Unsupervised language models for disease variant prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract

There is considerable interest in predicting the pathogenicity of protein variants in human genes. Due to the sparsity of high quality labels, recent approaches turn to *unsupervised* learning, using Multiple Sequence Alignments (MSAs) to train generative models of natural sequence variation within each gene. These generative models then predict variant likelihood as a proxy to evolutionary fitness. In this work we instead combine this evolutionary principle with pretrained protein language models (LMs), which have already shown promising results in predicting protein structure and function. Instead of training separate models per-gene, we find that a single protein LM trained on broad sequence datasets can score pathogenicity for any gene variant zero-shot, without MSAs or finetuning. We call this unsupervised approach **VELM** (Variant Effect via Language Models), and show that it achieves scoring performance comparable to the state of the art when evaluated on clinically labeled variants of disease-related genes.

1 Introduction

Understanding and quantifying the pathogenicity of human gene variants could transform healthcare, better inform treatment decisions, and enable new treatment modalities. However, relating specific missense variants to phenotypical disease indications is challenging, since the number of such variants (6.5 million) observed in the human population so far exceeds that which can be analyzed Karczewski et al. [2020]. Despite large-scale efforts to collate the disease relevance of gene variants Landrum and Kattman [2018], the majority of variants remain pathogenically unclassified Van Hout et al. [2020].

Computational methods offer the promise of at-scale interpretation of variants at speeds useful in a clinical setting Jagadeesh et al. [2019], Rentzsch et al. [2019]. However, many supervised models are trained on clinical labels of variable quality or with inconsistent clinical annotations resulting in inconsistent model performance. Unsupervised generative models avoid the labeling issues and have been successfully used to predict protein function and stability Hopf et al. [2014], Lapedes et al. [2012], Meier et al. [2021]. More recently, Frazer et al. [2021] introduced EVE, a family of variational autoencoders (VAEs) trained on protein Multiple Sequence Alignments (MSAs) for each gene of interest. EVE scores pathogenicity using variant probabilities as proxies for evolutionary fitness, and achieves current state-of-the art performance compared to other computational approaches without training on clinical labels.

In this work we describe **VELM** (Variant Effect via Language Models), an unsupervised approach for scoring variant pathogenicity using protein language models (LMs). Like prior unsupervised evolutionary approaches, VELM scores pathogenicity by using a sequence model to predict sequence likelihood. However, instead of training separate gene-specific generative models to estimate likelihoods, we use protein LMs pretrained by self-supervised learning on large open datasets of protein sequences. This training procedure produces models that capture statistical patterns across a broad

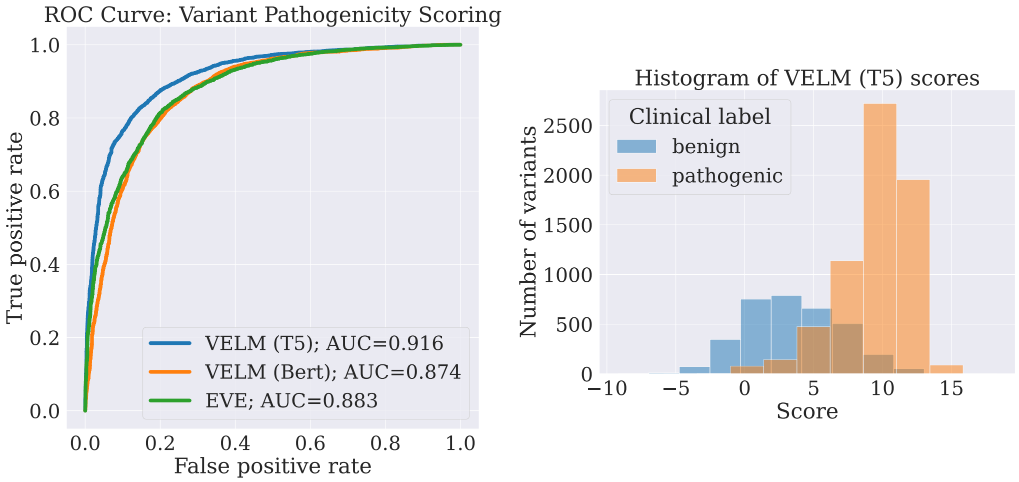

Figure 1: Left: Receiver Operating Characteristic (ROC) curve of VELM and EVE scores on our evaluation set of clinically labeled gene variants. VELM (T5) outperforms both the EVE score and VELM (Bert). Positive $\iff$ pathogenic, negative $\iff$ benign. Right: Histogram of VELM (T5) scores on clinically labeled variants. Broadly speaking, VELM assigns higher scores to pathogenic variants than for benign ones.

distribution of protein sequences, and enables estimating sequence likelihood for any gene variant zero-shot without finetuning on any MSAs. Thus, our approach uses a single model to efficiently scores pathogenicity for any gene variant of interest without having to train a new generative model per-gene. Ultimately, VELM allows us to efficiently predict pathogenicity for the large number of currently unlabeled variants across human disease-related genes.

