Toward trustable use of machine learning models of variant effects in the clinic

Mafalda Dias,[1,2,5] Rose Orenbuch,[3] Debora S. Marks,[3,4,5] and Jonathan Frazer[1,2,5,*]

There has been considerable progress in building models to predict the effect of missense substitutions in protein-coding genes, fueled in large part by progress in applying deep learning methods to sequence data. These models have the potential to enable clinical variant annotation on a large scale and hence increase the impact of patient sequencing in guiding diagnosis and treatment. To realize this potential, it is essential to provide reliable assessments of model performance, scope of applicability, and robustness. As a response to this need, the ClinGen Sequence Variant Interpretation Working Group, Pejaver et al., recently proposed a strategy for validation and calibration of in-silico predictions in the context of guidelines for variant annotation. While this work marks an important step forward, the strategy presented still has important limitations. We propose core principles and recommendations to overcome these limitations that can enable both more reliable and more impactful use of variant effect prediction models in the future.

The recent article by the ClinGen Sequence Variant Interpretation (SVI) Working Group titled "Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria"[1] reflects a change in attitude toward the value of computational tools for annotating genetic variants. The clinical community has been cautious to adopt in-silico predictions, in part due to the limited accuracy of early models but also due to historical over-optimism in model performance driven by flawed validation.[2] Indeed, until now, those following the guidelines put forward by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP) would treat all computational methods as a single source of "supporting" evidence—the weakest category.[2] Now, acknowledging the substantial progress in model accuracy in recent years, the ClinGen SVI group presents a calibration scheme to enable some models, given sufficient validation and subsequent recalibration, to provide "strong" evidence. However, despite presenting a practical way forward, the proposed guidelines still have severe limitations. The ClinGen SVI recommendations for mapping model scores to evidence weights, "supporting," "strong," etc., are centered on assessing a model's performance at predicting benign and pathogenic labels from ClinVar.[3] Importantly, these labels are pooled in the sense that they come from all available genes. In this brief commentary, we highlight some missed opportunities and pitfalls with this approach both in terms of implementation and sustainable future model integration. We discuss core principles for the trustable and efficient use of computational models and detail four areas where we as a community can improve their use in variant annotation.

Basic principles for reliable use of computational models

At the heart of our concerns and recommendations for possible ways forward reside two principles.

The assessment of model performance should mimic the downstream application as closely as possible. For example, in the case of variant annotation, it is important to distinguish if the task is to annotate previously unannotated variants in a disease-associated gene or if the model must be capable of generalizing to any gene, including those whose role in disease is not yet known. It is possible for a model to perform well at one of these tasks and not the other. Thus, to reliably assess the performance of a model on a specific task, the test must resemble that task as closely as possible.

Integrating model predictions with other sources of evidence requires consideration of what the model was trained on. The ACMG-AMP criteria for variant interpretation provides guidelines on how to combine multiple independent sources of evidence. As such, it is crucial to know exactly which data were used in training to avoid "double dipping." On the other hand, if two models are trained on distinct sources of evidence, it should be possible to incorporate both independently. More generally, to make the most effective use of a model requires consideration of what the model was trained on.

Four areas where assessment and calibration of variant effect prediction tools can be improved

The ClinGen approach to measuring model performance is based on predicting ClinVar labels pooled across genes. While this approach is ubiquitous in the literature, in the spirit of our first principle above, we want to stress that it is inadequate for all downstream tasks for which these

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain; [2]University Pompeu Fabra, Barcelona, Spain; [3]Department of Systems Biology, Harvard Medical School, Boston, MA, USA; [4]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[5]Senior author
*Correspondence: dias.frazer.marks@gmail.com

https://doi.org/10.1016/j.ajhg.2024.10.011.

