for the disease. We emphasise the scope: the coupling machinery is implemented, and its internal behaviour is unit-tested, but its empirical claim (that using the disease to resolve a VUS moves the variant towards its independent ground truth, and does so better than intrinsic evidence alone) is not evaluated in this work. It is pre-registered (see Pre-registration and reproducibility section below) and deferred to paired cohort data, for reasons established in the Results section on Differential diagnosis.

Calibration, abstention, and the safety interlock

The variant posterior is calibrated by isotonic regression. Calibration is estimated strictly out-of-sample: the isotonic fit is performed inside each training fold of a 5-fold stratified split, and expected calibration error (ECE) and Brier score are accumulated only on the held-out folds, so no variant is ever scored by a calibrator that saw its label. Fold disjointedness and complete coverage are asserted in code. All calibration and discrimination metrics include 95 percentile bootstrap CIs from 1,000 resamples. An abstention rule withholds a call when the posterior is insufficiently separated from its nearest competitor, routing instead to the cheapest next observation that would resolve the case. A safety interlock issues a hard stop when a planned management action is contraindicated for a sufficiently supported disease; the interlock is evaluated for both sensitivity and specificity.

Benchmarking design and the fair-comparison protocol

Two distinct eRepo surfaces are used in this study, and because their sizes differ by an order of magnitude, we define both here and refer to them consistently throughout this study. The bleeding panel surface is the expert panel classification set restricted to the in-scope genes: 2,239 single nucleotide variants, of which 425 missense variants carry a pathogenic or benign label and form the discrimination and calibration set. Every classification, calibration, and ceiling result is measured on this surface. The genome-wide surface is the complete eRepo export: 12,240 variants across 170 genes, carrying 38,802 applied evidence codes. It is used for one purpose only: to test whether the evidence partitions generalise beyond the genes it was written for. No performance claim is made on the genome-wide surface, and no partition coverage claim is made on the bleeding panel surface alone; the two figures answer different questions and are not alternative measurements of the same quantity. The primary truth surface was the FDA-recognised eRepo expert panel classifications [34, 35] rather than raw ClinVar [36, 37], and was restricted to the in-scope gene set. Because several comparator tools ingest ClinVar through criteria such as PP5/BP6, all tools were run ClinVar-blinded, and a time-split evaluation graded only variants whose expert classification post-dated the tools' bundled ClinVar snapshot (2021-05-01), so no tool could have memorised the label. The headline comparison surface was missense, where intrinsic ACMG tools differ. Comparators were REVEL [14] and AlphaMissense [15] as continuous predictors, InterVar [12] and GeneBe [38] as ACMG classifiers, with BIAS-2015 [13] positioned by its published eRepo metrics. AlphaMissense provides no score for one variant on each surface, so its entries in Table 2 use n = 424 and n = 382, respectively. Four variant-intrinsic criteria go unapplied throughout, for reasons that differ and that we state individually: PS1 and PM5 require same-

residue ClinVar lookups and are implemented but deliberately withheld, because supplying them under a ClinVar-blinded protocol would reintroduce the circularity we demonstrate in the ClinVar circularity result; PS4 is implemented but requires case-control counts absent from the annotation source; and PM1 is not implemented, because no VCEP specification in scope encodes hotspot or critical domain regions. Only the last is an engine limitation, and we note it as such. To make the calibration comparison like-for-like, REVEL and AlphaMissense were subjected to the same folds and isotonic protocol as DISCERN and are reported both raw and calibrated. AUROC differences between DISCERN and REVEL were tested using the DeLong's method [39] and a paired bootstrap on the same variants.

Table 2
Variant classification on the eRepo expert-panel surface.

Tool	Surface	n	AUROC (95% CI)	ECE (95% CI)	Brier	DeLong p vs DISCERN	Capabilities
DISCERN	eRepo-primary	425	0.939 (0.906–0.967)	0.017 (0.011–0.047)	0.064	n/a	C A T R B
REVEL (raw)	eRepo-primary	425	0.954 (0.931–0.972)	0.092 (0.079–0.121)	0.081	0.287	B
REVEL (calibrated)	eRepo-primary	425	0.954 (0.931–0.972)	0.043 (0.030–0.073)	0.073	0.287	C B
AlphaMissense (raw)	eRepo-primary	424	0.931 (0.906–0.953)	0.167 (0.135–0.200)	0.138	n/a	B
AlphaMissense (calibrated)	eRepo-primary	424	0.931 (0.906–0.953)	0.029 (0.024–0.065)	0.094	n/a	C B
GeneBe (circularity exhibit)	eRepo-primary	425	1.000	n/a	n/a	n/a	A T
DISCERN	time-split	383	0.927 (0.882–0.961)	0.019 (0.012–0.049)	0.065	n/a	C A T R B
REVEL (raw)	time-split	383	0.951 (0.924–0.973)	0.091 (0.080–0.125)	0.081	0.154	B
REVEL (calibrated)	time-split	383	0.951 (0.924–0.973)	0.034 (0.025–0.068)	0.077	0.154	C B
AlphaMissense (raw)	time-split	382	0.929 (0.901–0.955)	0.180 (0.146–0.216)	0.144	n/a	B
AlphaMissense (calibrated)	time-split	382	0.929 (0.901–0.955)	0.031 (0.024–0.067)	0.088	n/a	C B

Discrimination and calibration for DISCERN, REVEL and AlphaMissense, the latter two both raw and calibrated under the identical out-of-fold protocol, alongside InterVar, BIAS-2015 and GeneBe. Capabilities key: C, calibrated probability as delivered; A, ACMG classification; T, auditable criterion trail; R, abstention; B, run ClinVar-blinded; None, no listed capability. AUROC and ECE are point estimates with 95% confidence intervals in parentheses. The InterVar row is a prior in-house evaluation on a ClinVar-classified set, so it is not on a like-for-like basis with the AUROC column. BIAS-2015 was not re-run here, and its published evaluation does not report the quantities in the metric columns, so the row is retained for its capability comparison only. GeneBe is shown as a circularity exhibit, not a fair comparator.

Tool	Surface	n	AUROC (95% CI)	ECE (95% CI)	Brier	DeLong p vs DISCERN	Capabilities
GeneBe (circularity exhibit)	time-split	383	1.000	n/a	n/a	n/a	A T
InterVar (prior in-house run)	prior ClinVar set (missense subset)	364	0.811	n/a	n/a	n/a	A T B
BIAS-2015 (by citation)	published eRepo metrics	n/a	n/a	n/a	n/a	n/a	A T B

Discrimination and calibration for DISCERN, REVEL and AlphaMissense, the latter two both raw and calibrated under the identical out-of-fold protocol, alongside InterVar, BIAS-2015 and GeneBe. Capabilities key: C, calibrated probability as delivered; A, ACMG classification; T, auditable criterion trail; R, abstention; B, run ClinVar-blinded; None, no listed capability. AUROC and ECE are point estimates with 95% confidence intervals in parentheses. The InterVar row is a prior in-house evaluation on a ClinVar-classified set, so it is not on a like-for-like basis with the AUROC column. BIAS-2015 was not re-run here, and its published evaluation does not report the quantities in the metric columns, so the row is retained for its capability comparison only. GeneBe is shown as a circularity exhibit, not a fair comparator.

Differential diagnosis benchmark, baselines, and independent sensitivity

Differential diagnosis was evaluated on a curated benchmark of 42 published cases spanning all ten clusters, each represented by its causal gene together with the clinical and laboratory findings extracted from the report, with every case sourced to an NCBI-verified PMID and no reproduction of article text (Table S2, Additional file 3). Because the causal gene is supplied as input, we report the result against three phenotype-blind baselines: a gene lookup that ranks the cluster by placing first the diseases whose gene list contains the case gene (a static table, no phenotype, no likelihood ratios); a prior-only ranking that ignores the gene; and a uniform random floor. Results are stratified by whether the gene maps to more than one disease within its cluster: the subset in which something other than the gene must break the tie. Paired comparison uses the exact McNemar test plus a bootstrap CI for the difference. We additionally ran LIRICAL [40, 41] v2.4.1 in phenotype-only mode, supplying both observed and negated Human Phenotype Ontology (HPO) terms, in every case carrying at least one HPO term, and report it both genome-wide and restricted to the same cluster DISCERN ranks within. Exomiser [42] was not run: it requires a VCF, and seeding one per case, with zygosity matched to each report's stated inheritance, introduces a correctness risk we judged to be larger than the comparison's marginal value, given that LIRICAL had already established the phenotype-channel result on the same inputs. We state this as a scope decision, not as a handicapped run. Independent variant-arm sensitivity was assessed against the CDC Hemophilia A and B Mutation Project datasets (CHAMP and CHBMP) [43–45], a truth surface entirely distinct from ClinGen. Constructing the gene-only baseline revealed that the gene was not reaching the disease posterior, and we added the missing P(G|D) term described in previously (The joint model and its factorisation section). Because these 42 cases exposed the defect, post-correction performance on them is in-sample. We, therefore, report the

diagnosis arm in three states: phenotype-only with the gene withheld, which makes P(G|D) inert and is invariant to the correction; pre-correction; and post-correction, labelled in-sample. Only the first two are used for comparison against an external tool.

