Classification of In-frame Indels at Scale Using Calibrated Functional and Computational Evidence

Haneen Abderrazzaq[1], Ross Stewart[1], Malvika Tejura[2], Abbye E. McEwen[2,3,4], Silvia Casadei[3], Matthew W. Snyder[3], Nahum Smith[3], Alan F. Rubin[5], Douglas M. Fowler[2,3], Lea M. Starita[2,3], Predrag Radivojac[1]

[1]*Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts, U.S.A.* [2]*Department of Genome Sciences, School of Medicine, University of Washington, Seattle, Washington, U.S.A.* [3]*Brotman Baty Institute for Precision Medicine, University of Washington, Seattle, Washington, U.S.A.* [4]*Department of Laboratory Medicine and Pathology, School of Medicine, University of Washington, Seattle, Washington, U.S.A.* [5]*Collaborative Centre for Genomic Cancer Medicine, University of Melbourne, Melbourne, Victoria, Australia*

In-frame (non-frameshifting) insertion and deletion (indel) variants present a persistent challenge for variant classification as they require timely and accurate interpretation when observed in patients. We previously calibrated computational prediction tools for in-frame indels according to ACMG/AMP evidence guidelines, enabling their use in clinical variant classification. We additionally generated and curated high-throughput functional data corresponding to ten disease-associated genes, and similarly calibrated these data for clinical evidence strengths. Following an approach previously applied to missense variants, here we combine these two calibrated evidence sources to resolve in-frame indel variants of uncertain significance (VUS) across the ten genes. Applying this integrated evidence, we resolve up to 57% of the existing ClinVar VUS in these genes. We further apply this evidence to preclassify up to 6,000 in-frame indels not yet reported in ClinVar across these genes, providing a precomputed resource for future variant classification. Our results demonstrate that the combination of computational and functional evidence can substantially reduce the VUS burden of in-frame indels and will help improve clinical variant classification and decision-making. Finally, we show that difficulties in predicting and functionally characterizing the effects of in-frame indel variation across indel type and length present unique challenges that will require attention and customized solutions in the future.

Keywords: variant classification, in-frame indels, non-frameshifting indels, saturation genome editing, variant effect prediction

1. Introduction

Accurate variant classification is central to genomic medicine for rare diseases as it plays an integral role in guiding diagnosis and medical management for patients and their families.[1] As genetic testing has become more widespread, an increasing number of variants are identified in patients, yet most remain classified as variants of uncertain significance (VUS).[2] These unresolved classifications limit the clinical utility of genomic testing and prolong a patient's diagnostic odyssey. There is therefore a pressing need for strategies to rapidly resolve VUS.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

To standardize the practice of clinical variant classification, the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) have released evidence-based guidelines.[3] These guidelines define how different lines of evidence (e.g., population, functional, computational) are combined to classify a variant as pathogenic, likely pathogenic, variant of uncertain significance (VUS), likely benign, or benign. This system was later formalized into a quantitative, point-based framework that maps combinations of evidence to discrete point totals, which then determine each classification.[4,5] Applying this framework requires calibration of individual lines of evidence, the process of mapping raw scores (e.g., from variant effect predictors or high-throughput experimental assays) to discrete points such that the point totals correspond to the posterior probability thresholds for classification. Specifically, the classifications correspond to the following point totals and posterior probabilities of pathogenicity: pathogenic ($\geq$10 points; >0.99), likely pathogenic (6 to 9 points; 0.9–0.99), VUS (0 to 5 points; 0.1–0.9), likely benign (-6 to -1 points; 0.01–0.1), and benign (≤ -7 points; <0.01).

Substantial efforts have been devoted to studying the effects of missense variants, the most common class of coding variants, including the development and evaluation of computational models and the high-throughput functional characterization of their effects,[2,6,7] collectively improving missense variant classification. On the computational side, Pejaver et al.[8] introduced a framework for the calibration of missense predictors, subsequently extended to gene- and domain-aware calibration.[9] On the functional side, ExCALIBR[10] was developed to calibrate experimental assays, converting functional scores into variant-specific evidence points compatible with ACMG/AMP guidelines.[5] Applying both approaches, Tejura et al.[11] recently demonstrated the power of combining calibrated computational and functional evidence to reclassify missense VUS at scale. Across a set of 40 disease-associated genes, this approach achieved a discordance rate below 1% on ClinVar controls (i.e., variants previously classified as pathogenic or benign), resolved approximately 75% of existing VUS, and preclassified over 62% of variants from these genes not yet observed in ClinVar. This study provides a blueprint for a substantial reduction in VUS, including for variants not yet observed clinically.[11]

After missense variants, in-frame indels are the second largest class of protein-altering variants,[12] with approximately 285,000 rare in-frame indels in gnomAD.[13] Indels vary in type (insertion or deletion) and length, which complicates their study. Furthermore, clinical and population databases contain far fewer in-frame indels than missense variants, limiting the development of computational variant effect predictors. Due to these challenges, less attention has been devoted to generating functional data and developing computational tools for indels, leaving them understudied relative to missense variants. Nonetheless, a substantial number of indels still require clinical interpretation; there are approximately 29,000 in-frame indels across over 3,000 disease-associated genes (definitive, strong, or moderate gene-disease validity in GenCC[14]) in ClinVar, over 72% of which are VUS.

In this work, we adapt the approach of Tejura et al.[11] to apply calibrated functional (experimental) and computational evidence to resolve in-frame indel VUS and preclassify unobserved variants. Within the Impact of Genomic Variation on Function (IGVF) Consortium,[15] we have generated Saturation Genome Editing (SGE) for single amino acid deletions in seven genes:

BARD1, *BRCA2*, *CTCF*, *PALB2*, *RAD51D*, *SFPQ*, and *XRCC2*. Six of these genes, along with three previously published functional assays for *BAP1*,[16] *DDX3X*,[17] and *RAD51C*,[18] contain measurements for the existing ClinVar VUS, while the *SFPQ* functional data were only used for preclassification of variants yet to be recorded in ClinVar. We have recalibrated these experimental assays using ExCALIBR.[10] Separately, we previously calibrated computational predictors for in-frame indels and found that they often reach moderate evidence (+2 to +3 points) of pathogenicity and supporting evidence (−1 point) of benignity.[12] By combining this calibrated functional evidence and calibrated computational evidence, we resolve in-frame indel ClinVar VUS for these nine genes and preclassify thousands of unobserved variants across all ten genes.