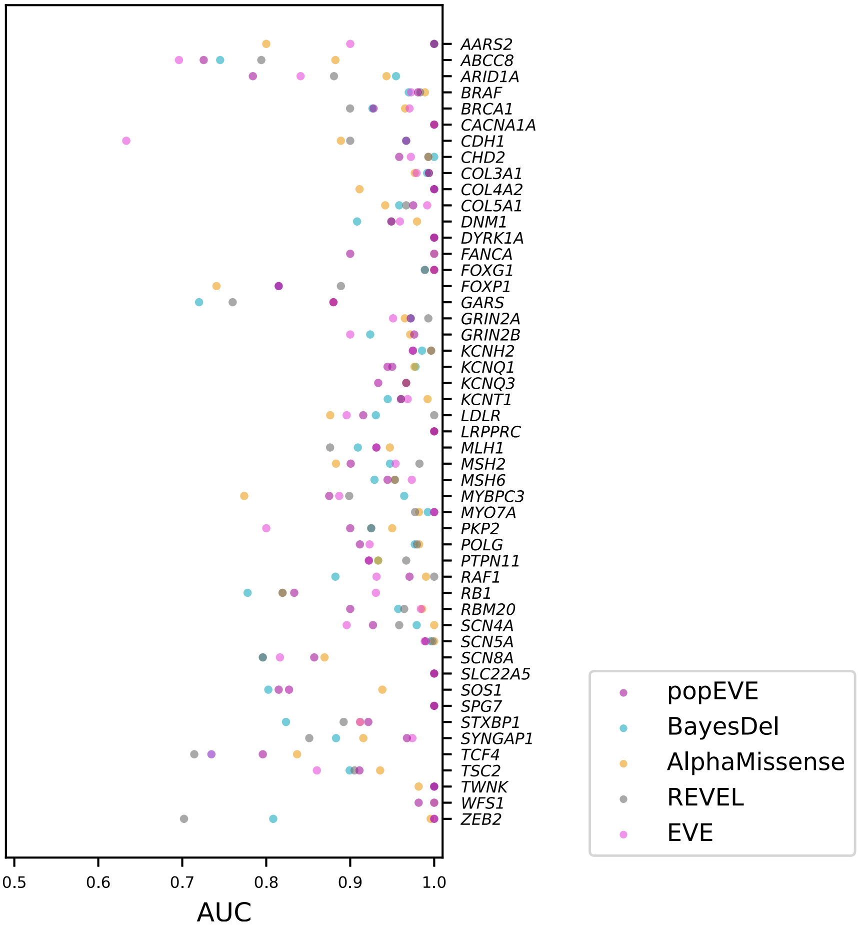

Figure 1. Performance of variant effect predictors varies across genes
Performance of state-of-the-art predictors[4,8–11] at separating benign from pathogenic variants from the ClinGen 2019 validation set[1] in the subset of genes for which there are at least five benign and five pathogenic variants. Performance is evaluated with the area under the receiver operating curve (AUC).

models are currently used. In what follows, we will use this assessment as a pertinent example to illustrate four points where model validation and calibration can be improved.

(1) Model performance varies across genes, so validation should be gene by gene where possible.

A basic observation we[4] and others[5,6] have made is that the impact of variants is more easily modeled in some genes than others. A model may perfectly separate benign and pathogenic variants in some genes and yet make numerous false-positive or false-negative predictions in others (Figure 1). There are many possible reasons for this. For example, how well a gene can be modeled depends on the number of sequences from that family available for training; models that learn from clinical annotations have more data for well-studied genes; models that use structural information only have these data for some genes; models that use deep mutational scans have more clinically relevant assays for some genes, and so on. In addition, some genes, depending on their product, structure and function, will be fundamentally simpler to model than others regardless of the training data or model's complexity. Taken together, this means that testing the performance with labels pooled across genes will both overestimate the performance in some parts of the proteome while underestimating performance in others. One ramification of this was recently discussed in Tejura et al. where the authors explored the consequences of inferring evidence weights based on pooled labels and found that inconsistencies arise as a result of heterogeneous model performance across genes.[7]

Recommendation: provided there are enough publicly known variant annotations to do so, it would be preferable to assess the performance on a gene-by-gene basis (or even domain by domain).

Of course, a major limitation of assessing variants on a gene-by-gene basis (or domain by domain) is that there are many genes for which there are insufficient labels to do so. This brings us to our next point.

(2) Data available for validation varies across genes, and this is a further reason to assess models gene by gene.

