Briefings in Bioinformatics, 2025, **26(1)**, bbae664
https://doi.org/10.1093/bib/bbae664
Problem Solving Protocol

Combining evolution and protein language models for an interpretable cancer driver mutation prediction with D2Deep

Konstantina Tzavella[1], Adrian Diaz[1], Catharina Olsen[1,2,3,‡], Wim Vranken[1,4,5,6,7,‡,*]

[1]Interuniversity Institute of Bioinformatics (IB2), Université Libre de Bruxelles, Vrije Universiteit Brussel (ULB-VUB), Triomflaan, Brussels 1050, Belgium
[2]Brussels Interuniversity Genomics High Throughput Core (BRIGHTcore), Vrije Universiteit Brussel (VUB), Université Libre de Bruxelles (ULB), Laarbeeklaan 101, Brussels 1090, Belgium
[3]Clinical Sciences, Research Group Genetics, Reproduction and Development (GRAD), Vrije Universiteit Brussel (VUB), Universitair Ziekenhuis Brussel (UZ Brussel), Laarbeeklaan 101, Brussels 1090, Belgium
[4]Structural Biology Brussels, Vrije Universiteit Brussel (VUB), Pleinlaan 2, Brussels 1050, Belgium
[5]Chemistry Department, Vrije Universiteit Brussel, Pleinlaan 2, Brussels 1050, Belgium
[6]AI Lab, Vrije Universtiteit Brussel, Pleinlaan 2, Brussels 1050, Belgium
[7]Biomedical sciences, Vrije Universiteit Brussel, Laarbeeklaan 101, Brussels 1090, Belgium

*Corresponding author. Interuniversity Institute of Bioinformatics (IB2), ULB-VUB, Triomflaan, Brussels 1050, Belgium. E-mail: wim.vranken@vub.be
‡Catharina Olsen and Wim Vranken contributed equally to this work.

Abstract

The mutations driving cancer are being increasingly exposed through tumor-specific genomic data. However, differentiating between cancer-causing driver mutations and random passenger mutations remains challenging. State-of-the-art homology-based predictors contain built-in biases and are often ill-suited to the intricacies of cancer biology. Protein language models have successfully addressed various biological problems but have not yet been tested on the challenging task of cancer driver mutation prediction at a large scale. Additionally, they often fail to offer result interpretation, hindering their effective use in clinical settings. The AI-based D2Deep method we introduce here addresses these challenges by combining two powerful elements: (i) a nonspecialized protein language model that captures the makeup of all protein sequences and (ii) protein-specific evolutionary information that encompasses functional requirements for a particular protein. D2Deep relies exclusively on sequence information, outperforms state-of-the-art predictors, and captures intricate epistatic changes throughout the protein caused by mutations. These epistatic changes correlate with known

Received: June 21, 2024. **Revised:** September 15, 2024. **Accepted:** December 7, 2024

to maintain protein function, have been shown to enhance mutation effect prediction [16–19]. Addressing higher-order epistasis, involving more than two amino acids in different positions in the sequence, remains a challenge despite numerous converging evidence on their crucial role in protein evolution and fitness [20–22]. To our knowledge, only EVE [23] and AlphaMissense [8] are able to incorporate such higher-order constraints. EVE relies on evolutionary information learned by amino acid sequences for each protein across different species using multiple sequence alignments (MSAs) and clusters the learned representations into benign or pathogenic groups based on their likelihood to occur. Because alternative isoforms of the same gene have identical homologs, it remains uncertain whether this approach can differentiate the effect of mutations on different isoforms [24]. Also, as an unsupervised model, it encounters challenges to discern protein-specific structural or functional alterations. To capture protein-specific nuances, the integration of direct functional and contextual data for a particular protein has been shown to lead to substantial improvements in prediction [25, 26]. AlphaMissense incorporates such data in a supervised setting, but considers frequently observed variants in human and primate populations as benign, whereas absent variants are labeled as pathogenic, resulting in an inherently noisy and biased training set, as rare variants can also be benign [27]. None of these predictors provide an interpretation of results that could support possible clinical decisions based on them.

