https://doi.org/10.1038/s41588-024-01821-8

Improving transparency of computational tools for variant effect prediction

Rachel Karchin, Predrag Radivojac, Anne O'Donnell-Luria, Marc S. Greenblatt, Michael Y. Tolstorukov & Dmitriy Sonkin

Efforts to integrate computational tools for variant effect prediction into the process of clinical decision-making are in progress. However, for such efforts to succeed and help to provide more informed clinical decisions, it is necessary to enhance transparency and address the current limitations of computational predictors.

Variant pathogenicity classification and interpretation is increasingly integrated into the process of clinical decision-making and has evolved into an important component of modern medicine. Multiple guidelines have been created to provide systematic and consistent protocols to perform variant pathogenicity classification and interpretation[1–3]. Such guidelines include the use of artificial intelligence in the form of computational tools for variant effect prediction. At the end of 2022, ClinGen published recommendations[4] on the use of computational tools for missense variant pathogenicity classification, augmenting the American College of Medical Genetics and Genomics (ACMG)/Association for Molecular Pathology (AMP) guidelines[1] for the classification of germline variants for monogenic conditions. The recommendations provide evidence for increasing the strength of criteria from computational tools for missense variant effect prediction. The important issue of double counting of evidence has been raised[4]; this arises from the potential use of information to train computational predictors and then subsequently using this very information as evidence within the ACMG/AMP guidelines. Population frequency, functional information, protein structure, mutational hotspots data, clinical annotations, as well as evolutionary sequence conservation are examples of such evidence types. These types of evidence are frequently included as features by variant effect predictors and are described in detail in most predictor publications. Therefore, to minimize double counting of evidence, careful and detailed examinations of evidence types used by computational predictors is required.

However, many key potential users of computational predictors, such as variant curators, clinical geneticists and medical doctors, are known to be under severe time constraints and cannot be realistically expected to review publications at the level of detail required or keep track of updates in variant effect predictors. To be able to apply computational predictors intelligently and to decrease double counting of evidence, diagnostic laboratory personnel and other users need easily accessible information about the features and 'reasoning' underlying a predictor of interest. Below, we outline the types of the additional information for which availability, especially if provided in a structured form, would be beneficial to increase the transparency and usefulness of computational tools for variant effect prediction.

Inputs and outputs of variant effect predictors

For predictors that use supervised learning (requires labeled training data), it is critically important to know if a variant, for which the prediction is provided, was used in the training data set, and if so, whether it contributed to this particular prediction; if this is the case, the predictor is likely to provide only limited additional evidence for overall classification. Furthermore, if a variant was in the training set for a predictor, it is important to know the criteria underlying its inclusion in the training set. In an extreme case, some databases used to construct training sets include variants that themselves have been classified as such on the basis of the results of a variant-effect prediction tool, producing circularity. This circularity may be difficult to measure and will require tightening of the criteria on reporting variant pathogenicity. For example, resources used to extract training sets may not list all the criteria used to assert variant pathogenicity, as is the case for ClinVar[5], and moreover, authors of publications used for reporting in the Human Gene Mutation Database[6] may implicitly use computational tools, but not credit them formally. The circularity issue has the potential to become more prominent as the use of generative artificial intelligence approaches enters the field of variant pathogenicity prediction, and in some cases limited attention is paid to the careful vetting of training data.

Both supervised and unsupervised (do not require labeled training data) predictors may include variant population frequency, functional evidence, mutational hotspots, evolutionary conservation and protein domain data as features, and, to mitigate double counting of evidence, computational prediction should not be used in conjunction with ACMG/AMP evidence used at full strength and based on the same features. It is also important to know the value and source of each feature, as well as databases (and versions) from which the information was extracted; for example, specifying variant population frequency source as filtering allele frequency from Genome Aggregation Database (gnomAD) v4.1.0. For all these evidence types, it would also be useful to know the relative weight of the evidence for an overall computational prediction, if such information is available and relevant to the prediction algorithm.

Furthermore, for variant classification and clinical interpretation, it is very important to know if a variant acts via a loss-of-function (LOF) or gain-of-function (GOF) mechanism. Therefore, computational predictors should specify if a prediction indicates LOF, GOF or makes no such claim. Table 1 lists examples of well-known computational predictors with key inputs and variant types covered. Of note, none of these predictors, with the exception of MutPred2, attempt to provide LOF or GOF inference.