ClinVar data is sparse—only a very small fraction (~2%) of all observed missense variants have any clinical annotation.[3] Furthermore, about 50% of all missense (likely) pathogenic labels reside in about 5% of genes known to be involved in disease, which corresponds to less than 1% of all human genes; see Figure 2A. Since these clinical labels are biased toward a small number of genes, as well as being sparse, estimation of model performance is significantly more accurate in some genes than others. For instance, for a gene like *TP53* (MIM: 191170) for which there are close to 350 missense variants with clinical annotations, one can get a much better estimate of the model's predictive performance than, say, for *KCNJ10* (MIM: 602208), which only has 16 labels.

This label bias is problematic when performing validation with labels

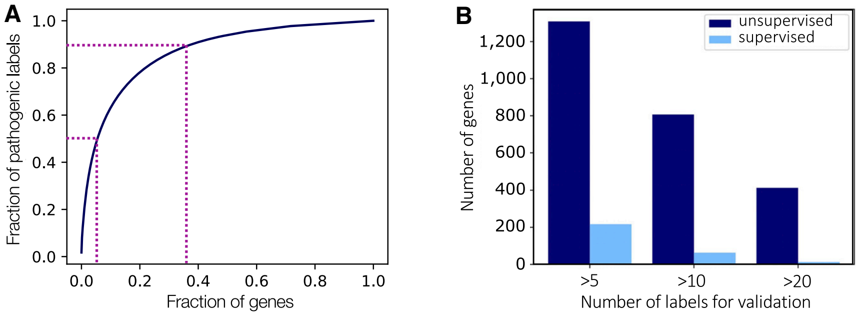

Figure 2. Bias and sparsity of clinical labels
(A) Clinical labels are not uniformly distributed across genes. Here, we show the fraction of pathogenic or likely pathogenic labels in ClinVar[3] as a function of the fraction of genes known to be involved in disease (~3,000 genes in total). 50% of the (likely) pathogenic variants are in 5% of the genes, and 90% of the variants are in just 38% of the genes.
(B) Many more genes are available for validation of unsupervised models compared to supervised models, as their clinical annotations were not used in training. Here, we show the mean number of genes that have sufficient labels for validation (5 or more, 10 or more, 20 or more), for unsupervised (dark blue) and supervised (light blue) models. We simulate a 90% train 10% test random split of all labels in ClinVar for the supervised models.[3]

pooled across all genes, as both model performance and the ability to measure model performance may be biased as a consequence, and worse still, correlated. This is particularly relevant for models supervised on clinical labels, as this test does not evaluate whether or not the model generalizes to regions of the proteome where few or no labels were seen in training.

Recommendation: the mapping of model scores to evidence weights for incorporating in schemes similar to the ACMG-AMP criteria should be done on a gene-by-gene basis whenever possible, and the weights should not only reflect the model performance but also the number of labels used to estimate this performance.

In the case of genes for which only a small number of labels currently exist, pooling labels across genes may be the best option. However, the resulting evidence weight should reflect the uncertainty in this performance estimate. Approaches like pooling across related genes only, and reweighting or downsampling labels, could help minimize the biases toward certain genes.

(3) Despite progress, circularity issues remain both when assessing model performance and when using models within ACMG-AMP guidelines.

A problem that plagues the use of machine learning in many areas of biology is circularity. Most models and approaches to assessing performance make the assumption that the data used in training and testing is independent and identically distributed (iid), but this is not the case for natural sequences nor clinical labels, which have a rich phylogenetic structure, as well as various acquisition biases. The ClinGen SVI group made considerable effort to remove some sources of circularity from their assessment; however, others remain. For instance, variants used for training the models assessed were removed from the validation set, but not those at the same position as those in training, potentially leading to inflated performance estimates. This is what is sometimes referred to as type 1 circularity[5,6]—variant-level leakage when the same variants (or homologous) are seen in both training and testing. When pooling variants across genes for validation, circularity also occurs if information leakage takes place at the gene-level, i.e., when variants of the same (or homologous) gene are seen in both training and testing. This so-called type 2 circularity[5,6] is intrinsically related to the gene-level biases described in point 2.

