



Protein Language Models in Variant Effect Prediction: Strengths, Failure Modes, and Benchmarks

Gabriela Mthembu^{1*}, Phong Barbosa², Ivana Zulu³

¹Department of Life Sciences, Zambezi Center for Life Sciences, Guyana

²Department of Life Sciences, Pearl River Academy of Life Sciences, Nigeria

³Department of Life Sciences, Periyar School of Life Sciences, South Africa

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijbc.2026.v1.i1.pp7-13](https://doi.org/10.62715/ijbc.2026.v1.i1.pp7-13)

Received: 07 March 2026

Accepted: 17 May 2026

Published: 30 July 2026

Abstract

Protein language models trained on large unlabelled sequence databases assign a likelihood to every amino acid at every position, and that likelihood has been adopted as a score for the effect of a missense substitution. The approach needs no alignment built for the protein in question and no labelled examples, which has made it attractive for the interpretation of variants of uncertain significance. This narrative review surveys what the published literature reports about these models when used in that way. It describes the scoring schemes in common use, sets them beside conservation measures and alignment-based generative models, and summarises the evidence from deep mutational scanning collections and from community benchmarks. Recurring failure modes are examined: shallow or biased sequence families, the conflation of molecular function with organismal fitness, insensitivity to cellular context such as expression and degradation, and the circularity that arises when evaluation sets are drawn from the same clinical databases that shaped earlier tools. The review concludes that these models are strong, general priors on tolerance to substitution rather than predictors of pathogenicity, and that benchmark design, calibration and clear statements of scope matter more than further gains in reported ranking accuracy.

Keywords: Protein language models, Variant effect prediction, Deep mutational scanning, Benchmarking, Variant interpretation, Zero-shot prediction

© 2026 Gabriela Mthembu et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

Sequencing a human exome returns many missense changes whose consequences are unknown. Most are never seen again in a clinical database and most have no functional measurement, so interpretation falls back on computation. For two decades the computational contribution came from conservation: a position that has not changed across evolution is assumed to matter, an idea implemented in tools such as SIFT (Ng and Henikoff, 2003) and PolyPhen-2 (Adzhubei et al., 2010) and placed in the supporting evidence tier of the interpretation guidelines issued by Richards et al. (2015).

Protein language models have changed the computational input. Trained with a masked or

* Corresponding author: Gabriela Mthembu, Department of Life Sciences, Zambezi Center for Life Sciences, Guyana. E-mail: gabriela.mthembu@zambezi.edu

© 2026 Gabriela Mthembu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

autoregressive objective on tens or hundreds of millions of sequences, they learn a distribution over amino acids at each position conditioned on the rest of the sequence. [Rives et al. \(2021\)](#) showed that representations learned this way encode secondary structure, contacts and remote homology without ever being told about them, and the later ESM-2 models of [Lin et al. \(2023\)](#) extended the scale. [Meier et al. \(2021\)](#) made the step that concerns this review: scoring a substitution by the model likelihood of the mutant residue relative to the wild type, with no task-specific training at all.

That proposal has been taken up quickly in both research and clinical settings, and the claims made for it range from a useful prior to something close to a pathogenicity predictor. This article reviews the published evidence. It is a narrative review of the modelling, benchmarking and clinical interpretation literature, written to separate what these models reliably provide from what is assumed of them, and it does not report a systematic search or any new evaluation.

2. Scope and Approach

The review covers models trained on protein sequence alone and used to score single amino acid substitutions, together with the alignment-based generative models they are usually compared against and the assay data used to judge them. Structure prediction is discussed only where structural output is used as a variant signal. Regulatory, splicing and copy number variation lie outside the scope, as do protein design applications. Work is grouped by the question it answers rather than by date, and preference is given to papers that report evaluation against experimental measurements rather than against curated clinical labels alone.

3. How a Language Model Scores a Variant

Three scoring schemes dominate. The masked marginal score replaces the position of interest with a mask token, reads off the model distribution over the twenty amino acids, and takes the difference in log probability between mutant and wild type. The wild-type marginal score skips the masking and reads the distribution from the unmodified sequence, which is cheaper and usually slightly worse. The pseudo likelihood score mutates the sequence and sums log probabilities across all positions, capturing the effect of the substitution on the rest of the sequence. [Meier et al. \(2021\)](#) compared these schemes and reported the masked marginal to be the most reliable of the simple options.

