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

Konstantina Tzavella[1], Adrian Diaz[1], Catharina Olsen[1,2,4†], Wim Vranken[1,2,3*†]

[1]Interuniversity Institute of Bioinformatics (IB2) , ULB-VUB, Brussels, Belgium.
[2]Brussels Interuniversity Genomics High Throughput Core (BRIGHTcore), Vrije Universiteit Brussel (VUB), Université Libre de Bruxelles (ULB), Laarbeeklaan 101, Brussels, 1090, Belgium.
[3]Structural Biology Brussels, Vrije Universiteit Brussel, Brussels, Belgium.
[4]Clinical Sciences, Research Group Reproduction and Genetics, Centre for Medical Genetics, Vrije Universiteit Brussel (VUB), Universitair Ziekenhuis Brussel (UZ Brussel), Laarbeeklaan 101, Brussels, 1090, Belgium.

*Corresponding author(s). E-mail(s): wim.vranken@vub.be;
Contributing authors: konstantina.tzavella@vub.be; adrian.diaz@vub.be; catharina.olsen@uzbrussel.be;
†These authors contributed equally to this work.

Abstract

Background: 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 predictors contain built-in biases and are often ill-suited to the intricacies of cancer biology. Most of them fail to offer result interpretation, creating a barrier to their effective utilization in the clinical setting.
Results: The AI-based D2Deep method we introduce here addresses these challenges by combining two powerful elements: i) a non-specialized 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 mutations in the clinical setting and can be used for the interpretation of results. The model is trained on a balanced, somatic training set and so effectively mitigates biases related to hotspot mutations compared to state-of-the-art techniques. The versatility of D2Deep is illustrated by its performance on non-cancer mutation prediction, where most variants still lack known consequences. D2Deep predictions and confidence scores are available via https://tumorscope.be/d2deep to help with clinical interpretation and mutation prioritization.
Conclusions: A purely protein sequence-based predictor to distinguish driver from passenger mutations in cancer outperforms predictors that utilize a plethora of contextual features, highlighting the necessity for somatic-specific, unbiased predictive models in the field of cancer research.

Keywords: Protein Large Language Models, Evolutionary Information, Machine Learning, Mutation effect prediction

Introduction

The increasing amount of DNA sequencing data that is becoming available has highlighted key similarities and differences between individuals related to biology and disease. Especially relevant are the exome regions of the DNA, as they are responsible for encoding the proteins that fulfill many cellular functions. They exhibit a wide range of missense mutations, where the amino acid encoding of DNA codons changes, so resulting in single amino acid variants (SAVs) at the protein level [1]. The vast majority of these are 'neutral' or 'benign' variants with no known detrimental effect on the phenotype of an individual, which are created randomly and are passed on by evolution [2]. Some SAVs, however, are pathogenic and cause critical protein malfunctions that have been linked to many Mendelian diseases [3] as well as predisposing to particular cancers [4],[5],[6],[7].

Among more than 4 million observed missense variants, only ∼2% have been clinically categorized as either pathogenic or benign [8]. For complex diseases such as cancer, the interpretation of the impact of SAVs is even more challenging. The behavior of cancer cells cannot be attributed to a single cellular process, but is rather the result of stochastic, somatic mutations that accrue over time, with additional contributions from germline and epigenetic variation [9]. Among these mutations, a small subset alters the cellular phenotype, and an even smaller subset confers a selective fitness advantage to clones, enabling them to evade the stringent cell control mechanisms that would normally result in cell death via apoptosis. These mutations, known as driver mutations, play a critical role in cancer development and progression, as opposed to selectively neutral passenger mutations [10]. The identification of driver mutations is essential for both understanding the molecular mechanisms of cancer and for developing targeted therapies. Pan-cancer genome analyses have revealed that,

on average, there are 4 to 5 driver mutations per cancer sample in both coding and non-coding genomic regions. However, from the sequenced samples collected from individuals with cancer, 5% of cases do not contain any of the known driver mutations [11]. In the remaining samples, while driver mutations are identified, the possibility for some mutations being overlooked remains. This implies that many cancer drivers remain to be discovered.