2. Materials and Methods

2.1. *ClinVar dataset*

We collected in-frame indels from ClinVar (September 2026) for the purpose of reclassification. Variants were filtered to indel length ≤ 50 bp, review status of at least one star, and SpliceAI[19] score < 0.2, following the same approach used in reclassification of missense variants.[11,20] Variants were restricted to ten genes that we had available SGE data for; however, no variants in *SFPQ* met our final filtering criteria. There were a total of 511 variants in this dataset, 109 of which also had functional data. Of the 511 variants, 23 had an existing clinical significance label; 14 were pathogenic or likely pathogenic (treating pathogenic/likely pathogenic as likely pathogenic) and nine were benign or likely benign (treating benign/likely benign as likely benign). 488 variants were VUS (447) or had conflicting classifications (41); for the purposes of this study, VUS and variants with conflicting classifications are collectively referred to as VUS. The distribution of ClinVar in-frame indels by gene is shown in Figure 1a-c. Figure 1d and 1e show the length distribution and the clinical significance breakdown of this dataset.

2.2. *Computational predictor selection and calibration*

We used calibrated evidence thresholds for in-frame indel computational predictors established by Abderrazzaq et al.[12] This approach first estimates the prior probability of pathogenicity for insertions and deletions separately, using labeled clinical variants and unlabeled population variants. These priors are then combined with the local positive likelihood ratios estimated from labeled clinical variants at a predictor's score s to estimate this predictor's local posterior probabilities of pathogenicity. The score range for each predictor is then discretized into points based on the estimated posteriors for each score s.[8]

From the indel predictors calibrated therein, we selected three tools with different approaches and strengths: CADD,[21] ESM1b,[22,23] and MutPred-Indel.[24] MutPred-Indel is one of the strongest-performing tools, reaching +3 evidence for pathogenicity for deletions and +2 for insertions, and −2 and −1 evidence for benignity for deletions and insertions, respectively. ESM1b is a protein language model, part of a class of tools of growing interest for variant effect prediction. It also reaches +3 evidence for pathogenicity for deletions and +2 for insertions, but does not reach the threshold for any benign-direction evidence. Finally, we included