Recommendation: for supervised models, when designing a test/train split of the data, it is important to consider if test variants are at the same position as those seen in training, in the same domain as those in training, or in the same (or homologous) protein as those in training. In fact, the first of these points relates to ACMG-AMP criterion PM5—which states that a known pathogenic variant at a given position may be regarded as moderate evidence for there being other pathogenic variants at that same position.

A perhaps even more serious circularity problem comes from the type of evidence and data that went into creating the labels used for assessing the performance in the first place. For example, as discussed by Livesey and Marsh,[5] it is circular to use a label established by leveraging population frequencies (criteria BS1 or PM2 of the ACMG-AMP criteria) to assess the performance of a model trained on the same population frequencies. This is true for all evidence, including ACMG-AMP criterion PM5 described above, as well as, critically, computational evidence itself. As we move into an era where computational models play an increasingly important role in variant classification, this form of circularity will become increasingly problematic unless we take measures to counter this.

Recommendation: to alleviate this issue, ClinVar should keep track of what sources of evidence are known to support the status of a given variant. One could then simply consider the strength of evidence supporting a variant in the absence of information the model will see at train time and build training and test sets accordingly.

(4) When integrating models with other evidence, greater emphasis should be put on the data used in training.

Finally, we would like to argue that the handling of models within pipelines such as those proposed by ACMG-AMP should reflect the sources of evidence used during training rather than relying on blanket criteria for "computational" or "in-silico" methods. This is for two reasons: (1) to avoid "double dipping"—independent sources of evidence should be accounted for only once, so which data were used for model building needs to be explicitly acknowledged; Pejaver et al.[1] also comment on this problem. So, for example, criteria PS1/PM5 should not be used in conjunction with a model that saw a label at that same position during training, and similarly criteria BS1/PM2 should not be used in conjunction with models trained on population frequencies; and (2) to allow for the effective use of combinations of models—if two models are trained with independent data, for instance, multiple sequence alignments from diverse species and deep mutational scanning data, then they provide independent evidence, whereas if two models are trained on exactly the same data, their consensus, or lack thereof, is informative as one source of evidence.

Recommendation: the weight with which computational model predictions should be integrated with other sources should not only depend on model performance but also reflect the evidence used in training.

Final thoughts

Many of the difficulties presented in this commentary are alleviated for models that have not used clinical labels as part of model building—referred to as unsupervised. First, not using any labels in training means that there are typically almost ten times more labels for validation, enabling more rigorous testing and calibration on a gene-by-gene basis for vastly more genes; see Figure 2B. Second, the issue of generalization to genes with fewer labels is removed entirely (quality of available validation, however, does of course remain an issue). Third, it is easy to track what information has been used in training, making integration with ACMG-AMP-style criteria straightforward and with minimal risk of "double dipping." We therefore see unsupervised modeling as a promising approach, which in fact currently delivers state-of-the-art performance.[4,8,9,12]

We would like to stress that the recommendations presented in this commentary, in particular the proposal for gene-by-gene assessments, need not be thought of as replacements for the approach suggested by the ClinGen SVI group but rather as opportunities for further refinement. For example, the ClinGen SVI group uses a Bayesian model, which could be augmented via a hierarchical strategy such that the calibration procedure they propose gets modified on a gene-by-gene basis, based on what labels are available, but reverts to something resembling the current model for genes with few or no labels.

Finally, we would like to discuss the limitations of training and assessing models with clinical labels and raise the question of whether we, as a community, are at risk of falling victim to Goodhart's law—"when a measure becomes a target, it ceases to be a good measure."[13] While classifying variants as benign or pathogenic might be of great value for diagnosis and guiding treatment, this binary classification is a false dichotomy that obscures many aspects of the role of genetic variation in disease. Fundamental questions for clinical genetics—such as how severe is the disease associated with a variant and how likely is its onset, how confident is the model about its predictions, and how do other variants impact these conclusions—cannot be addressed by models that have been optimized to place variants into two classes. At best, the scores from such models conflate these points, which can result in model pathologies such as overprediction of pathogenic variants and poor performance when assessing variants with milder effects.[1] However, in combination with the increasing amount of genotype-to-clinical-phenotype data, deep learning may be uniquely suited to addressing these questions. With careful consideration of training data and benchmarks, we have a great opportunity to develop in-silico variant predictions that may transform clinical genetics. With the principles and recommendations laid out in this commentary, we aim to highlight the avenues for further progress in ensuring the responsible and reliable use of such tools and in this way enhance their potential to impact our understanding of human disease.