When evaluated on a set of variants with known clinical labels from the ClinVar dataset [Landrum and Kattman, 2018], we find that the VELM score can discriminate variant pathogenicity with an AUC=0.92, exceeding the performance of EVE (AUC=0.89), see Figure 1.

2 VELM: Variant Effect via Language Models

Starting from a reference wildtype protein, our goal is to predict the pathogenicity of a given variant directly from its protein sequence. Following an unsupervised evolutionary approach, we leverage the relationship between sequence likelihood and evolutionary fitness to score variants without training on clinical labels (which can cause overfitting). We estimate sequence likelihood through protein language models (LMs) pretrained on large protein sequence datasets. Compared to EVE [Frazer et al., 2021], this removes the need to train separate per-gene generative models on processed MSAs. Indeed, we will show that a single pretrained LM can score any gene variant with no finetuning.

Inspired by techniques from natural language processing (NLP), protein language models are typically trained on large datasets of protein sequences with a masked language modeling objective. This trains the model to estimate the distribution over residues at particular positions given the *context* residues at surrounding positions. More precisely, these models compute $P(x_{i_1} = \bullet, \cdots, x_{i_m} = \bullet | x_{\setminus\{i_1,\cdots,i_m\}})$, where in practice the context $x_{\setminus\{i_1,\cdots,i_m\}}$ is created by masking the sequence at positions $i_1, \cdots, i_m$.

To define the VELM pathogenicity score, we need to use the protein LM to estimate a notion of variant likelihood (relative to the wildtype). We denote the wildtype sequence x^{wt} and variant sequence x^{mt}, and define the set of mutation positions $M = \{i : x_i^{\text{mt}} \neq x_i^{\text{wt}}\}$. Meier et al. [2021] found that the log odds ratio at mutated positions can effectively predict protein function. We define the VELM score using the same approach:

$$S(x^{\text{mt}}) := \sum_{i \in M} \log P(x_i = x_i^{\text{wt}} | x_{\setminus M}^{\text{wt}}) - \log P(x_i = x_i^{\text{mt}} | x_{\setminus M}^{\text{mt}}) \qquad (1)$$

where $x_{\setminus M}$ indicates masking x at all positions $i \in M$ (notably, $x_{\setminus M}^{\text{mt}} = x_{\setminus M}^{\text{wt}}$). Intuitively, $S(x^{\text{mt}})$ should be higher when the variant is *less* likely, indicating that it is more likely to be pathogenic.

Method	VELM (T5)	VELM (Bert)	EVE	REVEL	MA	DG2
mAUC ($\geq$ 1 labels)	0.901	0.858	0.917	0.934	0.888	0.895
mAUC ($\geq$ 3 labels)	0.912	0.876	0.930	0.946	0.895	0.901
mAUC ($\geq$ 5 labels)	0.933	0.892	0.936	0.956	0.904	0.916

Figure 2: Mean of AUCs (mAUC) over the evaluation set of disease-relevant genes (weighted by number of known labels). For each row, "$\geq N$ labels" means we restrict evaluation to genes that have at least N pathogenic and N benign labels for evaluating AUC. Note that VELM (ours), EVE [Frazer et al., 2021] and MA (MutationAssessor) are all unsupervised methods. DG2 (DEOGEN2) [Raimondi et al., 2017] is supervised by clinical labels, while REVEL [Ioannidis et al., 2016] is an ensemble method that combines the output of multiple individual tools.

Computing $S(x^{\mathrm{mt}})$ is relatively efficient and requires $|M|$ forward passes to evaluate a single variant. For reasonably small $|M|$, GPU batching leads to only a single forward pass in practice.

3 Experiments and Analysis

We apply VELM to missense variants of human disease-related genes whose sequence lengths are ≤ 512[1]. From the ClinVar dataset [Landrum and Kattman, 2018] there are known clinical labels for 10011 variants across the 1348 genes we consider[2]: 6613 variants are labeled "pathogenic" while 3398 are labeled "benign." For these clinically labeled variants, we compare our VELM score against the value of the label, and evaluate the effectiveness of using VELM score to classify variant pathogenicity.