Key tools for such discovery are Artificial intelligence (AI) approaches, which are now frequently used for helping to identify the phenotypic effect of single (or multiple) SAVs. Compared to rare disease mutation predictors, however, few somatic-specific diagnostic tools are available, despite evidence that they can outperform germline-trained predictors on cancer driver mutations [12]. Typically, such approaches are 'black boxes' that pay little attention to the interpretation of the results, which could reveal the underlying disease mechanisms. Moreover, some cancer genes are very frequently studied and much more prevalent in the datasets, which creates bias in the training data: predictors often do not interpret the mutation itself, but rather decide on whether the mutation occurs in a well-known cancer related gene [2]. Hence, there is a need for generic, unbiased and interpretable methodologies for identifying cancer driver mutations.

Evolutionary information (EI) is essential in the process of assessing the impact of amino acid changes to protein sequence fitness, with now over 250 million (putative) protein sequences available [13],[14]. Machine Learning advances [15],[16],[17] and notably AlphaFold2 [18], have shown that large supervised models, combining high-quality curated data with EI, can provide insights on biological mechanisms and solve previously intractable problems. A key molecular mechanism for which information can be extracted from EI is epistasis, where amino acids changes lead to compensatory adjustments in other parts of the sequence. Models that capture such protein-specific pairwise interactions have been shown to improve mutation effect prediction [19],[20],[21],[22]. Addressing higher-order epistasis between more than two positions remains a challenge despite numerous converging evidence on their crucial role in protein evolution and fitness [23],[24],[25]. Nevertheless, some of the still most commonly used approaches such as SIFT [26], PolyPhen2 [27] and FATHMM [28] only consider the conservation of amino acid in single sequence positions to predict the effects of mutations. To our knowledge, only EVE [29] and AlphaMissense [8] endeavor to incorporate such higher-order constraints. EVE learns the distribution of amino acid sequences for each protein across different species, and clusters the learnt representations into benign or pathogenic groups based on their likelihood to occur. 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 [30],[31]. AlphaMissense incorporates such data in a supervised setting, but considers the variants frequently observed in human and primate populations as benign, whereas absent variants are labelled as pathogenic, resulting in an inherently noisy and biased training set. None of these predictors provide an interpretation of results that could support the clinical decision.

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 EI from Multiple Sequence Alignments (MSAs) with protein-generic language models (LM). The LM inherently represent the probability of specific amino acids occurring in various contexts, in other words the grammar (or syntax) of the biological sequence, while MSAs represent evolutionary information that incorporates the effect of each mutation on protein function (semantic correctness). By employing a Gaussian Mixture Model (GMM) on the LM 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. To minimize bias, we meticulously curated a per-gene balanced training set of somatic mutations, and trained a generic multilayer perceptron (MLP) predictor that can score pathogenicity for any gene variant (Fig.1). Importantly, the features capture the effect of the mutation throughout the protein and can be used for the interpretation of results in the clinical setting, complemented by the provided confidence score. We validate this model on an independent test set, built to eliminate bias of cancer related genes, and performs to state-of-the-art standards whilst only using generic protein information. Its general applicability is further demonstrated in non-cancer mutation prediction, where still over 98% of variants lack known consequences.