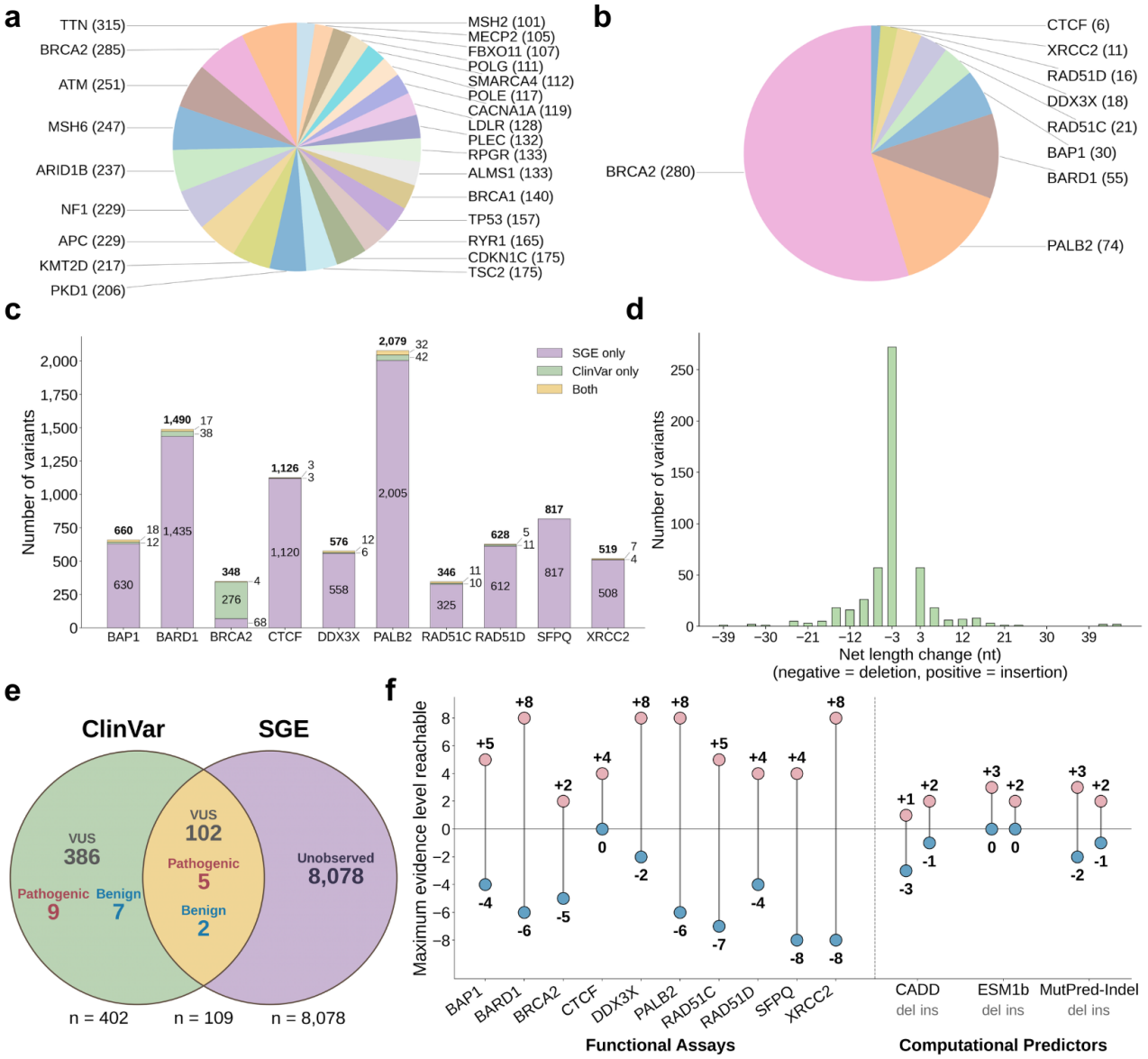


Fig. 1. a. Top 25 disease-associated genes in ClinVar by in-frame indel count. b. In-frame indels in ClinVar for each gene used in this study. c. Counts of variants per gene in ClinVar only, SGE only, and both datasets in this study. d. Length distribution of variants in ClinVar data. e. Overlap between ClinVar and SGE datasets, broken down by clinical significance. f. Maximum evidence level reachable for each functional assay (left) and each computational predictor (right) used in this study. Note that computational predictors have separate thresholds for deletions and insertions.

CADD, which reaches −3 and +1 evidence for deletions and −1 and +2 evidence for insertions. Among the three predictors, CADD provides the strongest benign-direction evidence, offering the greatest potential to push variants already classified as likely benign by functional evidence to a full benign classification (≤ -7 points) when combined. CADD is also widely used and easily accessible. VEST-indel[25] and PROVEAN[26] reached similar evidence levels to our selected tools but were unavailable at the time of carrying out this study. Figure 1f shows

the maximum evidence levels reached by each predictor used in this study.

2.3. *High-throughput functional data*

Saturation Genome Editing (SGE) data were generated according to methods described by Tejura et al.[11] for *BARD1*, *BRCA2*, *CTCF*, *PALB2*, *RAD51D*, *SFPQ*, and *XRCC2*; published SGE data were used for *BAP1*,[16] *DDX3X*,[17] and *RAD51C*.[18] Genes were selected based on clinical relevance; all genes except *SFPQ* have an established gene-disease relationship, defined by a definitive, strong, or moderate classification in GenCC.[14] *SFPQ* was added as a candidate disease gene due to its potential link to neurodegenerative disorders.[27] Following the criterion described above, we excluded variants with a SpliceAI score ≥ 0.2. There were a total of 8,187 variants with SGE scores in our final dataset, 109 of which were also in ClinVar. SGE data used in this study consisted of single amino acid deletions, with the exception of four multi-amino acid deletions in the published data sources. Variant counts per gene are shown in Figure 1c and dataset overlap with ClinVar by clinical significance is shown in Figure 1e.

2.4. *Functional assay calibration*

Functional scores for all ten genes were converted to ACMG/AMP evidence strengths[3,5] using ExCALIBR,[10] a statistical method for calibrating high-throughput experimental assay data and an improvement over the existing guidelines.[28] For each gene, ExCALIBR models functional scores from pathogenic, benign, population and synonymous variants to estimate the probability of pathogenicity of each variant, which is converted to integer evidence points between -8 (strongest benign evidence) and $+8$ (strongest pathogenic evidence). Figure 1f shows the maximum evidence levels reached by each functional assay.

2.5. *Combination of evidence*

We followed the same ACMG/AMP evidence combination workflow used by Tejura et al.[11] to combine experimental and computational evidence points. After summing evidence points from these two sources, variants were classified according to the following point ranges: benign ≤ -7, likely benign -6 to -1, VUS 0 to 5, likely pathogenic 6 to 9, pathogenic ≥ 10.[4,5] Each predictor's evidence points were combined separately with experimental evidence points; predictors were not combined with one another, as they are not independent and often share overlapping training data and features.[3]

3. Results

3.1. *Performance on ClinVar controls*

We first assessed the performance of functional and computational evidence using variants with an existing pathogenic or benign classification in ClinVar (Figure 2). Our ClinVar dataset contained 23 labeled (pathogenic, likely pathogenic, benign, or likely benign) in-frame indels across the nine genes with variants in ClinVar. We examined the performance of the functional data and each of the three predictors separately.

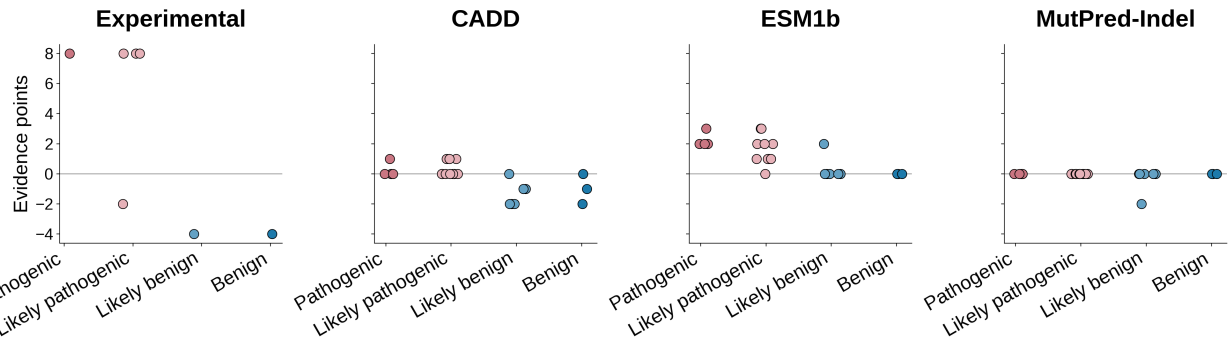


Fig. 2. Evidence points assigned to ClinVar variants labeled as pathogenic, likely pathogenic, benign, or likely benign by each method. Each predictor scored all 23 variants, but experimental data were only available for seven of these variants.

Seven of these 23 labeled variants had an available experimental score, and the experimental evidence agreed with the ClinVar label for six of the seven variants. The one exception was *DDX3X* p.Pro41del, labeled as likely pathogenic but scoring −2 from experimental data. The remaining four variants that had pathogenic or likely pathogenic classifications, all also in *DDX3X*, scored strongly positive (+8 points) with this assay. The two benign-labeled variants with experimental data, both *BAP1* variants, each scored −4.