Another deep-learning method for variant effect prediction utilizes protein language models, a technique originating from natural language processing. Protein language models (pLMs) do not rely on explicit homology and can estimate the likelihood of any possible amino acid sequence, learning from common patterns across protein families, allowing information to be generalized [28–30]. They have demonstrated the ability to implicitly learn how protein sequences influence various aspects of protein structure and function, including secondary structure and subcellular location [31, 32]. However, they have not yet been tested on the challenging task of cancer driver mutation prediction at a large scale.

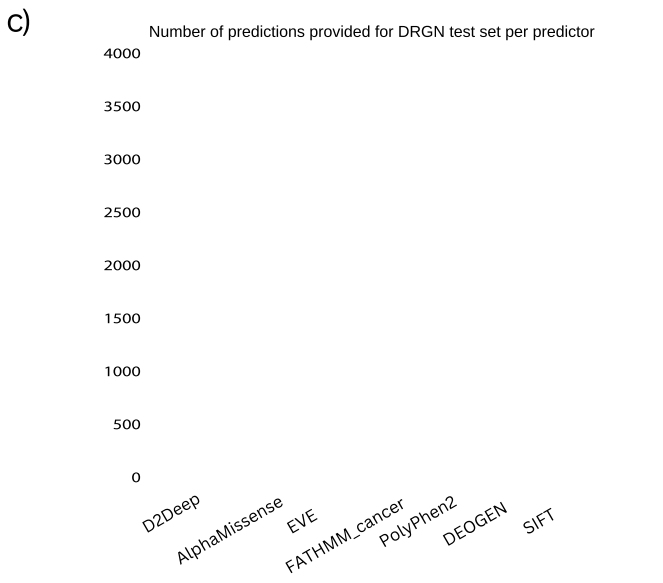
We present D2Deep, a protein sequence–only prediction method to distinguish driver from passenger mutations in cancer that is based on an original combination of protein-specific evolutionary information (EI) from MSAs with pLMs. pLMs inherently represent the probability of amino acids occurring in various contexts, in other words the grammar (or syntax) of natural protein sequences, while MSAs represent protein-specific evolutionary information that incorporates the effect of each mutation on protein function (semantic correctness). By employing a Gaussian mixture model (GMM) on the pLM features across evolutionary related sequences, we can therefore capture the site-specific evolutionary variation of proteins and so detect when sequences are semantically incorrect. The features extracted from this approach, which now combine general protein sequence principles with specific protein information, are subsequently used in a downstream supervised task for distinguishing between driver and passenger mutations in cancer. They capture the effect of the mutation throughout the protein and can be used for the interpretation of results in the clinical


effect

FATHMM_cancer


a)
100
80
60
40
20
0
Accuracy (%)
D2Deep
AIp
hotspot mutations
non-hotspot mutations


b)


c)
Number of predictions provided for DRGN test set per predictor
4000
3500
3000
2500
2000
1500
1000
500
0
D2Deep
AlphaMissense
EVE
FATHMM_cancer
PolyPhen2
DEOGEN
SIFT

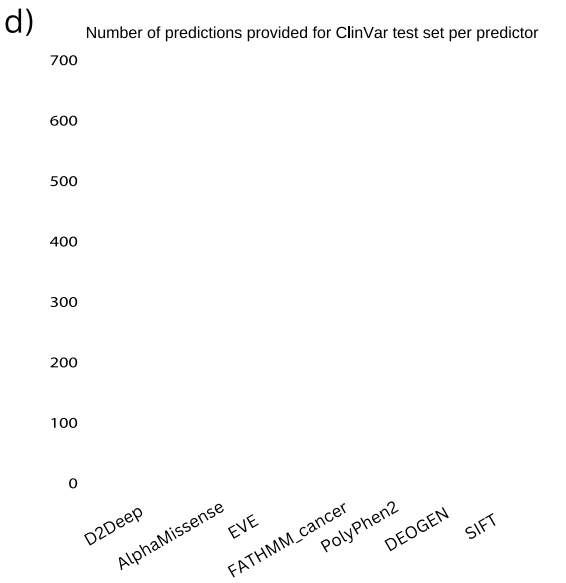

d)
Number of predictions provided for ClinVar test set per predictor
700
600
500
400
300
200
100
0
D2Deep
AlphaMissense
EVE
FATHMM_cancer
PolyPhen2
DEOGEN
SIFT