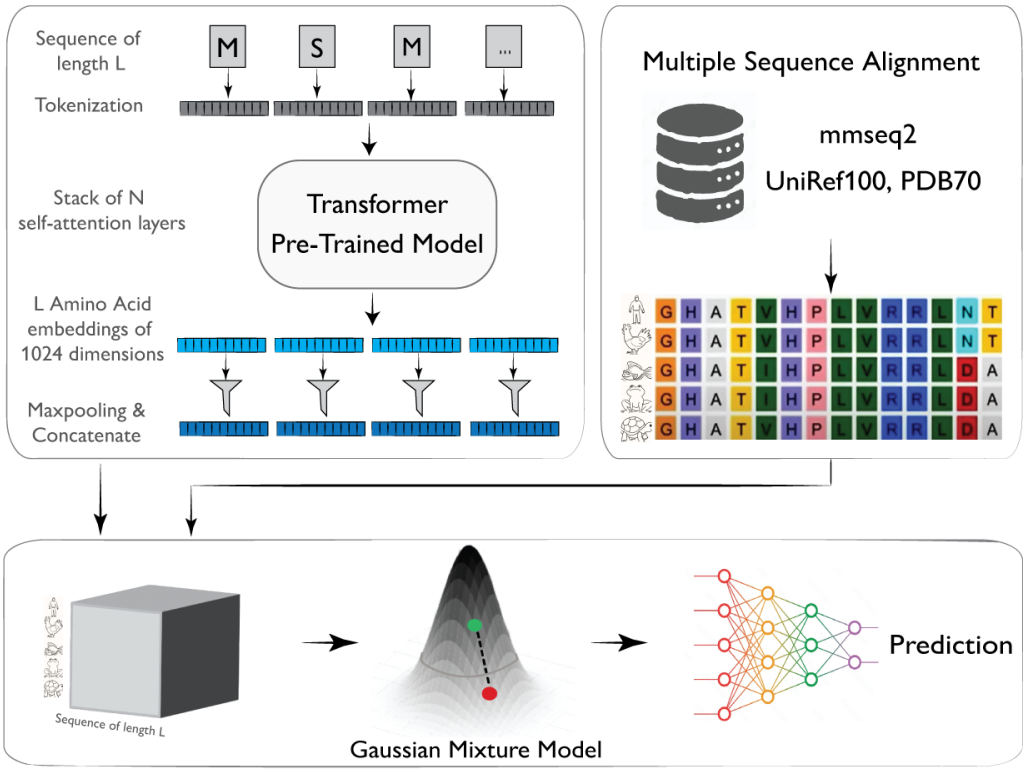

Fig. 1 Overview of D2Deep's pipeline. From the protein sequence input, each AA of the mutated sequence's MSA is mapped to the 20D LLM representations vector. Then, for each position a Gaussian Mixture Model is used to calculate the differences of log-probabilities between Wild-Type and Mutant, capturing the effect of the mutation over all positions of the sequence. The differences are then concatenated and used as features for the Fully-Connected Neural Network.

Results

Evolutionary Information aids Large Language models in capturing the specific effect of mutations

Language models for proteins are trained on extensive sequence databases such as those found in UniProt [32] and Pfam [33]. As a result, the representations they acquire primarily capture general contextual information applicable to a wide range of proteins, rather than being specifically tailored to the protein under study (Supplementary Fig.13). To probe whether the co-evolutionary information features captured by the LLM model are sufficient for the mutation prediction task, we tested the model using only the LLM-based features in the unbiased, independent DRGN test set (gray points, Fig.2e) and compared it with the model that integrates both LLM-based and EI features on a 5-fold cross-validation setting. The beneficial effect on model

performance obtained by i) introducing information from the MSA and ii) extracting features through our GMM-based approach (see Methods) is then evident from the coloured points in Fig.2e. Inclusion of evolutionary information from the MSA therefore does add additional protein-specific semantics in relation to the introduced mutation beyond the generic protein semantics captured by the language model.

D2Deep performance compared to state-of-the-art

We trained D2Deep on cancer balanced data, so avoiding reliance on prior knowledge about the gene being a known driver gene (or not). The mutated sequence's multiple sequence alignment is processed by mapping each amino acid to a 20-dimensional LLM representation vector. Subsequently, a Gaussian Mixture Model is employed for each position, calculating the differences in log-probabilities between the Wild-Type and Mutant, thereby capturing the mutation's impact across the entire sequence (Fig.1). To evaluate D2Deep's performance in classifying passenger from driver mutations, we calculated the predictions on the cancer specific, unbiased test set DRGN and compared its performance to 6 state-of-the-art predictors SIFT [26], DEOGEN2 [34], EVE [29], FATHMM-cancer [28], PolyPhen2 [35] and AlphaMissense [8]. Our method outperforms all but the FATHMM-cancer predictor as shown on Fig.2a-b. FATHMM-cancer's efficacy stems from its specialized training on the cancer mutation frequencies within the conserved domains of each protein. To study the inherent bias of the training, we segregated the mutation into the known hotspot cancer mutations and non-hotspot ones for the three best performing predictors, FATHMM, AlphaMissense and D2Deep. AlphaMissense, with the third highest performance, is trained using weak labeling, where mutations present in the general population are considered benign, and ones that are not observed labelled pathogenic. As shown in Fig.4b, the performance of AlphaMissense drops significantly between hotspot and non-hotspot mutations. D2Deep consistently performs well on the hotspot and non-hotspot mutations. Because reliable annotations for cancer related mutations are scarce, a part of the DRGN set is also included in our Training set (Supplementary Fig.14). To study how the inclusion of test mutations in the training set affects the performance of the algorithm we used cross-validation (5-folds) where the test set of each fold was excluded from the training set. Again D2Deep has comparable performance to the state-of-the-art. Here we have to note that the rest of the predictors were not retrained to exclude the test mutations and thus our predictor is more strictly tested than the others. As the predictors do not contain all protein isoforms of the test set, we show both the overlapping mutations present in all predictors (1868 mutations) and when the predictors are penalized for missing values (3603 mutations) in Fig.(3).