The predictors each scored all 23 control variants. CADD assigned likely benign evidence (−1 or −2 points) to seven of the nine benign or likely benign variants, concordant with their ClinVar classification, and assigned +1 evidence to four of the 14 pathogenic or likely pathogenic variants. ESM1b assigned between +1 and +3 points to 13 of the 14 pathogenic or likely pathogenic variants. ESM1b made one error on a likely benign variant, which received +2 points. MutPred-Indel scored every variant, but only one raw score crossed a calibrated threshold, giving −2 points to a variant labeled as likely benign.

Although these results are promising, there are not yet enough in-frame indels in these nine genes with definitive ClinVar classifications to reliably estimate the error rate of experimental or computational evidence. For missense variants, the error rate after combining these two lines of evidence for ClinVar controls was under 1%,[11] calculated on a substantially larger variant set.

3.2. *Classification of ClinVar VUS*

We next assessed how well functional and computational evidence could classify the 488 existing VUS in ClinVar across the nine genes (Figures 3–5). Although experimental data were available for only 102 of these variants, functional evidence can reach a substantially higher magnitude (up to ±8 points) than the computational predictors used here, which do not exceed ±3 evidence points (Figure 3). Experimental data alone resolved 71 of these 102 VUS (70%): 6 as benign, 54 as likely benign, and 11 as likely pathogenic (Figures 4 and 5).

Among the predictors, CADD assigned nonzero points to the majority of all VUS: 243 variants received negative points, each sufficient on its own for a likely benign reclassification, and 40 variants received positive points (Figure 3). ESM1b assigned nonzero evidence to 123

variants, all in the pathogenic direction, while MutPred-Indel assigned nonzero evidence to 120 variants, most in the benign direction (Figure 3).

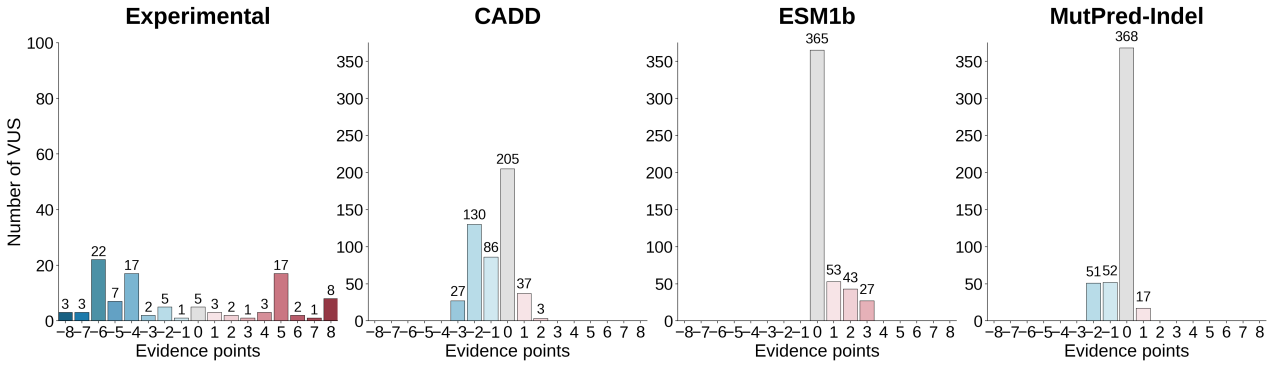


Fig. 3. Evidence points assigned to ClinVar VUS by method. Each predictor scored all 488 VUS, but experimental data were only available for 102 of these.

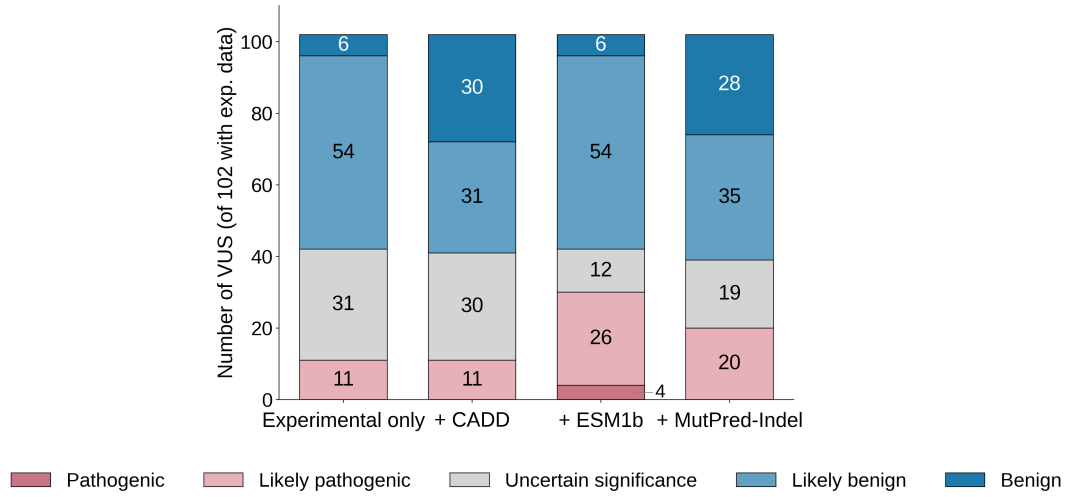


Fig. 4. Resulting classifications for ClinVar VUS with available experimental data. The leftmost bar uses experimental evidence only; each of the other three bars adds one predictor's evidence to the experimental evidence.