Figure 4. D2Deep confidence scores. Predictions for ClinVar annotations on two cancer driver genes (TP53, BRAF) when (a) all D2Deeps' predictions are shown. (b) High confidence predictions for ClinVar annotations are shown. (c) D2Deep confidence scores of correctly (intense colors) and incorrectly (soft colors) classified mutations for TP53 gene. We chose to demonstrate the predictions of genes with a relatively large number of high-quality labels in ClinVar. The rest of the available predictions can be found on our webserver.

To illustrate this behavior, we first compared the D2Deep features with long-range effects proposed by MAVISp [39] for the KRAS gene (Fig. 5). MAVISp uses the protein's structure to calculate the free energy changes caused by the mutations on the Switch I region, suggesting that some of these changes might contribute to protein activation through distal effects. In accordance with MAVISp findings, the D2Deep features have higher signal at positions corresponding to peaks in allosteric free energy changes, relying solely on sequence information.

Additionally, we collected ClinVar benign and pathogenic mutations on the TP53 gene and analyzed each mutation's impact on distant amino acids. Figure 6a shows the features of two pathogenic mutations, E258K and C238R. Their features have the highest values, and thus influence, distant in sequence amino

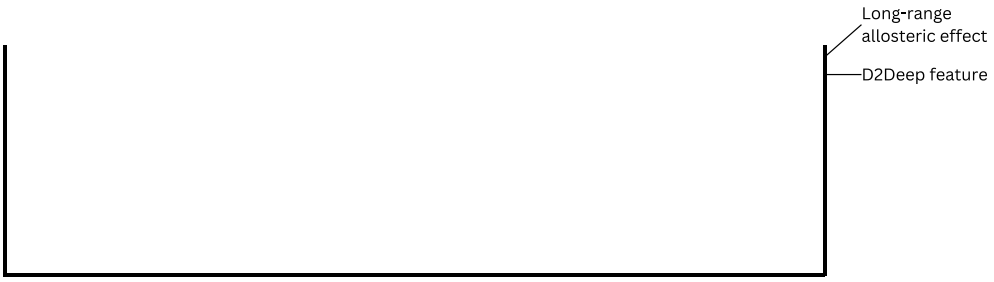

Long-range
allosteric effect
D2Deep feature

Figure 6. Long-range effect of TP53 mutations. (a) D2Deep features used in the downstream prediction of 149 benign and pathogenic ClinVar TP53 mutations. Two pathogenic mutations E258K and C238R were considered to study the influence of features in distant amino acid and capture potential co-evolution forces. (b) Predicted AlphaFold2 TP53 structure with focus on the surrounding regions of the mutations E258K and C238R. The amino acids that correspond to the highest D2Deep features are highlighted. (c) D2Deep features for benign (blue) and pathogenic (red) mutations of TP53. ClinVar benign mutations (352 mutations) have low feature signal while ClinVar pathogenic mutations (176 mutations) have high D2Deep signal. The VUS features follow the ground truths suggesting a correlation between known and unknown mutations. We consider the top 10 highest values as peaks. (d) D2Deep features of novel mutations Ala138Val, Ala276Ile, and Glu358Val. Only the sequence region where the D2Deep features are most affected by the mutation are shown, with the most influenced amino acids enriched in COSMIC mutations.