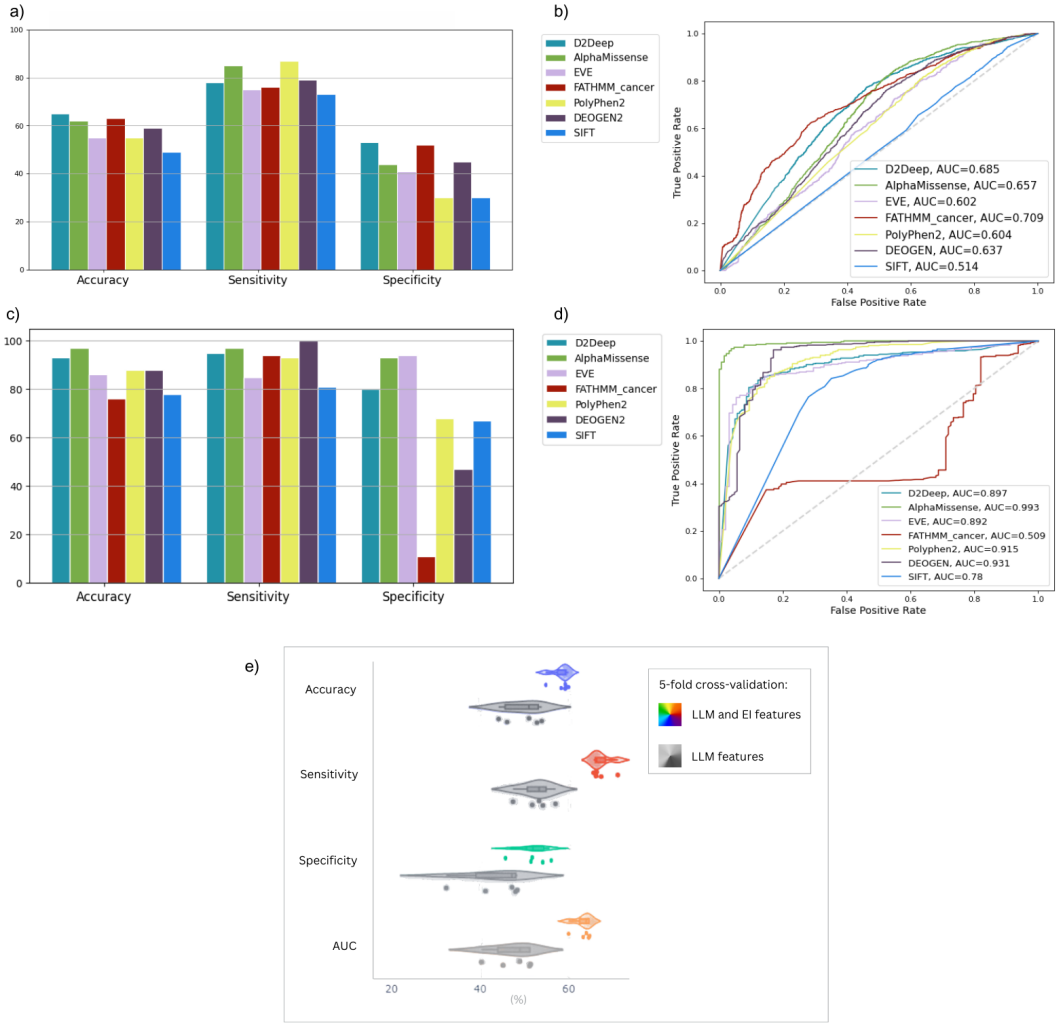

Fig. 2 a-b) D2Deep results for DRGN test set compared to existing predictors - c-d) D2Deep results for Clinvar annotations on cancer driver genes compared to existing predictors - e) Performance scores running D2Deep in 5-fold cross-validation on DRGN test set when both LLM Information and Evolutionary Information are used (different colors) compared to only LLM Information (grayscale)