Computing the VELM pathogenicity score (Eq. 1) requires a pretrained protein LM, for which there are multiple choices. Here we consider both ProtBert (420M parameters) and ProtT5 (3B parameters) [Elnaggar et al., 2021], both trained by masked language modeling on BFD [Steinegger and Söding, 2018] and UniRef [Suzek et al., 2015]. We will denote the results of scoring variants with each LM as **VELM (Bert)** and **VELM (T5)**, respectively. For comparison, we also evaluate the performance of other methods on the same set of variants:

1. EVE [Frazer et al., 2021]: An unsupervised evolutionary method that trains separate generative models on MSAs for each gene.
2. MutationAssessor (MA) Reva et al. [2011]: Another unsupervised scoring approach.
3. DEOGEN2 (DG2) [Raimondi et al., 2017]: A supervised method trained on clinical disease labels.
4. REVEL [Ioannidis et al., 2016]: An ensemble method that combines the output of multiple individual tools.

Aggregate Metrics: Figure 1 shows how the VELM score discriminates pathogenicity on our set of labeled variants. The VELM (T5) score has an AUC of 0.92, outperforming both EVE and VELM (Bert), with AUCs 0.88 and 0.87, respectively. The ROC Curve indicates that VELM (T5) produces pathogenicity scores with an overall better tradeoff beteween TPR and FPR compared to the other methods. The histogram of scores shaded by clinical label shows that the VELM scores is broadly capable of separating pathogenic and benign gene variants. Since the score is simply computed from the output of a protein LM, this indicates that the pretraining process learns statistical patterns in protein sequence that are relevant to predicting pathogenicity (via predicting likelihood).

Per-Gene Metrics: We can also evaluate how VELM scores discriminate pathogenicity on a per-gene basis. We calculate the *Mean AUC* (mAUC) by computing AUC for variants of each gene separately, then average the AUCs over genes weighted by the number of clinical labels available. Since many genes have just a few clinically labeled variants, per-gene evaluation statistics may be very noisy. We separately evaluate mAUC over genes with at least N pathogenic and benign labels, where $N = 1, 3,$ or 5. Figure 2 shows that REVEL generally achieves the highest Mean AUC on each evaluation set. Among non-ensemble methods, EVE generally performs best, though for the least noisy evaluation set of genes with ≥ 5 labels, VELM (T5) and EVE perform comparably.

[1] This is not a general limitation of VELM, but the particular protein LMs we use in this evaluation were only trained on sequences of length ≤ 512.

[2] We restrict to those labels with a ClinVar quality rating of at least one star.

3.1 Analysis

Overall, VELM (T5) achieves state of the art performance at predicting pathogenicity for arbitrary gene variants (aggregate AUC). It is comparable to other methods when scoring at a per-gene level (mean AUC), nearly matching state of the art for the least noisy evaluation set. These results are notable since VELM simply uses a pretrained protein LM to score any gene variant zero-shot, while other methods either train on clinical labels or on gene-specific evolutionary data. This leaves open the possibility for further improving performance by finetuning the protein LM on data pertaining to the disease-relevant genes of interest.

The fact that VELM (T5) outperforms VELM (Bert) falls in line with prior observations that ProtT5 outperforms ProtBert on a variety of structure and function prediction tasks Elnaggar et al. [2021]. This suggests that pathogenicity prediction may be yet another "downstream task" where performance can improved by simply pretraining better protein LMs.

4 Related Work

There has been extensive prior work in computational techniques to predict protein pathogenicity and in using large-scale self-supervised language models for protein sequences.

The literature on computational approaches for predicting protein pathogenicity is large and growing. Roughly speaking, these approaches can be categorized into *supervised* methods (e.g., Adzhubei et al. [2010], Raimondi et al. [2017]), *unsupervised methods* (e.g., Sim et al. [2012], Choi et al. [2012]), and supervised *meta-predictor* methods that use the outputs of both supervised and unsupervised methods as features (e.g., Ioannidis et al. [2016], Jagadeesh et al. [2016], Feng [2017], Qi et al. [2021], Ionita-Laza et al. [2016]). The unsupervised approach is favored in prior work which cites the variable quality of labels, bias in label availability, and sparsity of labels as difficulties in developing and validating supervised methods. In comparing to our work, the most recent and relevant such *unsupervised* approach is EVE Frazer et al. [2021], which is state-of-the-art. The key features distinguishing our work from that of Frazer et al. [2021] are: (a) we have one global protein LM instead of per-family sequence models (b) we train on a large database of protein sequences with no fine-tuning instead of EVE's individual MSAs, and (c) we perform zero-shot inference across all residue locations of a protein, instead of EVE's *focus* indices.

There has been a recent growth of interest in training language models on protein sequence datasets for the purposes of predicting protein structure and function [Alley et al., 2019, Lu et al., 2020, Madani et al., 2020, Elnaggar et al., 2021, Rives et al., 2021, Notin et al., 2022]. Most closely related to our work is ESM-1v [Meier et al., 2021], which used protein LMs and the log odds ratio at mutated positions to predict the effect of mutations on protein function zero-shot. Given the success of protein LMs for predicting structure and function, VELM explores their effectiveness for directly predicting pathogenicity in disease-relevant human genes.