The comparison class matters as much as the scoring rule. Alignment-based generative models fit a distribution to the homologous family rather than to all known protein sequence. Couplings models of the kind described by [Hopf et al. \(2017\)](#) capture pairwise epistasis within a family, the variational model of [Riesselman et al. \(2018\)](#) captures higher-order dependence, and the clinically oriented model of [Frazer et al. \(2021\)](#) adds a calibrated two-component mixture to turn the continuous score into a class. Hybrid approaches retrieve alignment statistics at inference time and combine them with an autoregressive language model, as in the method of [Notin et al. \(2022\)](#). Table 1 sets out the families and their characteristic weaknesses.

4. What the Models Do Well

The strongest and best replicated result is that a language model likelihood correlates with measured functional effect across a wide range of proteins without any protein-specific fitting. Deep mutational scanning, introduced as a general technique by [Fowler and Fields \(2014\)](#), produces thousands of measured substitutions for a single protein, and collections of such experiments have become the standard yardstick. [Livesey and Marsh \(2020\)](#) used them to compare predictors on a footing independent of clinical labels, and the assembled benchmark of [Notin et al. \(2023\)](#) extended that comparison across many assays and model families.

Generality is the practical strength. A conservation tool needs an alignment assembled for the protein at hand, and a supervised model needs labels; a language model needs only the sequence. For proteins with deep, diverse families the advantage over alignment-based generative models is modest, which several comparisons report, but the language model loses less ground when the family is shallow because

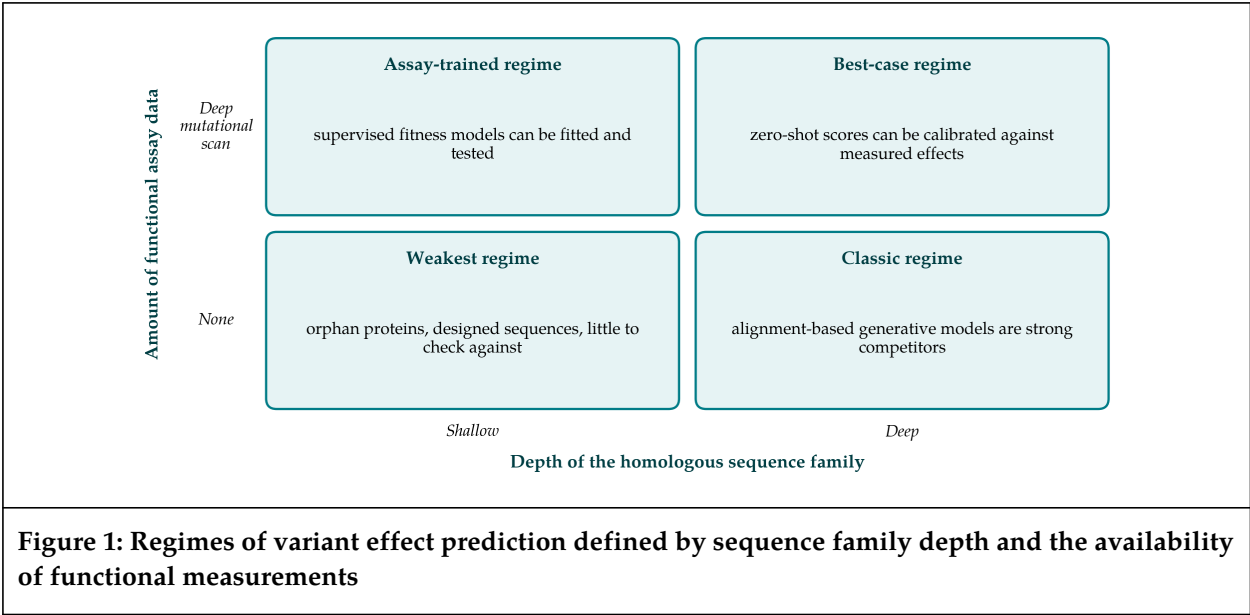
it has learned general constraints on protein sequence from everything else it saw. Figure 1 shows the two axes, family depth and assay availability, that together determine which approach is likely to help.

Table 1: Classes of missense variant effect predictor and the signal each one uses		
Class	Signal used	Known weakness
Position conservation	Column entropy in an alignment	Ignores epistasis; weak at tolerated but rare substitutions (Ng and Henikoff, 2003)
Supervised classifier on features	Curated labels plus structural features	Type 1 and type 2 circularity in evaluation (Grimm et al., 2015)
Family generative model	Dependencies within one alignment	Fails when the family is shallow (Riesselman et al., 2018)
Protein language model, zero shot	Likelihood under a global sequence model	Confuses rarity with damage; no cellular context (Meier et al., 2021)
Retrieval-augmented language model	Global model plus family statistics	Inherits both failure modes (Notin et al., 2022)
Structure-informed ensemble	Predicted structure plus sequence likelihood	Calibration depends on the training labels (Cheng et al., 2023)

A second strength is that the scores are cheap and exhaustive. A full substitution matrix for every position in a protein can be produced in one pass, which supports triage at proteome scale. The proteome-wide resource described by Cheng et al. (2023) combines this with structural context from the folding model of Jumper et al. (2021), and the structure database assembled by Varadi et al. (2022) makes such combinations routine.