Pre-registration and reproducibility

The disease-variant coupling endpoint, its analysis plan, and an explicit falsification condition were pre-registered before any cohort analysis [46]. All code required to reproduce the reported results is publicly available in the project repository [47]; a DOI-archived release *via* the GitHub-Zenodo integration accompanies this submission (see Availability of Data and Materials section). A supporting data record holds the derived results, configurations, and manifests required to reproduce every reported number. Third-party databases are referenced by versioned manifest rather than redistributed (Table S3, Additional file 4), and every software and database version is recorded in Table S4 (Additional file 5). Every result in this study is locked by a regression test in the public repository.

During the preparation of this manuscript, Claude (Anthropic) was used for language editing, for refining author-written text, and for coding assistance. No text or figure content was generated de novo by the tool; all AI-assisted output was applied to author-written material. The authors reviewed, verified and edited all AI-assisted output and take full responsibility for the content, accuracy and integrity of the manuscript.

Results

Variant classification on the expert panel surface

On the eRepo expert surface (2,239 single nucleotide variants; missense pathogenic/benign n = 425), DISCERN achieved an AUROC of 0.939 (95% CI 0.906–0.967), against REVEL 0.954 (0.931–0.972) and AlphaMissense 0.931 (0.906–0.953). Under the time-split (n = 383), the values were 0.927 (0.882–0.961), 0.951 (0.924–0.973), and 0.929 (0.901–0.955). Tested paired on the same variants, the difference from REVEL was not statistically significant on either surface: eRepo-primary $\Delta = -0.015$, DeLong $z = -1.06$, $p = 0.29$, paired bootstrap 95% CI − 0.044 to + 0.012; time-split $\Delta = -0.024$, $z = -1.43$, $p = 0.15$, CI − 0.060 to + 0.008. REVEL is DISCERN's own PP3 input, so the framework should track it, which it does; we make no claim to out-discriminate it (see Table 2 and Fig. 2).

Calibration, compared like-for-like

DISCERN's out-of-sample ECE was 0.017 (95% CI 0.011–0.047) on the eRepo-primary surface and 0.019 (0.012–0.049) under the time-split, with Brier scores of 0.064 and 0.065. That the calibration holds under the time-split is the strong form of the result, because it cannot be attributed to label leakage. Calibration is not, however, unique to DISCERN. Put through the identical folds and the identical isotonic protocol, the comparators calibrate too: REVEL reaches ECE 0.043 (0.030–0.073) from a raw 0.092, and AlphaMissense reaches 0.029 (0.024–0.065) from a raw 0.167. DISCERN remains best on every point estimate and on the Brier score, but the CIs overlap, and we do not assert a significant advantage. The distinction that matters is not whether a score can be calibrated, but who has to do it. Calibration is not a property of REVEL or

AlphaMissense; obtaining it requires fitting an isotonic layer to labelled expert classifications, a modelling step that those tools do not provide and that consumes the very labels a user is trying to predict. DISCERN is calibrated as delivered. More fundamentally, a calibrated score is still a score: it cannot emit an ACMG classification, an auditable criterion trail, a partition guarantee, an abstention decision, or a treatment-safety hard stop. InterVar and GeneBe emit categories that cannot be calibrated. The differentiator is the calibrated, auditable classification system, not calibration in isolation.

The partition reproduces expert reasoning without double-counting

Two claims must be kept apart here because they are supported by different evidence. That the partition assigns each criterion to exactly one factor is established by vocabulary coverage and by the inflation figure (Fig. 1B); it does not depend on per-criterion agreement at all. Per-criterion agreement supports the weaker and separate claim that the engine reproduces expert reasoning wherever it has the inputs to do so. We report all thirteen criteria encountered on this surface, grouped by mechanism rather than by value (Fig. 3; full counts, both agreement levels, and a per-row explanation in Table S5). Agreement is measured on the 1,265 variants of this surface carrying a pathogenic or benign expert classification, since a graded variant is required on both sides; every κ below reconciles with the applied counts reported alongside it at that denominator. Agreement was high for the two criteria that a sequence-and-annotation pipeline is built to derive: PVS1 (κ = 0.98) and PP3 (κ = 0.95). It was moderate for the frequency-based criteria, BA1 (κ = 0.64) and BS1 (κ = 0.30), with DISCERN being the more conservative of the two raters. It was low for PM2 (κ = 0.08) and BP4 (κ = 0.07). In both cases, the expert panels applied the criterion more often than DISCERN did: PM2 812 times vs DISCERN's 479, with 336 in common, and BP4 615 times vs DISCERN's 91. DISCERN requires an observed gnomAD frequency below the gene-specific threshold before applying PM2, and this annotation surface leaves that undetermined for many variants, so the engine withholds the criterion where the panels, working from the full curation record, applied it. Because a single κ conflates whether a criterion was applied at the same weight, we also report strength-level agreement among variants where both applied the criterion. It is 1.00 for BA1, BS1, and BP4, 0.97 for PP3, 0.94 for PM2, and 0.04 for PVS1. The PVS1 figure reflects an input limitation of this surface rather than a scoring policy: the PVS1 decision tree assigns strength from transcript geometry, namely whether the premature termination codon is predicted to trigger nonsense-mediated decay and what fraction of the protein is removed, and this annotation cache carries neither, so the tree downgrades conservatively to moderate on every loss-of-function variant here. Where those inputs are available, as in the CHAMP and CHBMP analysis below, the same tree assigns PVS1 at very strong. The two analyses use the same implementation and differ only in the evidence available to them. Seven criteria were applied zero times, for three distinct reasons that a single figure would obscure. PS3 is routed to the functional factor by the partition; of the three categories, it is the only one whose absence is itself evidence of the partition operating as specified on expert-classified data. PM1, PM5, PS1, and PS4 are variant-intrinsic and are not routed anywhere; they are absent because this pipeline has no input for them, since PM1 requires VCEP hotspot and critical domain annotation is not implemented across all in-scope genes, PM5 and PS1 require the same-residue lookups against ClinVar that the blinding protocol deliberately disables, and PS4 requires case-control counts that no public surface supplies at this scale. BP7 and BS2 are not implemented. BP7 matters here because the panels applied it 541 times: it addresses

synonymous variants, which fall outside the missense scope of every result reported here, so its absence is a scope boundary rather than a defect. The ceiling result below quantifies what restoring each category would change.

One consequence of this comparison strengthens the central result, and it does not depend on how often either side applied any criterion. Even if DISCERN applied PM2 to every variant that the expert panels did, the ceiling would remain unchanged: PM2 contributes 1 point, and PP3, at its maximum strength, contributes 2, so the intrinsic maximum remains 3.0 against the 6 required for likely pathogenic. The ceiling is a property of the point arithmetic, not of the application rate of any single criterion. Across the genome-wide partition analysis (12,240 eRepo variants, 170 genes, 38,802 applied evidence codes), partition vocabulary coverage was 100% with zero uncovered codes, and exclusive ownership prevented pathogenicity inflation in 4,068 variants, 33.2% (Wilson 95% CI 32.4–34.1%). Stated precisely: 33.2% of all eRepo variants carrying at least one applied code would be placed in a higher, more pathogenic band by a naive scorer that sums every applied criterion into a single factor, relative to the band produced by the variant-intrinsic factor alone. The comparison is against a coupled disease variant architecture in which a disease posterior computed from phenotype feeds PP4 back to the variant factor, so that the same observation can enter twice; conventional single-pass ACMG interpretation has no such feedback path, and no such risk, and the partition therefore addresses a hazard introduced by joint models rather than a defect in routine practice. The figure measures the evidence mass that a coupled model would be at risk of reusing. Among the 6,357 variants (51.9%) that carry a non-genetic owned code, the rate is 64.0%.

The ceiling that intrinsic evidence imposes on missense classification

The most consequential result in this work is a limitation. On the eRepo missense surface, using intrinsic sequence evidence alone, DISCERN reaches a maximum of 3.0 ACMG points against the 6 required for likely pathogenic, and assigns a pathogenic band to zero of 425 missense variants (Fig. 4A). Every band it resolves is benign-side, and at matched coverage, a REVEL threshold rule is at least as accurate (Table 3). This is not a defect, and it is not specific to DISCERN, but it has two separable causes, and it would be wrong to attribute it wholly to the partition. PM2 plus PP3 cannot reach six points. Of the remaining criteria that carry missense pathogenicity, PS3 and PP4 are owned by the functional and disease factors and are deliberately withheld from the intrinsic surface, whereas PM1, PM5, PS1 and PS4 are variant-intrinsic and go unapplied only because this evaluation supplies no input for them: no hotspot or domain annotation, no same-residue ClinVar lookup under the ClinVar-blinded protocol, and no case-control counts. We measured the split by adding to each variant's own points the criteria the expert panel applied that DISCERN did not. Restoring the intrinsic criteria that this pipeline cannot derive raises 27 of 425 to likely pathogenic; restoring the partition-routed criteria raises it to 52; restoring both raises it to 89. Neither cause alone explains the ceiling, and even together they account for only 89 of the 316 variants the panel called pathogenic (Fig. 4B; Table S6). These are lower bounds, because the arithmetic uses default code strengths and expert panels frequently apply strengths above default. The binding constraint is therefore what the ACMG framework requires as evidence for a missense variant, not any single routing decision. A REVEL threshold rule makes four times as many calls because an evidence framework does not bind it; at matched coverage, its calls

are at least as accurate as ours. Both facts stand, and the correct reading of 0 of 425 is not that the classifier never calls a missense variant; it is that, on sequence evidence alone, the guidelines do not permit a pathogenic call for this class, which is why the intended output is a calibrated abstention naming the specific next test rather than a forced band. We regard this as the strongest available argument for the coupling. The evidence the intrinsic surface lacks is exactly the evidence a disease model supplies. Of the 830 missense variants the expert panel left uncertain, DISCERN assigns a band to 188 and a ClinGen-threshold REVEL rule to 518, but neither number is an accuracy measure because eRepo VUS carry no independent truth. Closing that gap is what the pre-registered endpoint is for.