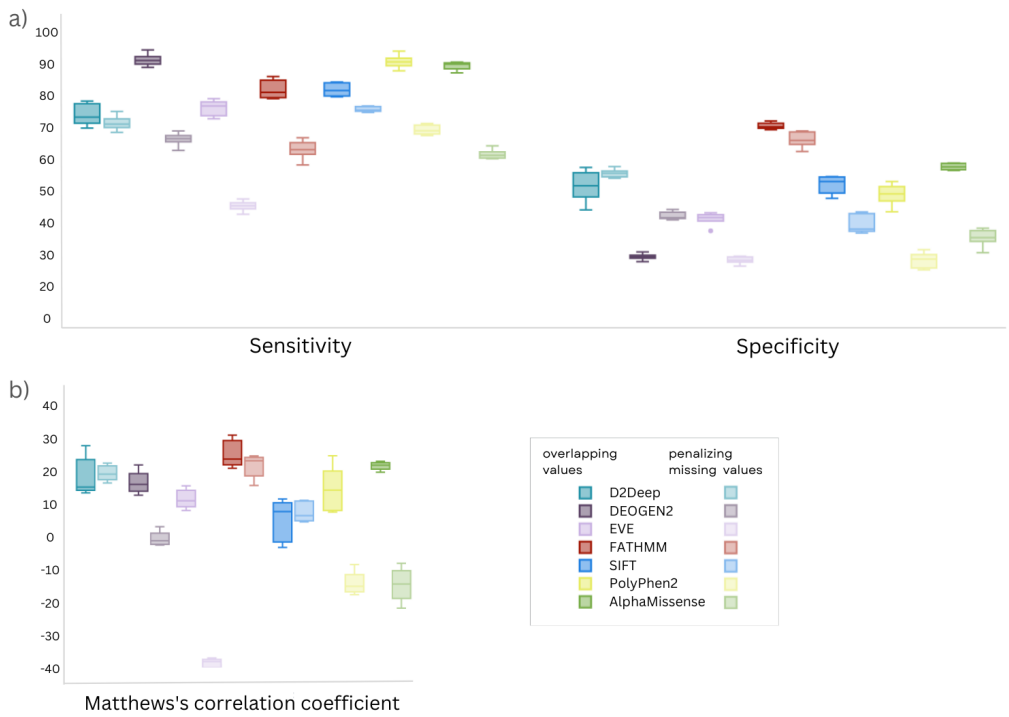

Fig. 3 5-fold validation performance of all six predictors on DRGN variants for overlapping values (intense colors) and when missing values are penalized (soft colors) - a) Sensitivity and Specificity metrics, b) Matthew's correlation coefficient

We further tested D2Deep's ability to predict fitness across cancer gene variants. To this end, we benchmarked our model on 690 confident labels from Clinvar, for which there are no conflicting annotations, for 5 cancer driver genes used in clinical setting. Here again D2Deep is comparable to state-of-the-art predictors while not trained on ClinVar data, as shown on Fig.2c-d. Overall, D2Deep predictor consistently performs well in both test sets (Fig.4a), establishing its position as a top-performing algorithm despite the fact that it exclusively relies on sequence information. In contrast, the highest-scoring predictors employ a range of contextual annotations. AlphaMissense, for example, relies on Structural context by using an AF-derived system. Polyphen-2 is trained on binding and functional site annotations as well as the output of other predictors. DEOGEN2, on the other hand, uses a variety of contextual information such as gene and pathway-oriented features. This highlights D2Deep's ability to generalize and deliver outstanding results in two substantially distinct test sets and underscores the importance of the training set utilized. Finally, D2Deep was tested on a dataset that includes only somatic and germline ClinVar mutations without conflicted annotations. The same amount of pathogenic and benign mutations were retained per protein for the germline and the somatic group, leading to a test set of 129 mutations for

each. The somatic performance was higher than for the germline mutations (Fig.4e as expected due to the somatic-specific training set used. Accordingly, excluding the germline mutations from the training set (resulting in 113 test mutations) led to a slight decrease in performance, not observed for the somatic ones. Most human genes undergo alternative splicing, which means that a single genetic variant can have different effects on various protein isoforms. Whether this variant is harmful or benign depends on its interactions with other parts of the protein. D2Deep's only requirement in sequence provides an advantage for prediction of different isoforms of the same protein and Variants of Unknown Significance compared to other predictors, as shown in the number of predicted mutations in DRGN and ClinVar tests sets respectively (Fig.4c).