Providing such level of detail will be especially challenging for meta-predictors, which use as input predictions from other computational predictors. By implicitly counting the same types of features

Table 1 | Examples of well-known computational predictors

Prediction algorithm	Variant types covered	Input data types	Website	Meta-predictor (list of predictors)
AlphaMissense	Missense	Reference protein sequence-based protein 3D structure predictions, amino acid frequency statistics, population frequency	https://github.com/google-deepmind/alphamissense	No
BayesDel	Single-nucleotide variants, inframe indels, non-coding variants	Sequence conservation, clinical annotations, other predictors	https://fenglab.chpc.utah.edu/BayesDel/BayesDel.html	Yes (PolyPhen-2, SIFT, FATHMM, LRT, MutationTaster, Mutation Assessor, phyloP, GERP++, SiPhy)
CADD	Single-nucleotide variants, multi-nucleotide substitutions, indels	Reference protein sequence, sequence conservation, protein functional annotations, other predictors	https://cadd.gs.washington.edu/snv	Yes (SIFT, PolyPhen, phastCons, phyloP, GERP++, mirSVR, targetScan)
CHASMplus	Missense	Somatic missense mutations from TCGA, hotspots, germline variation, sequence conservation, molecular function annotations, protein–protein interface annotations, sequence-biased regions, gene-level covariates	https://run.opencravat.org	No
Evolutionary Action	Missense	Sequence conservation	http://eaction.lichtargelab.org/	No
Mutation Assessor	Missense	Sequence conservation	http://mutationassessor.org/r3/	No
MutPred2.0	Missense, inframe indels (MutPredIndel), frameshifting indels and stop gains (MutPred-LOF)	Reference protein sequence, sequence conservation, homology, residue function, clinical annotations	http://mutpred.mutdb.org/index.html	No
PhD-SNPg	Single-nucleotide variants, inframe indels, non-coding variants	Reference protein sequence, sequence conservation	https://snps.biofold.org/phd-snpg/index.html	No
Primate-AI	Missense	Reference protein sequence, sequence conservation, population frequency, predicted protein secondary structure	https://github.com/Illumina/PrimateAI	No
PolyPhen-2	Missense	Reference protein sequence, sequence conservation, protein functional annotations, protein 3D structure	http://genetics.bwh.harvard.edu/pph2/	No
PROVEAN	Missense	Sequence conservation	http://provean.jcvi.org/genome_submit_2.php?species=human	No
REVEL	Missense	Sequence conservation, clinical annotations, other predictors	https://sites.google.com/site/revelgenomics	Yes (VEST, FATHMM, MutPred, PolyPhen, SIFT, PROVEAN, Mutation Assessor, MutationTaster, LRT, GERP, SiPhy, phyloP, phastCons)
SIFT	Missense, inframe indels	Sequence conservation	https://sift.bii.a-star.edu.sg/	No
VEST-4	Missense, inframe indels (VEST-indel)	Reference protein sequence, sequence conservation, clinical annotations, protein functional annotations, predicted protein secondary structure	https://run.opencravat.org	No

Indel, insertion and/or deletion.

and training examples multiple times, meta-predictors are at risk of overfitting and bias[7]. Efficient access in a centralized location to training data and feature details used by predictors would allow the creators of meta-predictors to identify these overlaps and to develop improved methods that result in less overfitting.

Many users find it difficult to interpret typical outputs of prediction methods, which often produce categorical results, such as 'pathogenic' or 'benign'[4]. However, these terms have specific definitions, and variant classification is intended to be based on careful evaluation of multiple criteria in relation to a particular condition or disease. The raw output of most prediction methods is a continuous value or 'score,' but many methods select thresholds on these scores, and it is important to understand or know how these thresholds were chosen. Other methods provide measures of statistical significance,

so that each score is associated with an adjusted *P* value, allowing the user to pick their own thresholds based on tolerance to false positives[4]. However, not all users might be confident in having to make such a choice. Thus, a transparent approach to calibrate predictor output is needed, such as the framework that mapped predictor scores onto nine non-overlapping intervals, corresponding to the nine evidential strength levels defined by the ACMG/AMP guidelines[4]. Widespread adoption of this and other calibration efforts[8] would be an important advance toward facilitating the clinical application of prediction methods.