Table 3
Decision quality at matched coverage.

Decision rule	Coverage	Accuracy on resolved	Pathogenic recall	Benign recall	n resolved	Pathogenic calls	Benign calls
DISCERN ACMG class (its own operating point)	0.188	0.938	0.000	0.688	80	0	80
DISCERN calibrated probability, coverage-matched	0.162	0.913	0.098	0.294	69	33	36
REVEL at ClinGen thresholds	0.807	0.956	0.851	0.541	343	277	66
REVEL, coverage-matched to DISCERN	0.186	0.975	0.123	0.349	79	39	40

Coverage, accuracy on retained variants, and pathogenic and benign recall for the DISCERN ACMG classification, the DISCERN calibrated probability at matched coverage, and REVEL at ClinGen-recommended thresholds and at matched coverage (n = 425 missense). No risk–coverage claim is made (the diagnosis arm is at ceiling at n = 42); this table reports coverage-matched decision quality only. Every band DISCERN resolves on this surface is benign (refer Table S6 and Fig. 4).

ClinVar circularity

A live-ClinVar classifier (GeneBe) scored a perfect AUROC of 1.0, matching the ClinVar direction on every gradeable variant, and continued to score 1.0 even under the time-split. Because a tool that queries current ClinVar always sees the present label, this is reproduction, not prediction, and it constitutes a direct demonstration that ClinVar-derived truth cannot fairly grade a ClinVar-consuming tool (Fig. 5). We therefore treat GeneBe as a circularity exhibit rather than a fair comparator.

Safety and confident errors

The treatment-divergence safety interlock fired with 100% sensitivity and 100% specificity across five treatment-divergent scenarios (desmopressin/type 2B VWD, splenectomy/BSS, related-donor transplant/RUNX1, platelet transfusion/Quebec, recombinant FVIII monotherapy/type 2N VWD), including the case in which the contraindicated disease occupies the leading position, and remained silent on a harmless alternative in all five (Figure S3; Table S7). Safety behaviour is invariant to the gene-term sensitivity sweep of the joint model, as reported in that section. No high-confidence errors were observed in the diagnosis evaluation (zero calls made at confidence ≥ 0.8 that were wrong). The diagnosis arm ranks the correct disease first in all 42 cases, including the 7% on which it declines to commit, so abstention has no errors to catch here and a risk-coverage curve is uninformative at n = 42. Diagnosis-posterior calibration was ECE 0.118 with mean confidence 0.882 against accuracy 1.00, that is, mildly under-confident (Figure S4). The abstention machinery is implemented and unit-tested; its operational value requires a benchmark that is not at the ceiling, which means a cohort.

Differential diagnosis, its baselines, and the coupling negative

On the curated 42-case benchmark spanning all ten clusters, DISCERN reached Top-1 100%, Top-3 100%, and abstention 7%. A phenotype-blind gene lookup (a static table with no findings or likelihood ratios) achieved 93% Top-1 on the same cases. The difference is + 7% (95% CI 0 to + 17%, exact McNemar p = 0.25), which is not significant. A prior-only ranking reached 21% and a uniform random floor 32%. The prior-only ranking sits below the random floor because it commits to the most prevalent disease in each cluster, whereas published case reports are skewed towards the rarer entities, and a fixed wrong answer scores worse than chance. Either way, it is the gene, not the phenotype channel, that carries the pooled result (Fig. 6A; Table 4). Stratifying makes this explicit, and the three strata partition the benchmark exactly. In 36 of 42 cases, the gene maps to exactly one disease within its cluster, and both DISCERN and the lookup are correct in all 36 cases. In 3 cases, the case carries no causal gene at all, so the lookup has nothing to key on (DISCERN 3/3, lookup 2/3). The informative subset (cases where a gene is present but maps to more than one disease in its cluster, so something other than the gene must break the tie) is n = 3, with DISCERN at 3/3 and the lookup at 1/3. 36 + 3 + 3 = 42. Three cases cannot support a diagnostic accuracy claim, and we do not make one either. In publicly curated cases, where the causal gene is supplied and phenotypic depth is minimal, the gene largely determines the disease. Building this baseline also revealed a defect in the model. As noted under earlier, the gene did not reach the disease posterior at all; eight cases failed for that reason, and the benchmark scored 81%. Adding the missing P(G|D) term corrected it, at a strength fixed by an architectural constraint and swept in full. Because these 42 cases exposed the defect, post-correction DISCERN has seen them and LIRICAL has not, so the post-correction number is in-sample and is not a head-to-head result. Two arms are free of that problem. The phenotype-only arm withholds the gene entirely, which also renders P(G|D) inert, making it provably identical before and after the correction; it is the only arm whose inputs match what LIRICAL receives. The pre-correction arm reproduces the engine as it stood before these cases informed it.

Table 4
Differential diagnosis against baselines, stratified by whether the gene determines the disease.

Method	n	Top-1	Top-3	Top-1 by stratum (unique / shared / no gene)	McNemar p vs gene lookup	In-sample
DISCERN post-correction	42	1.000	1.000	1.000 / 1.000 / 1.000	0.250	Yes
DISCERN pre-correction	42	0.810	1.000	0.778 / 1.000 / 1.000	n/a	No
Gene lookup (phenotype-blind)	42	0.929	1.000	1.000 / 0.333 / 0.667	n/a	No
Prior only (gene-blind)	42	0.214	n/a	0.167 / 0.333 / 0.667	n/a	No
Uniform random within cluster	42	0.323	0.899	n/a	n/a	No
DISCERN phenotype-only (no gene), vs LIRICAL subset	23	0.913	1.000	n/a	n/a	No
DISCERN HPO-representable only (no gene), vs LIRICAL subset	23	0.739	1.000	n/a	n/a	No
LIRICAL restricted to the same cluster	23	0.565	0.957	n/a	n/a	No

Top-1 and Top-3 accuracy for DISCERN and its baselines on the curated benchmark (n = 42), stratified into cases where the gene maps to one disease within its cluster (n = 36), cases carrying no causal gene (n = 3), and the informative subset where a gene is present but maps to more than one disease (n = 3). The stratum column gives Top-1 accuracy for the unique-gene, shared-gene and no-gene strata in that order. The exact McNemar test against the phenotype-blind lookup is reported for the post-correction arm; the remaining arms are not paired against the lookup. Strata partition the benchmark exactly: unique-gene 36 + shared-gene 3 + no-gene 3 = 42. The post-correction row is in-sample and must not be quoted against an external tool.

LIRICAL 2.4.1 ran in phenotype-only mode with negated terms supplied and could accept 23 of the 42 cases; the other 19 carry no HPO term at all. Restricted to the same cluster, the only fair ranking surface, it achieved Recall@1 57%, Recall@3 96%, and MRR 0.754. Withholding the gene equalises one input but not the others, so we separate the two claims. Given strictly the same evidence (the 13 findings HPO can express, gene withheld), DISCERN achieved Recall@1 74% (MRR 0.862): a lead of + 17 percentage points whose 95% CI spans zero (-4% to + 43%; McNemar 6 against 2, p = 0.29). We, therefore, do not claim that DISCERN reasons better than LIRICAL on identical evidence. Allowed additionally to read the 35 findings HPO cannot encode (among them the RIPA mixing study, the urokinase overexpression assay, and the prothrombinase assay), DISCERN reached 91% (MRR 0.957), a lead of + 35% (CI + 13% to + 57%; McNemar 9 against 1, p = 0.02). That result is significant, and is an architectural claim about what the model can represent, which is the same motivating negative as the coverage finding rather than a demonstration of better inference (Fig. 6B; all arms and paired tests in Table S8). On these same 23 cases, the pre-correction

arm reached 83%/100%/0.913 and the post-correction arm 100%/100%/1.000, the latter in-sample and quoted only for completeness. Genome-wide and without the gene, LIRICAL reached Recall@1 of 13%, which is not a contest: it ranks roughly 8,600 diseases, while DISCERN ranks within one of three to eight clusters. One ordering anomaly needs explaining: the phenotype-only arm (91%, no gene) scores above the pre-correction arm (83%, with the gene), so supplying information appears to have hurt. That is the expected signature of the defect itself. Before P(G|D) existed, the gene could not inform the disease posterior, yet it still entered the coupling term where an off-gene disease is penalised, so a supplied gene perturbed the ranking without informing it. Withholding it removed the perturbation. The anomaly is evidence that the marginalisation defect was real rather than that its correction was tuned.