Gene Census oncogene or in a tumor suppressor gene (namely, Tiers 1, 2, and 3 in COSMIC as described in Supplementary Fig. 7). Their distribution in cancer-related diseases in ClinVar can be seen in Supplementary Fig. 8. We additionally collected 2657 somatic variants with known molecular consequences from Cancer Genome Interpreter (CGI). To balance the (likely) pathogenic SAVs of each gene with a balanced benign set, we curated the UniProtKB/Swiss-Prot humsavar dataset (release 21 December 2021) keeping the 39 325 benign/likely benign SAVs. To this we added the ClinVar dataset with benign annotations that resulted in 43 030 benign SAVs. Additionally, we included common variants, SAVs frequently observed in the general population, from the gnomAD database, keeping mutations with allele frequencies (AF) >0.1%. Lastly, we mined single nucleotide polymorphisms (SNPs) from the Single Nucleotide Polymorphism database (dbSNP) excluding mutations with conflicted or uncertain interpretation, pathogenic, or risk associations. The mapping of the chromosomal position provided by dbSNP to protein coding positions was performed with the use of TransVar software [44]. After removing dataset overlap, 178 979 benign SAVs were retained. To establish a well-balanced training set at the gene level, we gathered both pathogenic and benign mutations for each gene while maintaining a ratio of not exceeding the 40%–60% class balance. To ensure fairness, for genes with a limited number of available mutations, we enforced a maximum difference of two mutations between the pathogenic and benign classes. After the filtering out the sequences with >2200 amino acids, the final training set contained 6608 mutations, 2956 deleterious and 3652 benign, internally balanced within 1012 genes. The script and data for the above workflow are publicly available.

Test sets

DRGN: To assess the construction bias affecting the *in silico* predictors, researchers introduced the DRGN set [2], comprising a total of 4093 variants. Among these variants, 1809 were identified as deleterious, while 2284 were categorized as passenger mutations. These variants were mapped to 153 driver genes. The deleterious variants were specifically selected from CGI, based on their experimental validation as cancer driver variants [45]. As for the passenger mutations, they encompassed 63 525 germline variants unrelated to cancer, sourced from Humsavar [35]. To ensure a comprehensive evaluation, the final test set included genes that had at least one positive (deleterious) and one negative (passenger) sample, comprising 3608 mutations.

Consensus pathogenic variants: We calculated the output of the method on 269 cancer driver genes included in Next Generation Sequencing (NGS) panel of biopsies of hematological and solid tumors from Compermed Guidelines (https://www.compermed.be/docs/Guidelinesjan2020-2.pdf). For the demonstration of predictions, we chose genes with a relatively large number of high-quality labels in ClinVar, three tumor suppressor genes (TP53, PTEN, CHEK2), and two oncogenes (BRAF, AR) with Review status: Practice guideline, Expert panel, Multiple submitters, Single submitter (downloaded in March 2023).

Transfer learning with pretrained model ProtT5-XL

Language models (LMs) are typically based on the Transformer architecture, which uses a mechanism called "self-attention" to weigh the importance of different words in a sentence when making predictions. This allows them to capture long-range dependencies in text. Protein LMs (pLMs) leverage vast datasets of protein sequences, similar to the training of language models. By training on billions of amino acids, these models learn to model statistical dependencies among amino acids based on their co-occurrence patterns across sequences. The learned representations have been shown to retain critical biophysical properties, making them valuable inputs for downstream tasks [31]. The best-performing pLM for our task was ProtT5-XL [31], trained on the BFD100 [46] and Uniref50 [47] datasets, from which we extracted 1024-dimensional (1024-D) amino acid representations.

Gaussian mixture model

The MSAs for both the training and test sets were generated using the mmseq2 algorithm with the Uniref100 and PDB70 databases [47]. The resulting alignments were filtered according to the protocol proposed by Hopf *et al.* [15]. In line with this protocol, we retained sequences that aligned to at least 50% of the target sequence and had at least 70% column occupancy. For five protein isoforms in our training set (out of a total of 1132), the filtering process resulted in fewer than two aligned sequences in the MSA. To address this, we applied a more lenient filtering cutoff, retaining sequences with at least 20% identity to the original protein.

Each amino acid in the MSA was then mapped to the 1024-D embedding learned during the pretraining phase. We applied a GMM from the sklearn library to each MSA column to measure how mutations deviate from the underlying distribution. Since the feature space (1024-D) exceeded the number of sequences in some MSAs, which could lead to overfitting, we performed dimensionality reduction. We reduced the feature dimensions using max-pooling with a kernel size of 50 and a stride of 50, resulting in 20-dimensional vectors (20-D) for each amino acid. The selection of 20-D was based on an evaluation of different dimensions to determine which provided the highest probability for the GMM and thus the best fit.