Data and code availability

This study did not generate datasets or code.

Acknowledgments

We thank all members of the Marks Lab and Dias & Frazer Group for valuable discussions. D.M. is supported by a Chan Zuckerberg Initiative Award (Neurodegeneration Challenge Network, CZI2018-191853) and an NIH Transformational Research Award (TR01 1R01CA260415). M.D. and J.F. are supported by the Spanish Ministry of Science and Innovation (PID2022-140793NA-I00). We are grateful to the CRG Core Technologies Programme for their support and assistance.

Declaration of interests

The authors declare no competing interests.

Web resources

ClinGen 2019, www.cell.com/cms/10.1016/j.ajhg.2022.10.013/attachment/67f2dc84-3e05-4502-9df0-70ec5499b17e/mmc2.xlsxpopEVE mutation portal, pop.evemodel.org

ClinVar, ncbi.nlm.nih.gov/clinvar

dbNSFP, sites.google.com/site/jpopgen/dbNSFP

References

1. Pejaver, V., Byrne, A.B., Feng, B.-J., Pagel, K.A., Mooney, S.D., Karchin, R., O'Donnell-Luria, A., Harrison, S.M., Tavtigian, S.V., Greenblatt, M.S., et al. (2022). Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am. J. Hum. Genet. *109*, 2163–2177. https://doi.org/10.1016/j.ajhg.2022.10.013.

2. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). 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*, 405–424. https://doi.org/10.1038/gim.2015.30.
3. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*, D1062–D1067. https://doi.org/10.1093/nar/gkx1153.
4. Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J.K., Brock, K., Gal, Y., and Marks, D.S. (2021). Disease variant prediction with deep generative models of evolutionary data. Nature *599*, 91–95. https://doi.org/10.1038/s41586-021-04043-8.
5. Livesey, B.J., and Marsh, J.A. (2024). Variant effect predictor correlation with functional assays is reflective of clinical classification performance. bioRxiv, 2024.05.12.593741. https://doi.org/10.1101/2024.05.12.593741.
6. Grimm, D.G., Azencott, C.-A., Aicheler, F., Gieraths, U., MacArthur, D.G., Samocha, K.E., Cooper, D.N., Stenson, P.D., Daly, M.J., Smoller, J.W., et al. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Mutat *36*, 513–523. https://doi.org/10.1002/humu.22768.
7. Tejura, M., Fayer, S., McEwen, A.E., Flynn, J., Starita, L.M., and Fowler, D.M. (2024). Calibration of variant effect predictors on genome-wide data masks heterogeneous performance across genes. Am. J. Hum. Genet. *111*, 2031–2043. https://doi.org/10.1016/j.ajhg.2024.07.018.
8. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science *381*, eadg7492. https://doi.org/10.1126/science.adg7492.
9. Orenbuch, R., Kollasch, A.W., Spinner, H.D., Shearer, C.A., Hopf, T.A., Franceschi, D., Dias, M., Frazer, J., and Marks, D.S. (2023). Deep generative modeling of the human proteome reveals over a hundred novel genes involved in rare genetic disorders. medRxiv, 2023.11.27.23299062. https://doi.org/10.1101/2023.11.27.23299062.
10. Feng, B.-J. (2017). PERCH: A Unified Framework for Disease Gene Prioritization. Mutat *38*, 243–251. https://doi.org/10.1002/humu.23158.
11. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am. J. Hum. Genet. *99*, 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016.
12. Gao, H., Hamp, T., Ede, J., Schraiber, J.G., McRae, J., Singer-Berk, M., Yang, Y., Dietrich, A.S.D., Fiziev, P.P., Kuderna, L.F.K., et al. (2023). The landscape of tolerated genetic variation in humans and primates. Science *380*, eabn8153. https://doi.org/10.1126/science.abn8197.
13. Goodhart, C.A.E. (1984). Problems of Monetary Management: The UK Experience, pp. 91–121. https://doi.org/10.1007/978-1-349-17295-5_4.