We searched for a genuinely held-out set and report that public data does not provide one. The GA4GH Phenopacket Store [48], a case corpus in the GA4GH Phenopacket schema [49], has a bleeding subset of 31 cases, of which 12 fall within a modelled gene, and 8 of those are F8 thrombophilia (elevated factor VIII, a different disease from hemophilia A that this cluster does not model), leaving 4 usable cases. Curating a held-out benchmark of newly published cases is the appropriate next step, but it is not something the present corpus supports. LIRICAL trails here not because it is a weak ranker, but because only 13 of the benchmark's 48 discriminating features have any HPO representation. The other 35 are the discriminating laboratory results, including the RIPA mixing study that separates plasma from platelet-derived components, the urokinase overexpression assay, and the prothrombinase assay. The HPO [50] encodes many assay-level abnormalities, including impaired ristocetin-induced aggregation, but not the discriminating results at this resolution (Fig. 6C). Median HPO terms per case is 1. Sparse phenotype, therefore, does not handicap every tool equally: the sparsity is systematic, and it removes the discriminating channel entirely while leaving the gene, which we have just shown is most of the answer. We report the coupling itself as a negative in public data. A circularity-safe proof-of-concept on the Phenopacket Store bleeding subset yielded only two usable cases (binary lift 0.0; continuous lift + 0.25, matched mean P(pathogenic or likely pathogenic) 0.334 against mismatched 0.082, directionally consistent but statistically powerless). The binary endpoint, which is the pre-registered one, is 0.0. Neither number supports a claim at n = 2. This is the same corpus thinness that constrains the public diagnosis comparison, arriving from a third direction, and it is precisely why the coupling requires paired cohort data in which each patient carries a phenotype, a candidate variant, and an independent ground-truth resolution simultaneously.

Independent sensitivity on CHAMP and CHBMP

Independent sensitivity on the CHAMP/CHBMP datasets was 91.2% on null variants (5,437 disease alleles, 2,130 null), with F8 at 97.7% and F9 at 65.8%; the lower F9 figure is explained by last-exon nonsense-mediated-decay escape [9] rather than by classifier error (Figure S5; Table S9).

Discussion

This work establishes, on open data, a variant-interpretation engine that matches the strongest available metapredictors on an FDA-recognised expert surface while adding properties that those tools do not attempt: an ACMG classification with an auditable criterion trail, an evidence partition that provably prevents double-counting and whose effect is quantified, a calibrated probability that survives a time-split, and a

treatment-safety interlock. The accompanying differential diagnosis engine and the independent CHAMP/CHBMP sensitivity benchmark show that the surrounding architecture behaves as intended. The coupling that gives the system its name is formalised and implemented, and is pre-registered for the evaluation that open data cannot support. The single most important empirical result here is negative and structural: under a correctly partitioned ACMG framework, intrinsic sequence evidence cannot assign a pathogenic band to a missense variant. Zero of 425, and that is not a shortcoming of this implementation; it is what the guidelines specify once one refuses to double-count, and it explains why VUS accumulate in exactly the class of variants where they are most burdensome. Every existing sequence-only classifier faces the same ceiling; tools that appear not to are generally either reading a ClinVar conclusion or thresholding a predictor score outside the evidence framework. The escape from that ceiling is an additional evidence stream, and the cheapest one available in a treatment-divergent domain is the disease itself. That is the basis of the coupling, and this study's contribution is to have measured the gap it must fill instead of assuming it. The mechanism underlying that ceiling is general; only its quantification is specific to this domain. The arithmetic does not reference any gene: the criteria obtainable from the sequence and its annotations alone cannot, in combination, reach the threshold for likely pathogenic, and every criterion that could close the gap is defined in terms of functional, segregational, case-control, or phenotypic observations. Any gene panel, in any disease area, applying the same guidelines under the same prohibition on double-counting will encounter the same wall. What will differ between panels is exactly where it falls, because that depends on how much of the annotation is available in practice, which criteria a given expert panel has specified, and how missense variants are distributed with respect to those specifications. The numbers reported here, zero of 425 and 89 recoverable of 316, are therefore a measurement of this domain and not a constant. Reproducing the measurement elsewhere requires one run of the released code against another panel's expert-classified variants; we would regard such replication, whether it agrees with these values or not, as the most useful immediate extension of this work.

For a diagnostic laboratory, the practical consequence is direct. A missense variant classified from sequence evidence alone will, in the great majority of cases, be reported as uncertain, and adopting a better *in silico* predictor will not change that, because the predictor relies on a single supporting criterion whose strength is capped by the framework itself. Reducing the uncertain fraction requires acquiring a different kind of evidence: a functional assay, segregation data, case-control counts, or a phenotype specific enough to carry PP4. The result quantifies how much of that evidence is needed, and the coupling described here is one route to supplying part of it from information a clinical service already holds. These contributions sit in a well-defined relationship to existing work. A comparable automated ACMG classifier, BIAS-2015, was published based on an eRepo benchmark and improvements over InterVar [13]; DISCERN's variant arm matches that surface and adds calibration, the quantified partition, and the safety layer, none of which the prior tool reports. The independent finding that ACMG treats phenotype as merely supporting PP4, limiting prioritisation, and that phenotype-integrating tools outperform sequence-only ones [17], is direct external support for the coupling theory. Gene- and syndrome-specific Bayesian variant models [18] validate the per-gene approach but do not discriminate diseases or couple the disease posterior back to the variant. Phenotype-driven diagnostic systems such as AI-MARRVEL [19] and DeepRare [20] address the breadth of rare-disease diagnosis and are validated on large patient cohorts, but they perform diagnosis rather than coupled VUS resolution; the agentic LLM systems have additionally produced unverifiable citations [20],

whereas DISCERN's value proposition in a treatment-divergent domain is precisely determinism, calibration, auditability, and a guaranteed safety stop.

Several limitations bound the present claims. First, we do not claim unique calibration. Post hoc isotonic calibration substantially closes the gap for REVEL and AlphaMissense, and our advantage over the calibrated comparators is not statistically significant (per the calibration results). Second, we do not claim superior discrimination. The AUROC difference from REVEL is not significant, and REVEL is our own input. Third, the diagnosis arm cannot carry an accuracy claim. A phenotype-blind gene lookup reaches 93% on the curated benchmark; the informative subset comprises three cases, and the arm therefore rests on abstention behaviour, the absence of confident errors, and the safety interlock. The P(G|D) correction is reported with its pre-correction score, so that a reader can see exactly what changed. Fourth, the comparison against rule-based ACMG classifiers is not made on a shared surface: InterVar is anchored to our own prior re-run on a ClinVar-classified set (missense n = 364), and BIAS-2015 to its published metrics, because dbNSFP and Nirvana data dependencies exceeded the available disk space, and the released BIAS-2015 pre-processing did not run in our environment. We claim competitiveness with published ACMG classifiers, not superiority over them on a shared surface. Fifth, and most importantly, the disease-variant coupling has not been proven in real data. It cannot be proven on the public corpus for structural reasons in the differential diagnosis outcome, and it cannot be proven on a single-disease cohort because the evidence that establishes disease identity is the same measurement that supports the variant, leaving no independent axis to test the coupling. Its evaluation is pre-registered with a falsification condition and depends on access to paired, multi-disease phenotype–genotype cohorts. Taken together, the formalism is the correct one, the variant interpretation engine and its auditability guarantees are established on an independent expert surface, and the quantitative ceiling we measure on intrinsic evidence is the clearest argument yet that the disease-variant coupling is the right next step. We pre-register its definitive cohort evaluation instead of asserting it.

Conclusions

Under a correctly partitioned application of the ACMG/AMP guidelines, intrinsic sequence evidence cannot reach a pathogenic classification for a missense variant. This is a property of the point arithmetic rather than of any implementation, so it holds in any gene and any disease area; what is specific to a domain is only where the ceiling falls, which is what we quantify here. For a diagnostic laboratory, the consequence is that adopting a stronger *in silico* predictor will not shrink the uncertain fraction, because the predictor feeds a single supporting criterion whose weight the framework caps. What shrinks it is evidence of another kind: a functional assay, segregation data, case-control counts, or a phenotype specific enough to carry PP4. The framework presented here supplies part of that evidence from information a clinical service already holds, and its variant interpretation engine, calibration, evidence partition and treatment-safety interlock are established on an independent expert surface. The coupling that motivates the architecture is implemented and pre-registered but is not claimed: its evaluation awaits paired phenotype-genotype cohort data, against an endpoint and a falsification condition fixed in advance. The measurement that matters most here is of a gap, and the contribution is to have made that gap precise, auditable and testable.