The addition of evidence from each predictor to functional evidence further improved VUS classification. Among the 102 variants with available experimental data, all three predictors strengthened existing classifications and resolved additional VUS (Figure 4). CADD moved 24 from likely benign to benign and one each from VUS to likely benign and likely pathogenic. ESM1b moved four variants from likely pathogenic to pathogenic and 19 from VUS to likely pathogenic. MutPred-Indel moved 22 from likely benign to benign, three from VUS to likely benign, and 10 from VUS to likely pathogenic. A small number of variants moved in the opposite direction: CADD and MutPred-Indel each returned one variant from likely pathogenic to VUS. Overall, adding computational evidence increased the number of resolved variants

from 71 with functional evidence alone to 72 (71%) with CADD, 90 (88%) with ESM1b, and 83 (81%) with MutPred-Indel.

Across all 488 VUS, CADD resolved the most when combined with functional evidence (279; 57%), followed by MutPred-Indel (141; 29%) and ESM1b (90; 18%) (Figure 5). Because the predictors provide at most +3 points, short of the +6 required for likely pathogenic, they can resolve VUS without functional evidence only in the benign direction. ESM1b, which provides only pathogenic evidence, therefore resolves VUS only in combination with experimental data. The number of resolved VUS here is limited by the low overlap between ClinVar VUS and available experimental data; given the results in Figure 4, we expect this fraction to increase as more functional data are generated and calibrated.

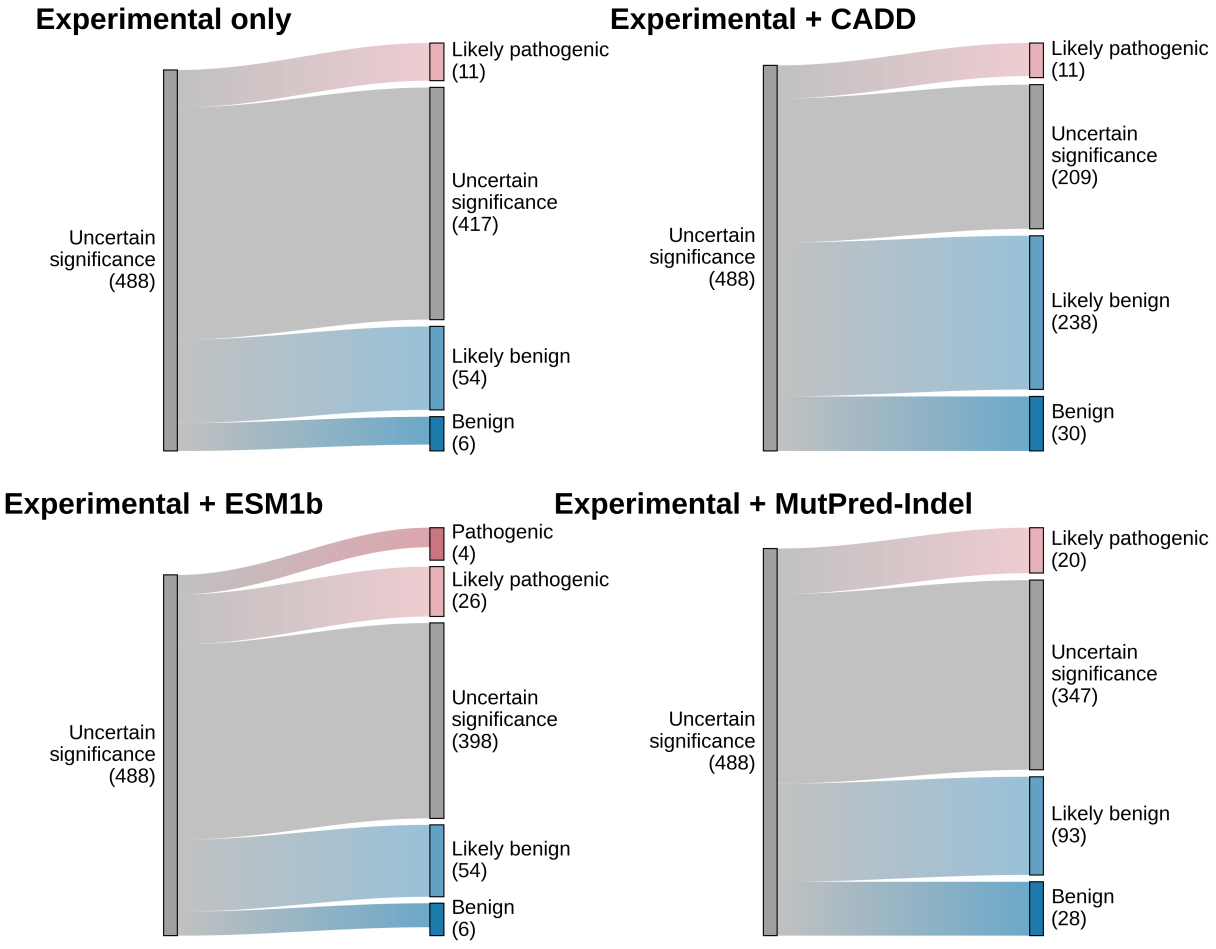


Fig. 5. Sankey diagram depicting classifications for all ClinVar VUS using experimental evidence alone and experimental evidence combined separately with each predictor.

3.3. *Preclassification of clinically unobserved variants*

We next assessed the potential of functional and computational evidence to preclassify in-frame indels not yet observed in ClinVar (Figure 6). Among the 8,078 SGE-covered indels not yet observed in ClinVar across our ten genes, functional evidence alone preclassified 5,145 of these variants (64%). Combining functional evidence with each predictor increased this to 5,262 (65%) with CADD, 6,027 (75%) with ESM1b, and 5,802 (72%) with MutPred-Indel. We note that 108 variants in *SFPQ* reached likely pathogenic when functional evidence was combined with ESM1b. However, because the *SFPQ* functional data were calibrated in a negative-unlabeled setting without pathogenic controls, this evidence should be interpreted as functional abnormality rather than a direct measure of pathogenicity.