Finally, it is important for a predictor and its calibration to be vetted repeatedly, independently and in a blinded setting. Errors in training protocols, the iterative nature of data access or simply biased training data can contribute to overly optimistic estimates of prediction performance that can propagate into training data for subsequent next-generation models and increase downstream circularity. Performance of predictors may also differ on the basis of the characteristics of a specific gene or disease, the germline versus somatic context and the type of genetic alterations (missense, insertions and/or deletions, splicing, non-coding, and so on)[8–10]. Community experiments such as the Critical Assessment of Genome Interpretation (CAGI) serve to inform the community of predictor quality, as well as provide valuable feedback to method developers and dataset providers[8]. CAGI is modeled on the Critical Assessment of Structure Prediction program[11], which was instrumental in promoting the development of algorithms to predict protein structures from protein sequences. More information on CAGI can be found at https://genomeinterpretation.org.

Conclusions

Here, we have discussed the need for increased transparency with respect to the inputs and outputs of variant effect predictors, which we believe is achievable in the near term. However, some computational predictors that use machine learning, including deep neural networks, further lack transparency with respect to their internal reasoning. Such methods are sometimes referred to as 'black boxes,' because their decision-making is encoded in parameters that are difficult to interrogate, not interpretable by humans and their outputs are not explainable. Interpretability and explainability of machine learning results are subject to intensive research in computer science, building in part upon methods such as attention models and interpretable queries[12], which could be incorporated into the next generation of methods for variant effect prediction[13].

We hope that this Comment will help to bring attention to current limitations in variant prediction tools, highlight the necessity to outline such limitations in recommendations for interpretation of sequence variants and stimulate efforts to make necessary improvements. Providing the name, value and source of each feature used by computational predictors, in addition to currently provided output, could be a valuable first step in such efforts. Another crucially important improvement for computational predictors would be to specify if a prediction indicates LOF or GOF, as such information is vital for proper variant classifications and clinical interpretations. With increased transparency and computational rigor, computational tools will help to provide more informed clinical decisions and more reliable variant pathogenicity classifications, benefiting both practitioners and patients in the realm of personalized medicine. Thus, despite the challenges ahead with regard to enhancing transparency of computational tools for variant effect prediction, they have the potential to improve genomic medicine.

Rachel Karchin [1,2] ✉, **Predrag Radivojac** [3], **Anne O'Donnell-Luria** [4,5,6], **Marc S. Greenblatt**[7], **Michael Y. Tolstorukov** [8] **& Dmitriy Sonkin** [9] ✉

[1]The Institute for Computational Medicine, The Johns Hopkins University, Baltimore, MD, USA. [2]Departments of Biomedical Engineering, Oncology, and Computer Science, The Johns Hopkins University, Baltimore, MD, USA. [3]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA. [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [5]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. [6]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [7]Department of Medicine and University of Vermont Cancer Center, University of Vermont, Larner College of Medicine, Burlington, VT, USA. [8]Department of Informatics and Analytics, Dana-Farber Cancer Institute, Boston, MA, USA. [9]National Cancer Institute, Division of Cancer Treatment and Diagnosis, Rockville, MD, USA.
✉e-mail: karchin@jhu.edu; dmitriy.sonkin@nih.gov

Published online: 2 July 2024

References

1. Richards, S. et al. *Genet. Med.* **17**, 405–424 (2015).
2. Horak, P. et al. *Genet. Med.* **24**, 986–998 (2022).
3. Li, M. M. et al. *J. Mol. Diagn.* **19**, 4–23 (2017).
4. Pejaver, V. et al. *Am. J. Hum. Genet.* **109**, 2163–2177 (2022).
5. Landrum, M. J. et al. *Nucleic Acids Res.* **44**, D862–D868 (2016).
6. Stenson, P. D. et al. *Hum. Genet.* **139**, 1197–1207 (2020).
7. Whalen, S., Pandey, O. P. & Pandey, G. *Methods* **93**, 92–102 (2016).
8. The Critical Assessment of Genome Interpretation Consortium. et al. *Genome Biol.* **25**, 53 (2024).
9. Riccio, C., Jansen, M. L., Guo, L. & Ziegler, A. *Hum. Genet.* **143**, 625–634 (2024).
10. Kuksa, P. P. et al. *Hum. Mol. Genet.* **31**, R62–R72 (2022).
11. Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. *Proteins* **23**, ii–v (1995).
12. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. *Proc. Natl Acad. Sci. USA* **116**, 22071–22080 (2019).
13. Abe, S. et al. *Cancers* **15**, 1118 (2023).

Author contributions

D.S. conceptualized the manuscript. D.S. and R.K. wrote the initial draft of the manuscript. All authors participated in the manuscript editing.

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Genetics* thanks Andreas Ziegler and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.