Abbreviations

ACMG
American College of Medical Genetics and Genomics
AMP
Association for Molecular Pathology
AUROC
area under the receiver operating characteristic curve
CHAMP
CDC Hemophilia A Mutation Project
CHBMP
CDC Hemophilia B Mutation Project
CI
confidence interval
ECE
expected calibration error
eRepo
ClinGen Evidence Repository
GA4GH
Global Alliance for Genomics and Health
HPO
Human Phenotype Ontology
ITP
immune thrombocytopenia
MRR
mean reciprocal rank
OSF
Open Science Framework
PMID
PubMed identifier
RIPA
ristocetin-induced platelet aggregation
VCEP
Variant Curation Expert Panel
VUS
variant of uncertain significance
VWD
von Willebrand disease
VWF
von Willebrand factor.

Declarations

Ethics approval and consent to participate

Not applicable. This study used only publicly available, de-identified variant data and published case reports. No human participants, biological samples, or patient-level records were involved.

Consent for publication

Not applicable. This manuscript contains no data from any person.

Availability of data and materials

The datasets generated and analysed in the current study are available in the Zenodo repository at [doi.org/10.5281/zenodo.21968627]. The deposit contains the benchmark outputs, configurations, data manifests, and the curated case set, that is, the inputs and measured outputs on which the conclusions rely. The pre-registered analysis plan for the disease-variant coupling, including its falsification condition, is available at the Open Science Framework [46]. Third-party databases are referenced *via* a versioned manifest with a source URL, version, snapshot date, and checksum, and are not redistributed. No patient-level data are included. Curated published cases are distributed as PubMed identifiers with extracted HPO terms; no article text is reproduced.

The software described in this article is available as follows:

- Project name: DISCERN
- Project home page: https://github.com/ahmedanees-m/discern
- Archived version: https://doi.org/10.5281/zenodo.21968627
- Operating system(s): platform independent
- Programming language: Python
- Other requirements: Python 3.11 or higher
- License: MIT
- Any restrictions to use by non-academics: none

Competing interests

The authors declare that they have no competing interests.

Funding

This work was financially supported by Vellore Institute of Technology, Vellore under the Faculty Seed Grant (RGEMS) (Sanction Order No. SG20250050).

Authors' contributions

AAMA (CRediT): Conceptualisation; Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Visualisation; Writing – original draft; Writing – review and editing.

RD (CRediT): Software; Investigation; Validation; Writing – review and editing.

EJRN (CRediT): Project administration; Supervision; Resources; Validation; Writing – review and editing.

All authors read and approved the final version of the manuscript.

Acknowledgements

The authors gratefully acknowledge Vellore Institute of Technology, Vellore, India, for access to computational resources and for the institutional support under which this study was carried out. This study used publicly available data from the ClinGen Evidence Repository, ClinVar, gnomAD, the Human Phenotype Ontology, the GA4GH Phenopacket Store, and the CDC Hemophilia A and B Mutation Projects; we thank the expert panels, submitters, and curators whose work these resources represent.

References

1. Fowler DM, Rehm HL. Will variants of uncertain significance still exist in 2030? Am J Hum Genet. 2024;111(1):5–10. 10.1016/j.ajhg.2023.11.005.
2. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405–24. 10.1038/gim.2015.30.
3. Connell NT, Flood VH, Brignardello-Petersen R, Abdul-Kadir R, Arapshian A, Couper S, et al. ASH ISTH NHF WFH 2021 guidelines on the management of von Willebrand disease. Blood Adv. 2021;5(1):301–25. 10.1182/bloodadvances.2020003264.
4. Othman M. Platelet-type von Willebrand disease: a rare, often misdiagnosed and underdiagnosed bleeding disorder. Semin Thromb Hemost. 2011;37(5):464–9. 10.1055/s-0031-1281030.
5. Stepensky PY, Wolach B, Gavrieli R, Rousso S, Ben Ami T, Goldman V, et al. Leukocyte adhesion deficiency type III: clinical features and treatment with stem cell transplantation. J Pediatr Hematol Oncol. 2015;37(4):264–8. 10.1097/MPH.0000000000000228.
6. Balduini CL, Savoia A, Seri M. Inherited thrombocytopenias frequently diagnosed in adults. J Thromb Haemost. 2013;11(6):1006–19. 10.1111/jth.12196.
7. Young G, Luban N, White JG. Giant platelet disorders in African-American children misdiagnosed as idiopathic thrombocytopenic purpura. J Pediatr Hematol Oncol. 1999;21(3):231–6. 10.1097/00043426-199905000-00012.

8. Ross JE, Zhang BM, Lee K, Mohan S, Branchford BR, Bray P, et al. Specifications of the variant curation guidelines for ITGA2B/ITGB3: ClinGen Platelet Disorder Variant Curation Panel. Blood Adv. 2021;5(2):414–31. 10.1182/bloodadvances.2020003712.
9. Abou Tayoun AN, Pesaran T, DiStefano MT, Oza A, Rehm HL, Biesecker LG, et al. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. Hum Mutat. 2018;39(11):1517–24. 10.1002/humu.23626.
10. Tavtigian SV, Greenblatt MS, Harrison SM, Nussbaum RL, Prabhu SA, Boucher KM, et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet Med. 2018;20(9):1054–60. 10.1038/gim.2017.210.
11. Tavtigian SV, Harrison SM, Boucher KM, Biesecker LG. Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. Hum Mutat. 2020;41(10):1734–7. 10.1002/humu.24088.
12. Li Q, Wang K. InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. Am J Hum Genet. 2017;100(2):267–80. 10.1016/j.ajhg.2017.01.004.
13. Eisenhart C, Brickey R, Nadon B, Mewton J, Bayat V, Automating. Genome Med. 2025;17(1):148. 10.1186/s13073-025-01581-y. ACMG variant classifications with BIAS-2015 v2.1.1: algorithm analysis and benchmark against the FDA-approved eRepo dataset.
14. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99(4):877–85. 10.1016/j.ajhg.2016.08.016.
15. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381(6664):eadg7492. 10.1126/science.adg7492.
16. Pejaver V, Byrne AB, Feng BJ, Pagel KA, Mooney SD, Karchin R, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am J Hum Genet. 2022;109(12):2163–77. 10.1016/j.ajhg.2022.10.013.
17. Ghasemnejad T, Liang Y, Jahanian KH, Eidi M, Salmaninejad A, Abedini SS, et al. Comprehensive evaluation of ACMG/AMP-based variant classification tools. Bioinformatics. 2026;42(2):btaf623. 10.1093/bioinformatics/btaf623.
18. Walsh R, Mazzarotto F, Whiffin N, Buchan R, Midwinter W, Wilk A, et al. Quantitative approaches to variant classification increase the yield and precision of genetic testing in Mendelian diseases: the case of hypertrophic cardiomyopathy. Genome Med. 2019;11(1):5. 10.1186/s13073-019-0616-z.
19. Mao D, Liu C, Wang L, Ai-Ouran R, Deisseroth C, Pasupuleti S, et al. AI-MARRVEL: a knowledge-driven AI system for diagnosing Mendelian disorders. NEJM AI. 2024;1(5):AIoa2300009. 10.1056/AIoa2300009.
20. Zhao W, Wu C, Fan Y, Qiu P, Zhang X, Sun Y, et al. An agentic system for rare disease diagnosis with traceable reasoning. Nature. 2026;651(8106):775–84. 10.1038/s41586-025-10097-9.
21. Biesecker LG, Harrison SM, ClinGen Sequence Variant Interpretation Working Group. The ACMG/AMP reputable source criteria for the interpretation of sequence variants. Genet Med. 2018;20(12):1687–8. 10.1038/gim.2018.42.

22. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen: the Clinical Genome Resource. N Engl J Med. 2015;372(23):2235–42. 10.1056/NEJMsr1406261.
23. ClinGen Criteria Specification Registry. https://cspec.genome.network. Accessed 8 Aug 2026.
24. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature. 2024;625(7993):92–100. 10.1038/s41586-023-06045-0.
25. Genome Aggregation Database (gnomAD). https://gnomad.broadinstitute.org. Accessed 8 Aug 2026.
26. Megy K, Downes K, Simeoni I, Bury L, Morales J, Mapeta R, et al. Curated disease-causing genes for bleeding, thrombotic, and platelet disorders: Communication from the SSC of the ISTH. J Thromb Haemost. 2019;17(8):1253–60. 10.1111/jth.14479.
27. Ross JE, Mohan S, Zhang J, Sullivan MJ, Bury L, Lee K, et al. Evaluating the clinical validity of genes related to hemostasis and thrombosis using the Clinical Genome Resource gene curation framework. J Thromb Haemost. 2024;22(3):645–65. 10.1016/j.jtha.2023.11.011.
28. Almarza Novoa E, Kasbekar S, Thrasher AJ, Kohn DB, Sevilla J, Nguyen T, et al. Leukocyte adhesion deficiency-I: a comprehensive review of all published cases. J Allergy Clin Immunol Pract. 2018;6(4):1418–e142010. 10.1016/j.jaip.2017.12.008.
29. Godley LA, Shimamura A. Genetic predisposition to hematologic malignancies: management and surveillance. Blood. 2017;130(4):424–32. 10.1182/blood-2017-02-735290.
30. Seidizadeh O, Peyvandi F, Mannucci PM. Von Willebrand disease type 2N: an update. J Thromb Haemost. 2021;19(4):909–16. 10.1111/jth.15247.
31. Kaplan J, De Domenico I, Ward DM. Chediak-Higashi syndrome. Curr Opin Hematol. 2008;15(1):22–9. 10.1097/MOH.0b013e3282f2bcce.
32. Blavignac J, Bunimov N, Rivard GE, Hayward CP. Quebec platelet disorder: update on pathogenesis, diagnosis, and treatment. Semin Thromb Hemost. 2011;37(6):713–20. 10.1055/s-0031-1291382.
33. Muszbek L, Katona É. Diagnosis and management of congenital and acquired FXIII deficiencies. Semin Thromb Hemost. 2016;42(4):429–39. 10.1055/s-0036-1572326.
34. ClinGen E. Repository. https://erepo.clinicalgenome.org. Accessed 8 Aug 2026.
35. US Food and Drug Administration. FDA recognition of public human genetic variant databases. https://www.fda.gov/medical-devices/precision-medicine/fda-recognition-public-human-genetic-variant-databases. Accessed 12 Aug 2026.
36. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062–7. 10.1093/nar/gkx1153.
37. ClinVar. https://www.ncbi.nlm.nih.gov/clinvar. Accessed 8 Aug 2026.
38. Stawiński P, Płoski R. Genebe.net: implementation and validation of an automatic ACMG variant pathogenicity criteria assignment. Clin Genet. 2024;106(2):119–26. 10.1111/cge.14516.
39. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45.

10.2307/2531595.

40. Robinson PN, Ravanmehr V, Jacobsen JOB, Danis D, Zhang XA, Carmody LC, et al. Interpretable clinical genomics with a likelihood ratio paradigm. Am J Hum Genet. 2020;107(3):403–17. 10.1016/j.ajhg.2020.06.021.
41. LIRICAL. likelihood ratio interpretation of clinical abnormalities. https://github.com/TheJacksonLaboratory/LIRICAL. Accessed 8 Aug 2026.
42. Smedley D, Jacobsen JO, Jäger M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. 2015;10(12):2004–15. 10.1038/nprot.2015.124.
43. Payne AB, Miller CH, Kelly FM, Soucie JM, Hooper WC. The CDC Hemophilia A Mutation Project (CHAMP) mutation list: a new online resource. Hum Mutat. 2013;34(2):E2382–91. 10.1002/humu.22247.
44. Li T, Miller CH, Payne AB, Hooper WC. The CDC Hemophilia B mutation project mutation list: a new online resource. Mol Genet Genomic Med. 2013;1(4):238–45. 10.1002/mgg3.30.
45. CDC Hemophilia Mutation Projects (CHAMP and CHBMP). Centers for Disease Control and Prevention. https://www.cdc.gov/hemophilia/mutation-project/index.html. Accessed 12 Aug 2026.
46. Mahaboob Ali AA. DISCERN: pre-registration of the disease-variant coupling endpoint. OSF Registries. 2026. https://doi.org/10.17605/OSF.IO/5MQCV. Accessed 11 Aug 2026.
47. DISCERN source code repository. GitHub. https://github.com/ahmedanees-m/discern. Accessed 8 Aug 2026.
48. GA4GH Phenopacket Store. https://github.com/monarch-initiative/phenopacket-store. Accessed 8 Aug 2026.
49. Jacobsen JOB, Baudis M, Baynam GS, Beckmann JS, Beltran S, Buske OJ, et al. The GA4GH Phenopacket schema defines a computable representation of clinical data. Nat Biotechnol. 2022;40(6):817–20. 10.1038/s41587-022-01357-4.
50. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Res. 2021;49(D1):D1207–17. 10.1093/nar/gkaa1043.

Figures

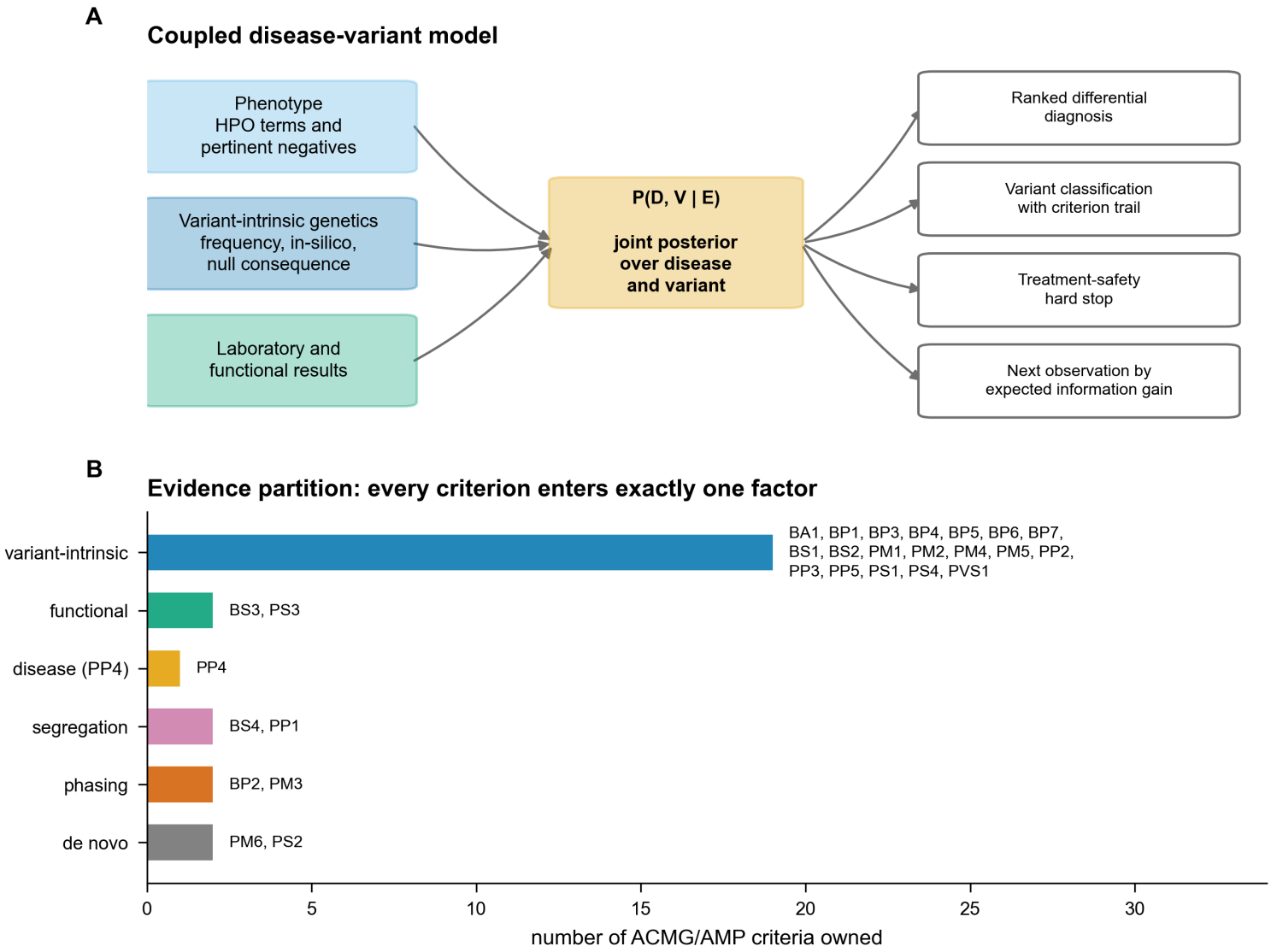


Figure 1

The DISCERN framework and its evidence partition.

(A) The coupled model. Phenotypic, variant-intrinsic, and functional evidence streams enter a joint posterior over disease and variant; outputs comprise a ranked differential diagnosis, a calibrated variant classification with its criterion trail, a treatment-divergence safety flag, and the recommended next observation. (B) The evidence partition. Each ACMG/AMP criterion is owned by exactly one factor, so no criterion contributes to the joint posterior more than once.