5. Failure Modes

Four failure modes recur in the literature, and they are not simply matters of accuracy.



5.1. Rarity is not damage

A language model scores how expected a residue is given the sequence context, which is a statement about the sequences in the training set and not about molecular function. Residues that are rare because of phylogenetic sampling, codon usage or gene duplication history receive low likelihood whether or not

they impair the protein. The converse also holds: a substitution common in a bacterial orthologue may be scored as acceptable in a human protein where the surrounding network differs. The underlying constraint learned is tolerance across evolution, which is correlated with function and not identical to it.

5.2. Family depth and bias

Performance degrades for proteins with few homologues, for fast-evolving families and for regions that are intrinsically disordered. Sequence databases are also unevenly sampled across the tree of life, so some human proteins sit in families dominated by a narrow set of lineages. Retrieval-based hybrids (Notin *et al.*, 2022) and the constraint information in population datasets such as the one described by Karczewski *et al.* (2020) can partly compensate, but neither creates evolutionary signal where none exists.

5.3. Missing cellular context

A substitution can be benign for the folded enzyme and still cause disease through reduced expression, mislocalisation, altered degradation or disrupted interaction with a partner. Nothing in a sequence likelihood addresses dosage, tissue-specific expression or haploinsufficiency. Equally, a substitution that abolishes catalysis in vitro may be tolerated in vivo for a protein with redundancy. Functional assays have the same limitation, which is why the guidance of Brnich *et al.* (2020) insists that an assay be shown to model the disease mechanism before its output is used as evidence.

5.4. Circular evaluation

Evaluating any predictor on clinical labels invites circularity. Grimm *et al.* (2015) described two forms: training and test variants drawn from the same genes, and datasets in which almost all variants in a gene share one label, so that a predictor can score well by learning the gene rather than the variant. Labels in public archives such as the one maintained by Landrum *et al.* (2018) were themselves assigned with computational support, which compounds the problem. Assay-based evaluation avoids it, at the cost of covering few proteins.

6. Benchmarks and What They Can Settle

Benchmarking in this area has improved considerably. Early comparisons used heterogeneous clinical sets; current practice favours collections of deep mutational scans with standardised processing, of which the resource described by Notin *et al.* (2023) is the most widely used, alongside earlier protein transfer learning benchmarks (Rao *et al.*, 2019) that established the format. These allow a ranking question to be answered: given two predictors, which orders substitutions within a protein more faithfully against measurement.

Several questions remain outside what such benchmarks settle. They say little about calibration, since a rank correlation is unchanged by any monotone transformation of the score, yet a clinical user needs a threshold with known error rates. They weight proteins by which ones happen to have been assayed, and those tend to be small, soluble and of established interest. And because each assay measures one property in one system, aggregate performance across assays conflates fitness, stability, binding and abundance. Reporting per-assay results with the measured property named is more informative than a single averaged figure.

For clinical use the framework of Richards *et al.* (2015) assigns computational predictions supporting weight only, and subsequent calibration work has examined how much weight a well-validated predictor can carry. Starita *et al.* (2017) set out the case for functional assays as the stronger evidence type, and the recommendations of Brnich *et al.* (2020) describe how an assay is validated for that purpose. A language model score is best understood as input to that process rather than as a substitute for it.

7. Gaps and Future Work

Three gaps stand out. The first is calibration to interpretable quantities. Converting a log odds ratio into a likelihood ratio with stated performance, as attempted by the mixture formulation of Frazer *et al.* (2021), is more useful than another increment in rank correlation. The second is coverage beyond the single

substitution: insertions, deletions, multi-residue changes and frame-preserving rearrangements are common in patients and poorly handled.

The third is the integration of modalities that sequence cannot carry. Predicted structure, binding interfaces, post-translational modification sites, expression and population constraint all inform tolerance, and the current practice of concatenating them into an ensemble is pragmatic rather than principled. Representation learning across sequence and structure jointly, of the kind the models of [Lin et al. \(2023\)](#) and [Elnaggar et al. \(2022\)](#) begin to approach, is a natural direction, provided that evaluation keeps pace and stays independent of clinical labels.