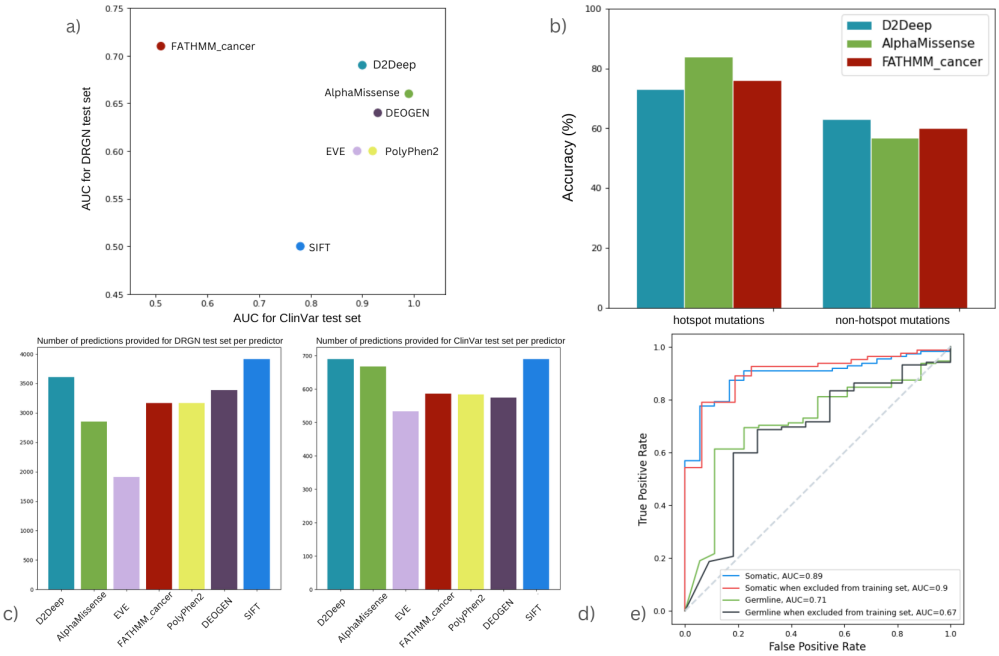

Fig. 4 a) Comparison of predictors for both Clinvar and DRGN test sets based on AUC metric - b) Performance of AlphaMissense and D2Deep on hotspot cancer mutations compared to mutations not on hotspot regions - c-d) Number of valid predictions for DRGN test set for all predictors - e) Comparison of D2Deep's performance on somatic versus germline ClinVar mutations. Somatic have higher performance that the germline ones and without drop in performance when excluded from the training set

Certainty and Confidence

In a clinical context such as cancer it is crucial to be able to assess the reliability of a prediction. We therefore explored whether we could use the log-likelihood measure inherent to Gaussian Mixture Models, which quantitatively assesses the goodness of

fit of a data point, as a proxy to reliability. A higher log-likelihood indicates a better fit, suggesting that the GMM effectively captures the underlying structure of the data. We calculated the confidence score equation using grid search as described in Methods. The weighted accuracy on the DRGN test set was increased from 64.2% to 65.3% and the high confidence mutations for genes TP53, AR, PTEN, CHEK2 and BRAF have increased performance as shown in Table Table 1. In Fig.5(column b) and Supplementary Fig.9 we show the qualitative performance when the most confident predictions (with a threshold of 65%) are kept.

Table 1 D2Deep performance on ClinVar annotations

Gene	ClinVar annotations	Accuracy (%)	Weighted Accuracy (%)	VUS annotations
TP53	280	85	91.5	627
BRAF	109	89	96.9	243
AR	116	86.2	93.5	67
CHEK2	10	90	94.2	1498
PTEN	175	81.7	96.1	586

Increase of D2Deep's performance for known cancer driver genes TP53, BRAF, AR, CHEK2 and PTEN when confidence score is used.