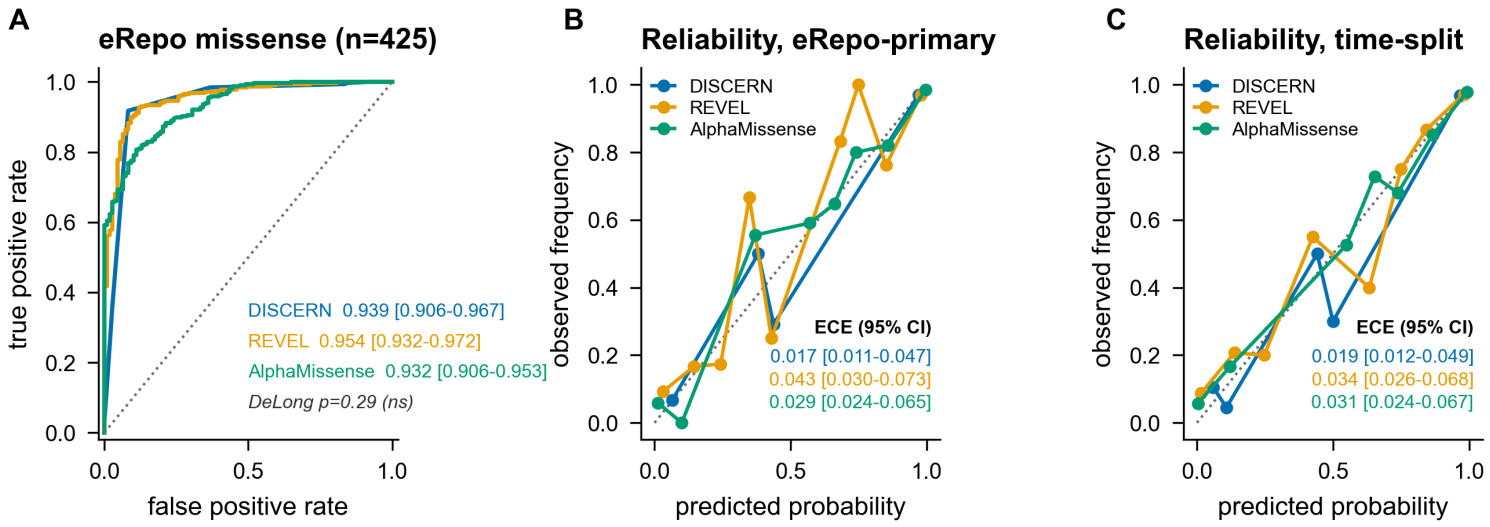


Figure 2

Variant discrimination and calibration on the ClinGen eRepo expert panel surface.

(A) Receiver operating characteristic curves for DISCERN, REVEL, and AlphaMissense on missense variants (n = 425; 424 for AlphaMissense); areas under the curve with bootstrap 95% CIs, and the paired DeLong p value for DISCERN against REVEL, are annotated on the panel. (B) Reliability diagrams for DISCERN, REVEL, and AlphaMissense calibrated under the identical out-of-fold isotonic protocol, with expected calibration error and 95% CIs. (C) As in (B), on the time-split panel, no tool could have observed the expert classification.

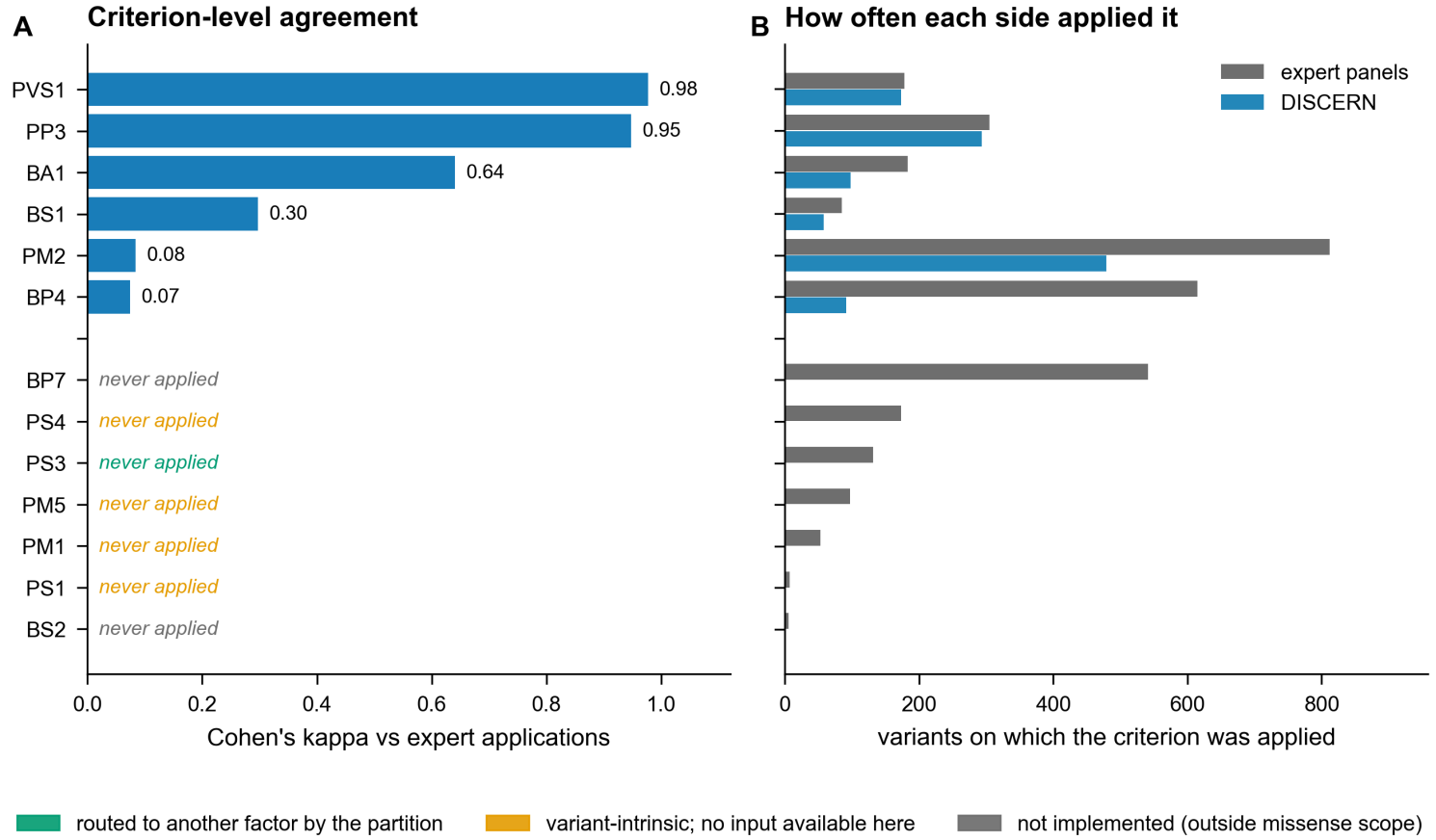


Figure 3

Per-criterion agreement with expert panel applications.

(A) Cohen's kappa between DISCERN and ClinGen eRepo expert panel criterion applications across the bleeding and platelet gene set, computed on the 1,265 variants carrying a pathogenic or benign expert classification and ordered from strongest agreement to weakest. Criteria applied zero times are shown beneath, by cause, in three categories that a single figure would conflate: PS3 is routed to the functional factor by the evidence partition, so its absence is expected behaviour, not a gap; PM1, PM5, PS1, and PS4 are variant-intrinsic criteria for which this pipeline supplies no input under the ClinVar-blinded protocol; and BP7 and BS2 are not implemented, BP7 because it addresses synonymous variants, which lie outside the missense scope of every reported result. (B) The number of variants on which each side applied the criterion, showing the direction and size of each disagreement. An unmodified code is resolved to its criterion's default strength before comparison; strength-level agreement is given in Table S5.

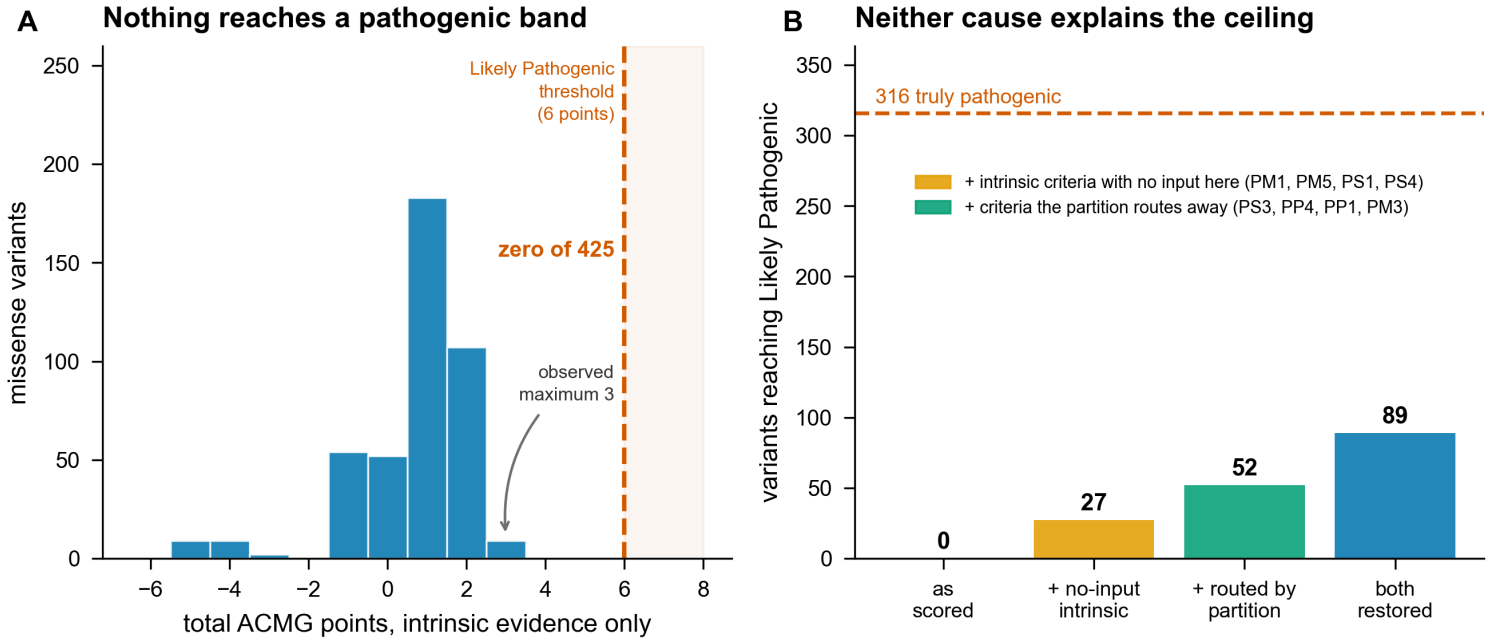


Figure 4

The ceiling that intrinsic evidence imposes on missense classification.