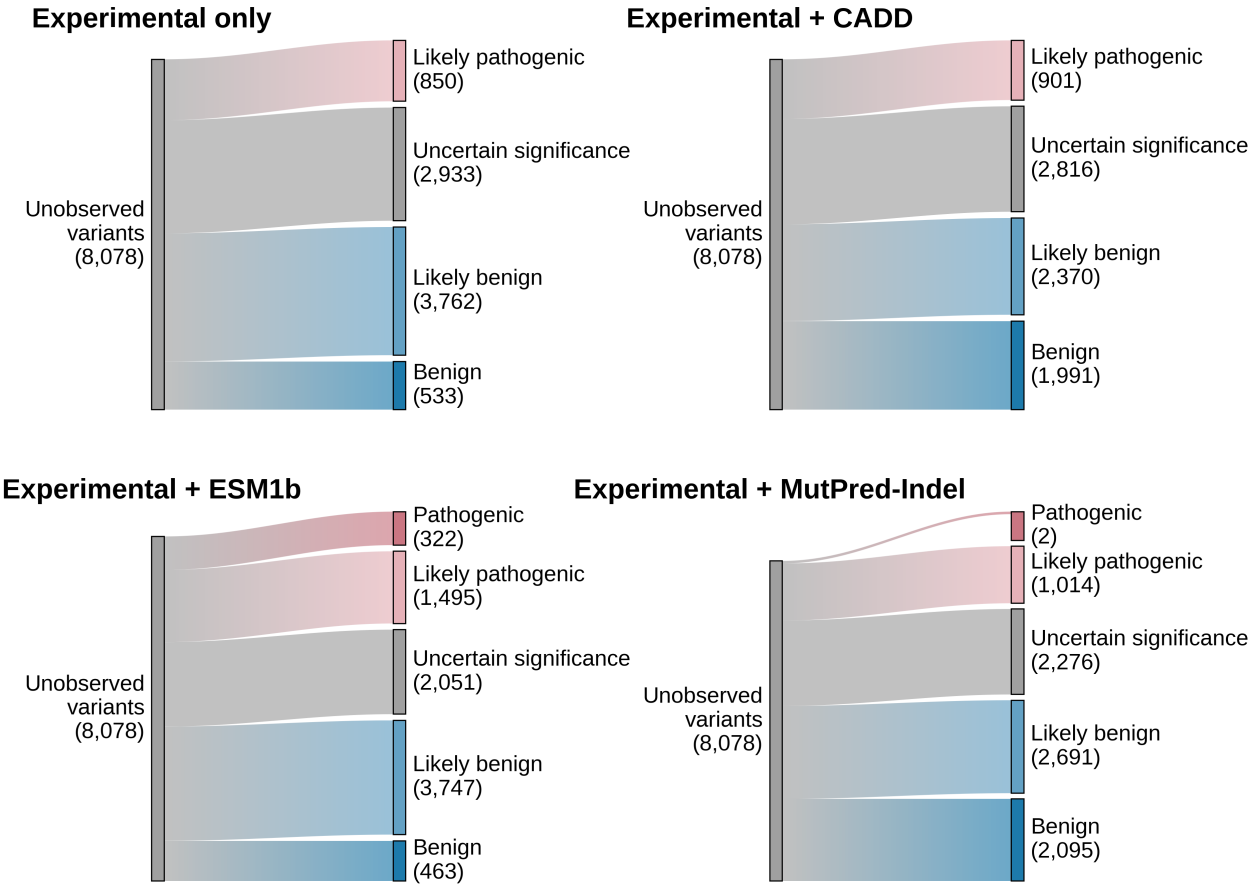


Fig. 6. Sankey diagram depicting preclassifications for clinically unobserved variants using experimental evidence alone, and experimental evidence combined separately with each predictor. Only indels with available experimental data were considered for preclassification due to an extremely large number of potential in-frame indels shorter than 50 bp.

3.4. *Agreement across scoring methods*

We evaluated the concordance between functional and computational evidence by determining the disagreement between method pairs, defined as one method assigning positive evidence while the other assigns negative evidence (Figure 7). Across both the VUS and clinically unobserved variant sets, nearly all method pairs disagreed less than 5% of the time. ESM1b had the highest disagreement with experimental data among unobserved variants (10.3%), and tied with MutPred-Indel for the highest disagreement among VUS (3.9%). CADD had the fewest disagreements with functional data, at 2.0% for VUS and 3.2% for unobserved variants.

Figure 8 shows the final classifications reached by combining each predictor with functional evidence, across all ClinVar VUS and unobserved variants (8,566 in total), broken down by evidence source and the frequency of conflicts between the two lines of evidence. Because ESM1b contributes only pathogenic evidence, all of its disagreements occurred when the functional evidence was benign. For MutPred-Indel, 58% of disagreements (225 of 388) resulted in a final VUS classification; nearly all of these involved moderately positive experimental evidence offset by −1 or −2 points from MutPred-Indel. The remaining disagreements were distributed across likely pathogenic, likely benign, and benign classifications, with none clearly predominating. CADD disagreements resulted in a final VUS classification 37% of the time (94 of 257), similarly involving moderately positive functional evidence offset by −1 or −2 CADD points. Most of the remaining disagreements were likely benign outcomes (85 of 257).