To prevent bias from the GMM's structure in each dimension, we employed a threshold-based approach. We fitted a GMM to the 20-D samples of each MSA column, computed the log-likelihood of the WT features to be in the model and the log-likelihood threshold, below which 1% of the samples lie. Then, we computed the distance between the log-likelihood and this threshold (WT-threshold). When a mutation is introduced, the pretrained embeddings of the entire sequence are affected. To capture its effect, we calculated the distance between the mutation's log-likelihood and above 1% threshold (MUT-threshold). This allowed us to quantify the mutation's impact, even when the WT and mutation features were similar in magnitude but opposite in sign.

The final feature was obtained by computing the difference between WT-threshold and MUT-threshold at each sequence position, yielding one value per position. These differences were concatenated across the sequence to form the feature set used in the downstream supervised learning task.

Model architecture and training

We built a deep-learning model for the mutation classification. The model receives the features produced by the GMM part of the algorithm and predicts the mutation's pathogenicity as a probability value from 0 to 1. The classifier is composed of two fully connected (FC) layers followed by a batch normalization for faster and more stable training. Because the FC layers must have a defined input length, we chose as maximum length the 2200 amino acids, padding shorter sequences with zero, which did not affect 90% of the cancer gene panel (Supplementary Fig. 9). To select the hyperparameters of D2Deep, we performed a grid search using 90% of the training data to train the model and the remaining 10% as

the validation set to select the hyperparameters (Supplementary Fig. 10, Supplementary Table 1). The test sets were not used for hyperparameter selection. The model was trained for 200 epochs. The number of epochs to train were selected based on the early stopping technique choosing the epoch on each the validation error starting increasing while the training error continues decreasing. Dropout layers of 0.3 are applied during training to all layers to avoid overfitting, followed by a Rectified Linear Unit (ReLU) activation function. The final FC layer is used to decide pathogenicity with the use of a sigmoid activation function incorporated in the BCEWithLogitsLoss loss function. The use of BCEWithLogitsLoss is recommended in the PyTorch documentation (https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html), over a plain Sigmoid followed by a BCELoss, as numerically more stable. The weights of the classifier were initialized using Xavier normal initialization, and the batch size was set to 64. Instead of depending only on the current gradient to update the weights in every step, a gradient descent with momentum of 0.1 aggregates current and past gradients. The optimizer used was AdamW with a learning rate of 0.00003 and a scheduler with a warmup period of 1000 steps. The model was trained in the PyTorch framework.

pLM comparison

For the comparison between pLMs and pLMs/EI, we follow the workflow proposed by ESM-1v [37]. We first calculated the pLM embeddings for the training dataset using the EMS-1v (esm1v_t33_650M_UR90S_1.ptk) pLM. As suggested by the authors, we kept a mean representation of 1280 dimensions for each mutated sequence. We then reduced the dimensions to 60 using principal component analysis, which was used for downstream supervised learning. The authors' grid search identified support vector regressor as the best model for variant prediction, which we also used. For leave-one-out cross-validation, we used the Training set described in Methods and materials. The same workflow was followed for the Prot-T5, using 1024-D embeddings as proposed by the authors.

Calculation of confidence score

We calculated the average GMM log-probabilities of the amino acids present in each position of the MSA of each protein sequence. We performed a grid search and selected the confidence formula that optimized the weighted performances across all test sets. This procedure resulted in Equation (1) for pathogenic mutations and Equation (2) for benign mutations.

$$\text{overall\_confidence}_{\text{pathogenic}}(x) = \frac{1}{M}\sum_{i=1}^{M}\left(\log\left(p_{GMMx}(i)\right)\right) \quad (1)$$

$$\text{overall\_confidence}_{\text{confidence}}(x) = 1 - \frac{1}{M}\sum_{i=1}^{M}\left(\log\left(p_{GMMx}(i)\right)\right)\cdot 1.3 \quad (2)$$

where M is the number of samples of fitted GMM (equal to number of sequences in the MSA), $\log(p_{GMMx})$ is the log-probability of each sample under the current model, and x is the position of the mutation. The inversion of overall_confidence for the benign mutations can be interpreted by Supplementary Fig. 11 where it is shown that the benign mutations have more confident predictions when the average log-probability of the samples on the position is smaller.