(A) Distribution of total ACMG points across missense variants (n = 425) under intrinsic evidence alone, against the six-point threshold for likely pathogenic; the observed maximum is 3.0 points, and no variant crosses the threshold. (B) Attribution of the ceiling. Restoring the variant-intrinsic criteria that this pipeline cannot annotate lifts 27 variants to likely pathogenic; restoring the criteria that the partition routes to other factors lifts 52; restoring both lifts 89 of the 316 variants the expert panels classified as pathogenic. Neither cause accounts for the ceiling alone.

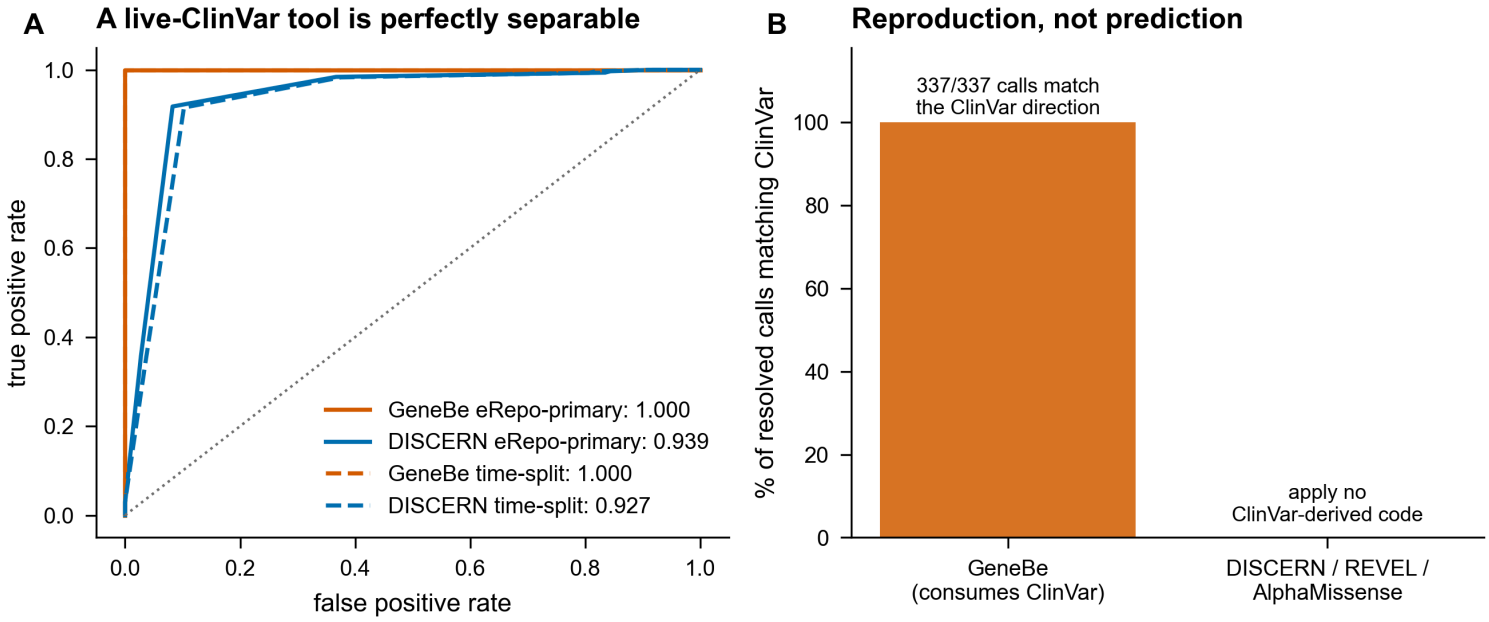


Figure 5

ClinVar circularity in a live-database classifier.

(A) Receiver operating characteristic for GeneBe on the eRepo surface and on the time-split panel, against DISCERN for reference; GeneBe attains an area under the curve of 1.0 on both surfaces. (B) Concordance with ClinVar: GeneBe reproduces the ClinVar direction on every gradeable call, whereas DISCERN, REVEL and AlphaMissense apply no ClinVar-derived criterion and so have no such call to reproduce. A tool querying live ClinVar cannot be fairly graded against ClinVar-derived truth.

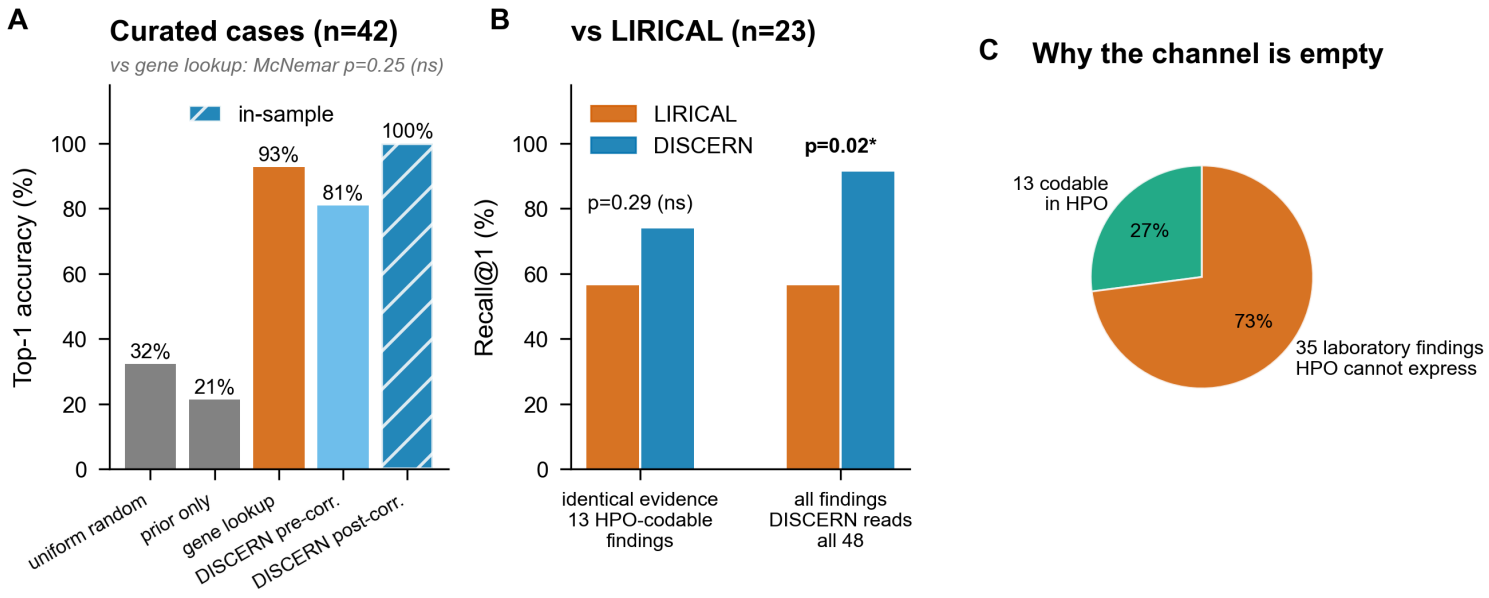


Figure 6

Differential diagnosis against phenotype-blind and external baselines.

(A) Top-1 accuracy on the curated benchmark (n = 42) for a uniform random floor, a prior-only ranking, a phenotype-blind gene lookup, and DISCERN before and after the disease-posterior correction; the post-correction bar is hatched to denote an in-sample result, as these cases exposed the defect that the correction fixed. (B) Comparison with LIRICAL on the 23 cases carrying any phenotype term, under matched evidence (the 13 findings the HPO can express, gene withheld from both tools) and under full evidence (including the 35 laboratory findings HPO cannot encode). (C) Representability of the benchmark's 48 discriminating features in HPO.

Supplementary Files

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

- Additionalfile1.pdf
- Additionalfile2.xlsx
- Additionalfile3.xlsx
- Additionalfile4.xlsx
- Additionalfile5.xlsx