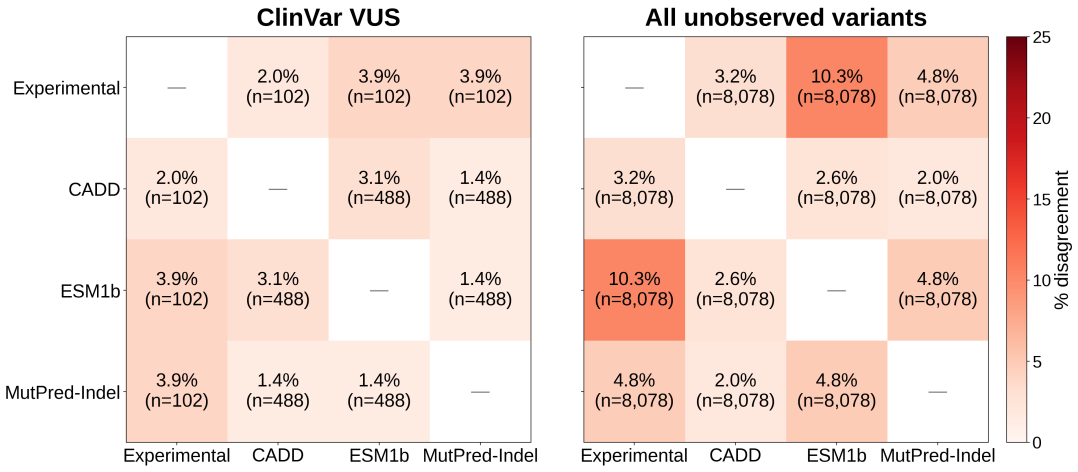


Fig. 7. Pairwise disagreement rates among experimental evidence and the three computational predictors, for ClinVar VUS (left) and unobserved variants (right). Disagreement is defined as one method assigning positive evidence while the other assigns negative evidence.

4. Discussion

We demonstrate that the high-throughput functional data alone are highly informative for classifying in-frame indels. When available, experimental evidence was sufficient to move a majority of VUS to determinate classifications (benign, likely benign, or likely pathogenic),

specifically 70% of VUS with SGE data in this study. We also found that experimental evidence alone was able to preclassify 64% of variants with SGE data that were not observed in ClinVar. Although experimental data are instrumental for variant classification, they have important limitations. Assays such as SGE typically measure one aspect of variant effect and may not capture other functional consequences, so a variant that appears normal in one assay could still be pathogenic through a different mechanism. Conversely, abnormal assay results do not always indicate clinical pathogenicity. Functional data generation is also constrained by cost and the time required for assay optimization.

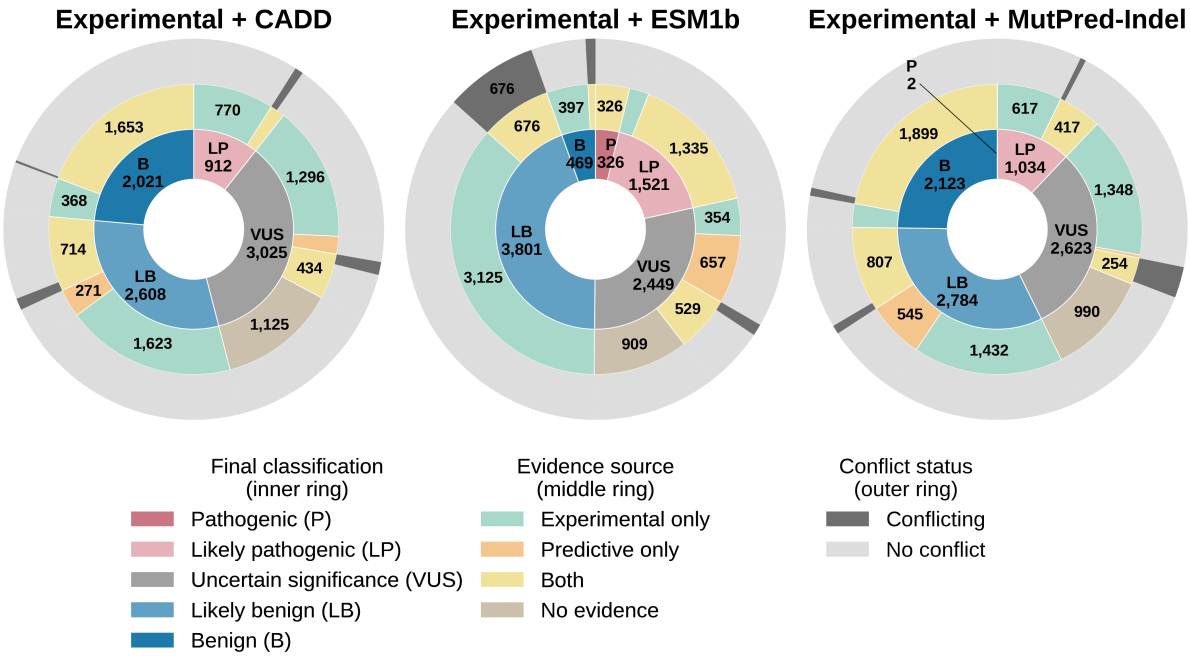


Fig. 8. Breakdown of final classifications by evidence source and conflict status, for each predictor combined with experimental evidence, across all ClinVar VUS (488) and unobserved variants (8,078).

Although computational predictors cannot reach evidence as strong as experimental data on their own, they add substantial value for classification. First, we found that computational evidence supplements experimental evidence, frequently contributing additional evidence points that rarely disagree in direction with experimental evidence (Figure 7). Computational evidence is especially useful for pushing variants classified by functional evidence as likely pathogenic into a pathogenic classification, as functional evidence is capped at +8 points,[4] short of the 10 points needed for pathogenic classification. We observe this effect in 322 unobserved variants upon the addition of evidence from ESM1b (Figure 6). Second, computational predictors can score a much wider range of variants than current functional assays cover. Although our SGE datasets provide substantial functional data for single amino acid deletions, this leaves a gap in experimental coverage for studying the effects of all insertions and longer deletions, which are more challenging to generate with high-throughput functional assays. Computational predictors, however, can easily score insertions and longer

deletions, extending classification potential beyond the available functional data alone. This is particularly valuable for likely benign classification, which requires only a −1 evidence point according to the forthcoming SVCv4 guidelines. Although we only generated predictor scores for the subset of unobserved variants with available SGE data, predictors could in principle be applied more broadly to additional variants not covered by our SGE data, classifying them as likely benign. We note that our comparison of the three predictors in this study is to characterize their relative performance and utility for indel classification; the choice of which predictors to adopt for clinical use is left to diagnostic laboratories.