We tested the D2Deep predictions on independent Deep Mutational Scanning data [36], where all mutations in all protein positions are tested in relation to an experimental fitness measure such as cell function. Advantageous mutations have positive Enrichment Scores (ES), neutral mutations close to zero, while deleterious for the function of the cell have negative Enrichment Scores. For the TP53 case, D2Deep predictions have a lower value for benign and higher ($\geq$ 0.5) value for pathogenic mutations. As a consequence, a negative correlation with the Enrichment Scores is observed with a Pearson correlation of -0.46. Furthermore, we selected the confidently benign experimental mutations (ER score $\geq$ 0.2) and grouped them in groups based on D2Deep predictions as benign or pathogenic. We did the same for the confidently pathogenic experimental mutations (ER score $\leq$ -0.2) that were correctly and incorrectly classified by our predictor. The correctly predicted mutations have significantly higher confidence scores than those incorrectly predicted in both groups with Wilcoxon rank sum p-value equal to 1.52e-11. (Fig.5c).

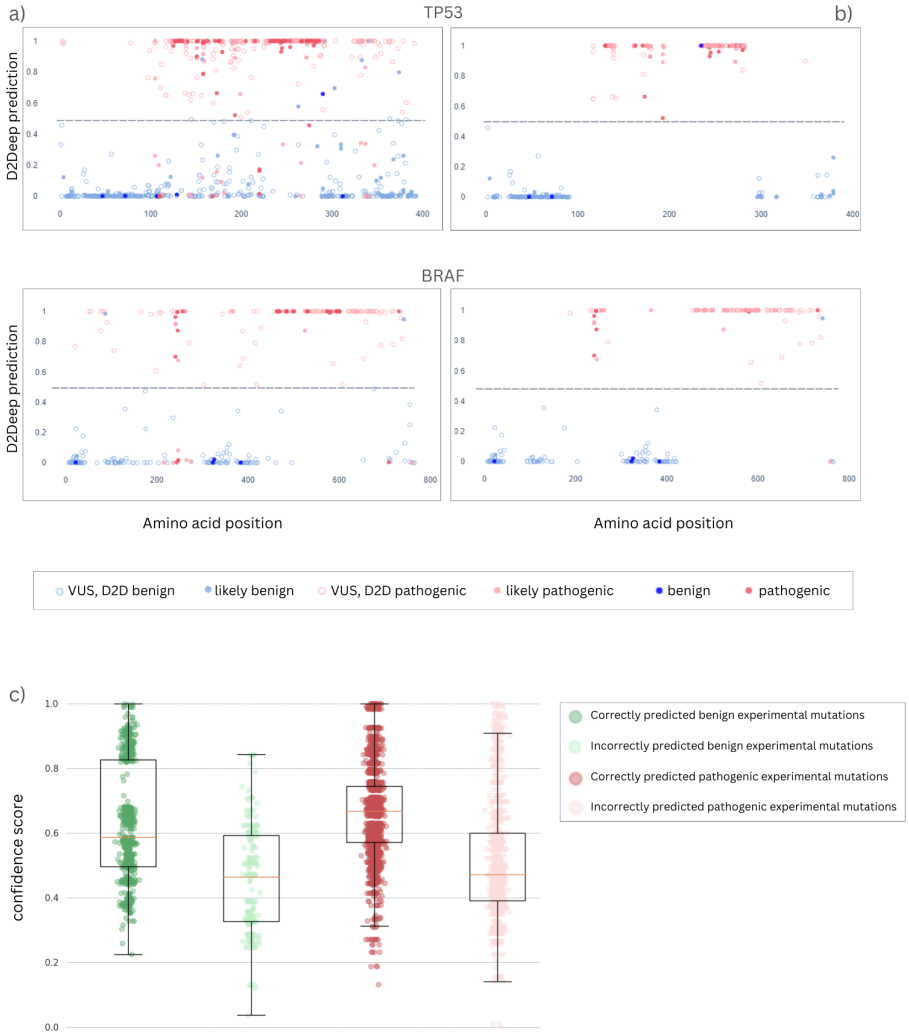

Fig. 5 D2Deep results for Clinvar annotations on 2 cancer driver genes (TP53, BRAF) when a) all D2Deeps' predictions are shown - b) high confidence predictions for Clinvar annotations are shown - c) D2Deep confidence scores of correctly (intense colors) and incorrectly (soft colors) classified mutations for TP53 gene