The calculation of the weighted performance was done by multiplying each performance metric by its corresponding weight (i.e. overall_confidence). This ensures that the contribution of each prediction to the overall performance is proportional to its confidence level. We then summed up the weighted performance metrics for all predictions and divided the sum of the weighted performance metrics by the sum of the weights as shown in Equation (3):

$$\frac{\sum_{i=1}^{n} w_i \cdot p_i}{\sum_{i=1}^{n} w_i} \quad (3)$$

where $w_1, w_2, \ldots, w_n$ are the weights assigned to each prediction and $p_1, p_2, \ldots, p_n$ are the performance metrics for each prediction. We proceeded by min–max normalizing the confidence of each prediction to obtain a range of 0%–100% for all genes.

Key Points

- The detection of cancer-driving mutations from genomics data is increasingly feasible, but many challenges remain, particularly detecting such mutations beyond the well-known hotspot regions of cancer-related genes.
- We present the first large-scale benchmark of protein language models for predicting cancer-driving mutations.
- Our findings highlight the necessity of including evolutionary information in the features captured by protein language models, which we achieve with the D2Deep method.
- The D2Deep features capture the epistatic effect of a mutation, which correlates with known mutations in the clinical setting and long-range allosteric effect from other tools.
- We introduced a statistical approach to generate confidence scores, crucial for clinical interpretation and mutation prioritization.

Supplementary Data

Supplementary data is available at *Briefings in Bioinformatics* online.

Author contributions

K.T. developed the computational method and performed data curation and validations. K.T. and W.V. wrote the manuscript. A.D. implemented the webserver. C.O. reviewed and edited the manuscript and supervised the research project along with W.V.

Conflict of interest

None declared.

Funding

Vrije Universiteit Brussel Research Council under the Interdisciplinary Research Program TumorScope [IRP20 to K.T.]; European Union's Horizon 2020 research and innovation program under grant agreement No. 101016834 (HosmartAI) to K.T., W.V.; Research Foundation Flanders (FWO) International Research Infrastructure [I000323N to W.V.].

Code availability

All code required to run D2Deep is available via our public GitHub repository (https://github.com/KonstantinaT/D2Deep/).

Data availability

All data used for training and testing the model are available in the public Zenodo repository: https://zenodo.org/doi/10.5281/zenodo.8200795 (DOI 10.5281/zenodo.8200795).