One limitation of our evaluation is that there are too few ClinVar control variants with associated functional data to estimate an error rate for the classification of indels; only seven labeled variants were available for evaluation and six of these were assigned experimental evidence in-line with their ClinVar classification. However, our calibration[9,10] and evidence combination approach[11] have demonstrated a very low error rate for ClinVar variants overall, suggesting that this method is likely to reliably generalize to indels and achieve necessary rigor in variant classification.[29] While currently available in-frame indel predictors provide useful evidence for classification, they do not yet reach the evidence strengths achieved by predictors for missense variants.[12] Additionally, only a subset of existing predictors have been calibrated to ACMG/AMP evidence strengths and are therefore usable within this framework.[12] Continued development of indel-specific computational methods, along with sustained maintenance of existing tools, will help close this gap and improve indel classification.

5. Conclusion

This study demonstrates the potential of combining functional and computational evidence to resolve in-frame indel VUS and preclassify thousands of clinically unobserved variants. This represents an important step toward reducing the growing VUS burden of in-frame indels, a variant class that has lacked at-scale classification methods. Beyond the classification of VUS in ClinVar, our preclassification approach allows determinate evidence to be immediately available upon a variant's first clinical observation. We plan to expand this study as more functional and clinical data accumulate and as additional computational tools, including newer models, become available.

6. Acknowledgments

This work was supported by the NSF GRFP award 2023361442 (H.A.) and the NIH National Human Genome Research Institute (NHGRI) awards U01HG012022 (P.R.) and UM1HG011969 (L.M.S., D.M.F.).

7. Data Availability

All SGE data are available in the IGVF catalog.[30] Additional data are available at `https://github.com/habderrazzaq/indel_classification`.

References

1. M. H. Wojcik *et al.*, Genome sequencing for diagnosing rare diseases, *N Engl J Med* **390**, 1985 (2024).
2. A. E. McEwen *et al.*, Multiplexed assays of variant effect for clinical variant interpretation, *Nat Rev Genet* **27**, 137 (2026).
3. S. Richards *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* **17**, 405 (2015).
4. S. V. Tavtigian *et al.*, Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework, *Genet Med* **20**, 1054 (2018).
5. S. V. Tavtigian *et al.*, Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines, *Hum Mutat* **41**, 1734 (2020).
6. P. Katsonis *et al.*, Genome interpretation using in silico predictors of variant impact, *Hum Genet* **141**, 1549 (2022).
7. Y. J. Lin *et al.*, Variant Impact Predictor database (VIPdb), version 2: trends from three decades of genetic variant impact predictors, *Hum Genomics* **18**, p. 90 (2024).
8. V. Pejaver *et al.*, Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria, *Am J Hum Genet* **109**, 2163 (2022).
9. Y. Chen *et al.*, Gene- and domain-aware calibration increases the clinical utility of variant effect predictors, *bioRxiv 2026.02.17.706269* (2026).
10. D. Zeiberg *et al.*, Gene-based calibration of high-throughput functional assays for clinical variant classification, *bioRxiv 2025.04.29.651326* (2025).
11. M. Tejura *et al.*, A scalable approach to resolving variants of uncertain significance, *bioRxiv 2026.02.14.705848* (2026).
12. H. Abderrazzaq *et al.*, Calibration of in-frame indel variant effect predictors for clinical variant classification, *bioRxiv 2026.04.15.718599* (2026).
13. K. J. Karczewski *et al.*, The mutational constraint spectrum quantified from variation in 141,456 humans, *Nature* **581**, 434 (2020).
14. M. T. DiStefano *et al.*, The Gene Curation Coalition: a global effort to harmonize gene-disease evidence resources, *Genet Med* **24**, 1732 (2022).
15. IGVF Consortium, Deciphering the impact of genomic variation on function, *Nature* **633**, 47 (2024).
16. A. J. Waters *et al.*, Saturation genome editing of BAP1 functionally classifies somatic and germline variants, *Nat Genet* **56**, 1434 (2024).
17. E. J. Radford *et al.*, Saturation genome editing of DDX3X clarifies pathogenicity of germline and somatic variation, *Nat Commun* **14**, p. 7702 (2023).
18. R. Olvera-León *et al.*, Comprehensive saturation genome editing of RAD51C provides precise functional classification within a spectrum of variant effects, *Cell* **187**, 5719 (2024).
19. K. Jaganathan *et al.*, Predicting splicing from primary sequence with deep learning, *Cell* **176**, 535 (2019).
20. L. C. Walker *et al.*, Using the ACMG/AMP framework to capture evidence related to predicted and observed impact on splicing: recommendations from the ClinGen SVI Splicing Subgroup, *Am J Hum Genet* **110**, 1046 (2023).
21. M. Schubach *et al.*, CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions, *Nucleic Acids Res* **52**, D1143 (2024).
22. A. Rives *et al.*, Biological structure and function emerge from scaling unsupervised learning to

250 million protein sequences, *Proc Natl Acad Sci U S A* **118**, p. e2016239118 (2021).
23. N. Brandes *et al.*, Genome-wide prediction of disease variant effects with a deep protein language model, *Nat Genet* **55**, 1512 (2023).
24. K. A. Pagel *et al.*, Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome, *PLoS Comput Biol* **15**, e1007112 (2019).
25. C. Douville *et al.*, Assessing the pathogenicity of insertion and deletion variants with the Variant Effect Scoring Tool (VEST-Indel), *Hum Mutat* **37**, 28 (2016).
26. Y. Choi *et al.*, Predicting the functional effect of amino acid substitutions and indels, *PLoS One* **7**, p. e46688 (2012).
27. P. M. Gordon *et al.*, A conserved role for the ALS-linked splicing factor SFPQ in repression of pathogenic cryptic last exons, *Nat Commun* **12**, p. 1918 (2021).
28. S. E. Brnich *et al.*, Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework, *Genome Med* **12**, p. 3 (2019).
29. B. Rost *et al.*, Protein function in precision medicine: deep understanding with machine learning, *FEBS Lett* **590**, 2327 (2016).
30. D. Li *et al.*, The IGVF catalog-from genetic variation to function, *Nucleic Acids Res* **54**, D1437 (2026).