5 Conclusion

In this work, we investigate the effectiveness of pretrained protein language models for assessing variant pathogenicity, a problem of great clinical interest. We introduce an unsupervised method called VELM that scores variant sequences by using protein LMs to estimate sequence likelihood, and show that it matches state of the art predictive performance. VELM is computationally efficient and flexible, using a single model to score variants of any gene with no finetuning.

The current work can be improved along multiple directions. First, the current protein LMs were trained on sequences of limited length, restricting our evaluation to sequences of length ≤ 512. Aside from removing this technical limitation, results can likely be improved by using better pretrained LMs such as ESM [Hsu et al., 2022], or by finetuning the LMs on relevant sequences (to human disease-related genes).

References

I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. A method and server for predicting damaging missense mutations. *Nature* *methods*, 7(4):248–249, 2010.

E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G. M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan. Predicting the functional effect of amino acid substitutions and indels. 2012.

A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, et al. Prottrans: towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

B.-J. Feng. Perch: a unified framework for disease gene prioritization. *Human mutation*, 38(3): 243–251, 2017.

J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, and D. S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

T. A. Hopf, C. P. Schärfe, J. P. Rodrigues, A. G. Green, O. Kohlbacher, C. Sander, A. M. Bonvin, and D. S. Marks. Sequence co-evolution gives 3d contacts and structures of protein complexes. *elife*, 3: e03430, 2014.

C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, and A. Rives. Learning inverse folding from millions of predicted structures. *bioRxiv*, 2022. doi: 10.1101/2022.04.10.487779. URL `https://www.biorxiv.org/content/early/2022/04/10/2022.04.10.487779`.

N. M. Ioannidis, J. H. Rothstein, V. Pejaver, S. Middha, S. K. McDonnell, S. Baheti, A. Musolf, Q. Li, E. Holzinger, D. Karyadi, et al. Revel: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*, 99(4):877–885, 2016.

I. Ionita-Laza, K. McCallum, B. Xu, and J. D. Buxbaum. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature genetics*, 48(2):214–220, 2016.

K. A. Jagadeesh, A. M. Wenger, M. J. Berger, H. Guturu, P. D. Stenson, D. N. Cooper, J. A. Bernstein, and G. Bejerano. M-cap eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature genetics*, 48(12):1581–1586, 2016.

K. A. Jagadeesh, J. M. Paggi, J. S. Ye, P. D. Stenson, D. N. Cooper, J. A. Bernstein, and G. Bejerano. S-cap extends pathogenicity prediction to genetic variants that affect rna splicing. *Nature genetics*, 51(4):755–763, 2019.

K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

M. J. Landrum and B. L. Kattman. Clinvar at five years: Delivering on the promise. *Human mutation*, 39(11):1623–1630, 2018.

A. Lapedes, B. Giraud, and C. Jarzynski. Using sequence alignments to predict protein structure and stability with high accuracy. *arXiv preprint arXiv:1207.2484*, 2012.

A. X. Lu, H. Zhang, M. Ghassemi, and A. Moses. Self-supervised contrastive learning of protein representations by mutual information maximization. *BioRxiv*, 2020.

A. Madani, B. McCann, N. Naik, N. S. Keskar, N. Anand, R. R. Eguchi, P.-S. Huang, and R. Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.

J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

P. Notin, M. Dias, J. Frazer, J. M. Hurtado, A. N. Gomez, D. Marks, and Y. Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pages 16990–17017. PMLR, 2022.

H. Qi, H. Zhang, Y. Zhao, C. Chen, J. J. Long, W. K. Chung, Y. Guan, and Y. Shen. Mvp predicts the pathogenicity of missense variants by deep learning. *Nature communications*, 12(1):1–9, 2021.

D. Raimondi, I. Tanyalcin, J. Ferté, A. Gazzo, G. Orlando, T. Lenaerts, M. Rooman, and W. Vranken. Deogen2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic acids research*, 45(W1):W201–W206, 2017.

P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*, 47(D1):D886–D894, 2019.

B. Reva, Y. Antipin, and C. Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39(17):e118–e118, 2011.

A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng. Sift web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research*, 40(W1):W452–W457, 2012.

M. Steinegger and J. Söding. Clustering huge protein sequence sets in linear time. *Nature communications*, 9(1):1–8, 2018.

B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and U. Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

C. V. Van Hout, I. Tachmazidou, J. D. Backman, J. D. Hoffman, D. Liu, A. K. Pandey, C. Gonzaga-Jauregui, S. Khalid, B. Ye, N. Banerjee, et al. Exome sequencing and characterization of 49,960 individuals in the uk biobank. *Nature*, 586(7831):749–756, 2020.