References

1. Won DG, Kim DW, Woo J, *et al.* 3Cnet: Pathogenicity prediction of human variants using multitask learning with evolutionary constraints. *Bioinformatics* 2021;**37**:4626–34. https://doi.org/10.1093/bioinformatics/btab529.
2. Raimondi D, Passemiers A, Fariselli P, *et al.* Current cancer driver variant predictors learn to recognize driver genes instead of functional variants. *BMC Biol* 2021;**19**:3. https://doi.org/10.1186/s12915-020-00930-0.
3. Jin SC, Homsy J, Zaidi S, *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat Genet* 2017;**49**:1593–601. https://doi.org/10.1038/ng.3970 Nature Publishing Group.
4. Agarwal SK, Beth Kester M, Debelenko LV, *et al.* Germline mutations of the MEN1 gene in familial multiple endocrine neoplasia type 1 and related states. *Hum Mol Genet* 1997;**6**:1169–75. https://doi.org/10.1093/hmg/6.7.1169.
5. Sevenet N, Sheridan E, Amram D, *et al.* Constitutional mutations of the hSNF5/INI1 gene predispose to a variety of cancers. *Am J Hum Genet* 1999;**65**:1342–8. https://doi.org/10.1086/302639.
6. Lamlum H, Al Tassan N, Jaeger E, *et al.* Germline APC variants in patients with multiple colorectal adenomas, with evidence for the particular importance of E1317Q. *Hum Mol Genet* 2000;**9**:2215–21. https://doi.org/10.1093/oxfordjournals.hmg.a018912.
7. Venkitaraman AR. Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* 2002;**108**:171–82. https://doi.org/10.1016/S0092-8674(02)00615-3.
8. Cheng J, Novati G, Pan J, *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 2023;**381**:eadg7492. https://doi.org/10.1126/science.adg7492.
9. Roy DM, Walsh LA, Chan TA. Driver mutations of cancer epigenomes. *Protein Cell* 2014;**5**:265–96. https://doi.org/10.1007/s13238-014-0031-6.
10. Greaves M, Maley CC. Clonal evolution in cancer. *Nature* 2012;**481**:306–13. https://doi.org/10.1038/nature10762.
11. Aaltonen LA, Abascal F, Abeshouse A, *et al.* Pan-cancer analysis of whole genomes. *Nature*. Nature Publishing Group, 2020;**578**:82–93. https://doi.org/10.1038/s41586-020-1969-6 .
12. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;**31**:3812–4. https://doi.org/10.1093/nar/gkg509.
13. Adzhubei IA, Schmidt S, Peshkin L, *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**:248–9. https://doi.org/10.1038/nmeth0410-248.
14. Rentzsch P, Witten D, Cooper GM, *et al.* CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 2019;**47**:D886–94. https://doi.org/10.1093/nar/gky1016.
15. Shihab HA, Gough J, Cooper DN, *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* 2013;**34**:57–65. https://doi.org/10.1002/humu.22225.
16. Hopf TA, Ingraham JB, Poelwijk FJ, *et al.* Mutation effects predicted from sequence co-variation. *Nat Biotechnol* 2017;**35**:128–35. https://doi.org/10.1038/nbt.3769 Nature Publishing Group.
17. Breen MS, Kemena C, Vlasov PK, *et al.* Epistasis as the primary factor in molecular evolution. *Nature* 2012;**490**:535–8. https://doi.org/10.1038/nature11510 Nature Publishing Group.
18. Figliuzzi M, Jacquier H, Schug A, *et al.* Coevolutionary landscape inference and the context-dependence of mutations in Beta-lactamase TEM-1. *Mol Biol Evol* 2016;**33**:268–80. https://doi.org/10.1093/molbev/msv211.
19. Pejaver V, Urresti J, Lugo-Martinez J, *et al.* Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat Commun* 2020;**11**:5918. https://doi.org/10.1038/s41467-020-19669-x Nature Publishing Group.
20. Weinreich DM, Lan Y, Wylie CS, *et al.* Should evolutionary geneticists worry about higher-order epistasis? *Curr Opin Genet Dev* 2013;**23**:700–7. https://doi.org/10.1016/j.gde.2013.10.007.
21. Domingo J, Diss G, Lehner B. Pairwise and higher order genetic interactions during the evolution of a tRNA. *Nature* 2018;**558**:117–21. https://doi.org/10.1038/s41586-018-0170-7.
22. Echave J, Wilke CO. Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence. *Annu Rev Biophys* 2017;**46**:85–103. https://doi.org/10.1146/annurev-biophys-070816-033819 NIH Public Access.
23. Frazer J, Notin P, Dias M, *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* 2021;**599**:91–5. https://doi.org/10.1038/s41586-021-04043-8 Nature Publishing Group.
24. Brandes N, Goldman G, Wang CH, *et al.* Genome-wide prediction of disease variant effects with a deep protein language model. *Nat Genet*. Nature Publishing Group, 2023;**55**:1512–22. https://doi.org/10.1038/s41588-023-01465-0.
25. Wittmann BJ, Yue Y, Arnold FH. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Syst* 2021;**12**:1026–1045.e7. https://doi.org/10.1016/j.cels.2021.07.008.
26. Hsu C, Nisonoff H, Fannjiang C, *et al.* Learning protein fitness models from evolutionary and assay-labeled data. *Nat Biotechnol* 2022;**40**:1114–22. https://doi.org/10.1038/s41587-021-01146-5.
27. Quaio CRD, Ceroni JRM, Cervato MC, *et al.* Parental segregation study reveals rare benign and likely benign variants in a Brazilian cohort of rare diseases. *Sci Rep* 2022;**12**:7764. https://doi.org/10.1038/s41598-022-11932-z.
28. Rives A, Meier J, Sercu T, *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* 2021;**118**:e2016239118.
29. Alley EC, Khimulya G, Biswas S, *et al.* Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* 2019;**16**:1315–22. https://doi.org/10.1038/s41592-019-0598-1 Nature Publishing Group.
30. Rao R, Meier J, Sercu T. *et al.* Rives a.: Transformer protein language models are unsupervised structure learners. *bioRxiv* 2020;2020.12.15.422761. Section: New Results. Available from: https://doi.org/https://www.biorxiv.org/content/10.1101/2020.12.15.422761v1.
31. Elnaggar A, Heinzinger M, Dallago C, *et al.*