8. Conclusion

Protein language models supply a general, inexpensive and genuinely zero-shot prior on which amino acid substitutions a protein will tolerate, and on assay-based benchmarks they match or exceed alignment-based methods across a wide range of proteins. Their weaknesses follow from what they are. They score how expected a residue is, not how damaging it is; they lose power where sequence families are shallow; they know nothing of expression, dosage or cellular context; and evaluation against clinical labels remains vulnerable to circularity. Treating these scores as a well-characterised prior, reported with calibration and with an explicit statement of the regime in which they were validated, is the use the evidence supports.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010). [A method and server for predicting damaging missense mutations](#). *Nature Methods*, 7(4): 248-249.
- Brnich, S.E., Abou Tayoun, A.N., Couch, F.J., Cutting, G.R., Greenblatt, M.S., Heinen, C.D., Kanavy, D.M., Luo, X., McNulty, S.M., Starita, L.M., Tavtigian, S.V., Wright, M.W., Harrison, S.M., Biesecker, L.G. and Berg, J.S. (2020). [Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework](#). *Genome Medicine*, 12(1): 3.
- Cheng, J., Novati, G., Pan, J., Bycroft, C., Zemgulyte, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., Schneider, R.G., Senior, A.W., Jumper, J., Hassabis, D., Kohli, P. and Avsec, Z. (2023). [Accurate proteome-wide missense variant effect prediction with AlphaMissense](#). *Science*, 381(6664): eadg7492.
- Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D. and Rost, B. (2022). [ProtTrans: Toward understanding the language of life through self-supervised learning](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 7112-7127.
- Fowler, D.M. and Fields, S. (2014). [Deep mutational scanning: A new style of protein science](#). *Nature Methods*, 11(8): 801-807.
- 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(7883): 91-95.
- 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., Duncan, L.E. and Borgwardt, K.M. (2015). [The evaluation of tools used to predict the effect of missense variants is hindered by two types of circularity](#). *Human Mutation*, 36(5): 513-523.

- Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Scharfe, C.P.I., Springer, M., Sander, C. and Marks, D.S. (2017). [Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35\(2\): 128-135.](#)
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P. and Hassabis, D. (2021). [Highly accurate protein structure prediction with AlphaFold. *Nature*, 596\(7873\): 583-589.](#)
- Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., Gauthier, L.D., Brand, H., Solomonson, M., Watts, N.A., Rhodes, D., Singer-Berk, M., England, E.M., Seaby, E.G., Kosmicki, J.A., Walters, R.K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J.X., Samocha, K.E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A.H., Minikel, E.V., Weisburd, B., Lek, M., Ware, J.S., Vittal, C., Armean, I.M., Bergelson, L., Cibulskis, K., Connolly, K.M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M.E., Neale, B.M., Daly, M.J. and MacArthur, D.G. (2020). [The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581\(7809\): 434-443.](#)
- Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J.B., Kattman, B.L. and Maglott, D.R. (2018). [ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46\(D1\): D1062-D1067.](#)
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S. and Rives, A. (2023). [Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379\(6637\): 1123-1130.](#)
- Livesey, B.J. and Marsh, J.A. (2020). [Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Molecular Systems Biology*, 16\(7\): e9380.](#)
- Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T. and Rives, A. (2021). [Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34: 29287-29303.](#)
- Ng, P.C. and Henikoff, S. (2003). [SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31\(13\): 3812-3814.](#)
- Notin, P., Dias, M., Frazer, J., Marchena-Hurtado, J., Gomez, A., Marks, D.S. and Gal, Y. (2022). [Tranception: Protein fitness prediction with autoregressive transformers and inference-time retrieval. *Proceedings of the 39th International Conference on Machine Learning*, 16990-17017.](#)
- Notin, P., Kollasch, A.W., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, Y. and Marks, D.S. (2023). [ProteinGym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36: 64331-64379.](#)
- Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P. and Song, Y.S. (2019). [Evaluating protein transfer learning with TAPE. *Advances in Neural Information Processing Systems*, 32: 9689-9701.](#)
- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K. and Rehm, H.L. (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. *Genetics in Medicine*, 17\(5\): 405-424.](#)
- Riesselman, A.J., Ingraham, J.B. and Marks, D.S. (2018). [Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, 15\(10\): 816-822.](#)
- Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J. and Fergus, R.

- (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15): e2016239118.
- Starita, L.M., Ahituv, N., Dunham, M.J., Kitzman, J.O., Roth, F.P., Seelig, G., Shendure, J. and Fowler, D.M. (2017). Variant interpretation: Functional assays to the rescue. *American Journal of Human Genetics*, 101(3): 315-325.
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zidek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D. and Velankar, S. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1): D439-D444.