33. Raimondi D, Tanyalcin I, Ferté J, *et al.* DEOGEN2: Prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res* 2017;**45**:W201–6. https://doi.org/10.1093/nar/gkx390.
34. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* 2013;**Chapter 7**:Unit7.20. https://doi.org/10.1002/0471142905.hg0720s76.
35. The UniProt Consortium. UniProt: The universal protein knowledgebase. *Nucleic Acids Res* 2017;**45**:D158–69. https://doi.org/10.1093/nar/gkw1099.
36. Sonnhammer EL, Eddy SR, Durbin R. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins Struct Funct Bioinf* 1998;**28**:405–20. https://doi.org/10.1002/(sici)1097-0134(199707)28:3⟨405::aid-prot10⟩3.0.co;2-l.
37. Meier J, Rao R, Verkuil R. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv* 2021;2021.07.09.450648. Section: New Results. Available from: https://doi.org/https://www.biorxiv.org/content/10.1101/2021.07.09.450648v2.
38. Dunham AS, Beltrao P. Exploring amino acid functions in a deep mutational landscape. *Mol Syst Biol* 2021;**17**:e10305. https://doi.org/10.15252/msb.202110305.
39. Arnaudi M, Beltrame L, Degn K. *et al.* MAVISp: A modular structure-based framework for genomic variant interpretation. *bioRxiv* 2022;2022.10.22.513328. Section: New Results. Available from: https://doi.org/https://www.biorxiv.org/content/10.1101/2022.10.22.513328v4.
40. Garziera M, Cecchin E, Canzonieri V, *et al.* Identification of novel somatic TP53 mutations in patients with high-grade serous ovarian cancer (HGSOC) using next-generation sequencing (NGS). *Int J Mol Sci* 2018;**19**:1510. https://doi.org/10.3390/ijms19051510.
41. Saha G, Singh R, Mandal A, *et al.* A novel hotspot and rare somatic mutation p.A138V, at TP53 is associated with poor survival of pancreatic ductal and periampullary adenocarcinoma patients. *Mol Med* 2020;**26**:59. https://doi.org/10.1186/s10020-020-00183-1.
42. Wang Y, Goh KY, Chen Z, *et al.* A novel TP53 gene mutation sustains non-small cell lung cancer through mitophagy. *Cells* 2022;**11**:3587. https://doi.org/10.3390/cells11223587.
43. Forbes SA, Bhamra G, Bamford S, *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* 2008;**Chapter 10**:Unit 10.11. https://doi.org/10.1002/0471142905.hg1011s57.
44. Zhou W, Chen T, Chong Z, *et al.* TransVar: A multilevel variant annotator for precision genomics. *Nat Methods* 2015;**12**:1002–3. https://doi.org/10.1038/nmeth.3622 Nature Publishing Group.
45. Tamborero D, Rubio-Perez C, Deu-Pons J, *et al.* Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med* 2018;**10**:25. https://doi.org/10.1186/s13073-018-0531-8.
46. Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2 Nature Publishing Group.
47. Suzek BE, Wang Y, Huang H, *et al.* UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 2015;**31**:926–32. https://doi.org/10.1093/bioinformatics/btu739.

Briefings in Bioinformatics, 2025, **26(1)**, bbae664
https://doi.org/10.1093/bib/bbae664
Problem Solving Protocol