Interpretation of epistatic features of known and novel mutations

There are few predictors that consider the inter-dependencies between amino acids. D2Deep does so by taking into account the changes inflicted throughout the sequence when a mutation is introduced. To illustrate D2Deep's behaviour we collected benign and pathogenic mutations on the TP53 driver gene and depicted the magnitude of the effect of each mutation on other amino acids in respect to the distance from the

mutation site. Each mutation is influencing amino acids beyond its direct sequence neighborhood, so illustrating that the LLM and EI based feature space captures potential co-evolution forces. In Fig.6a we show the features of two randomly selected pathogenic mutations, E258K and C238R. Their features have the highest values, and thus influence, distant in sequence amino acids. Nevertheless, when depicting the AlphaFold2 predicted structure, it becomes apparent that the peaks of these features are, in reality, indicative of nearby amino acids in three-dimensional space (Fig.6b). This suggests that the mutation features could be used to interpret the predictions and associate novel mutations with already confirmed ones. To further investigate this, we tested the features of novel somatic driver mutations on the TP53 driver gene from three recent publications [37], [38], [39]. Two of them (Ala138Val and Ala276Ile) are predicted by D2Deep as pathogenic, with confidence 69% and 90% respectively, while Glu358Val is predicted as benign but with 50% confidence. To explore the mechanisms influencing the predictions, we again depict the features of mutated sequences Ala138Val, Ala276Ile and Glu358Val. The features capture the potential impact of the mutation on the rest of the protein's fitness and exhibit elevated values in positions coinciding with established driver mutations sourced from the COSMIC database (Fig.6d).

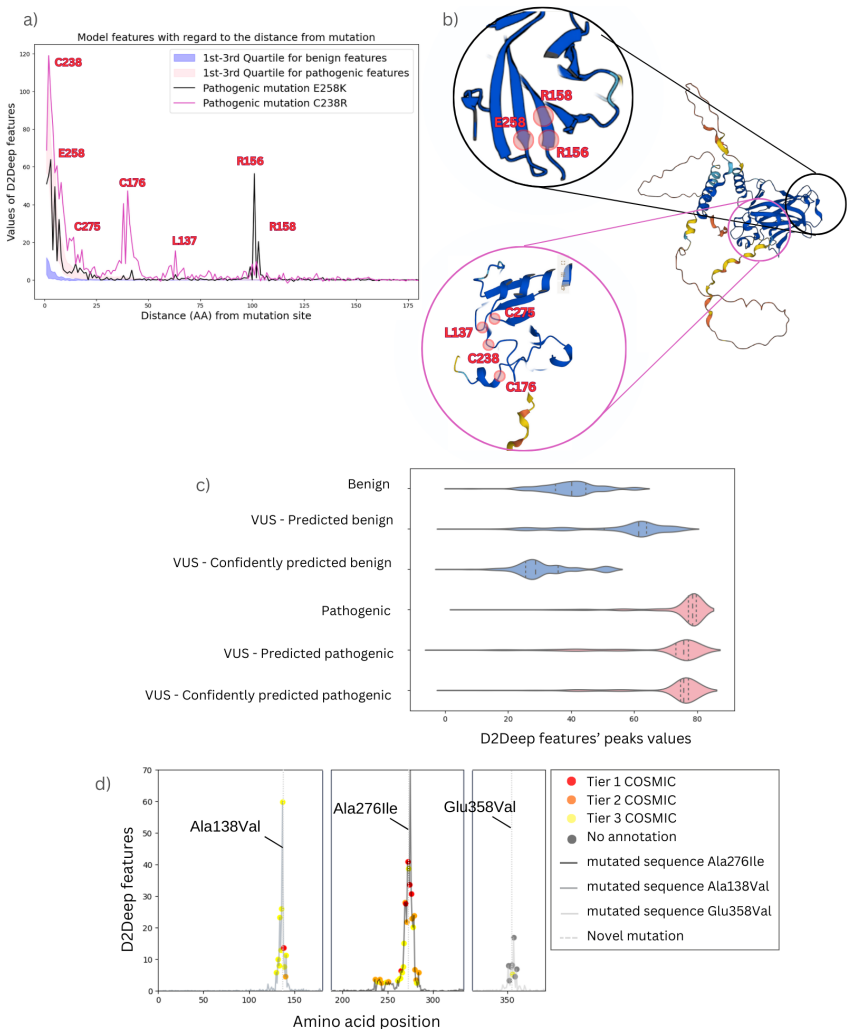

Fig. 6 a) D2Deep features used in the downstream prediction of 149 benign and pathogenic ClinVar TP53 mutations. Two pathogenic mutations E258K and C238R were considered to study the influence of features in distant amino acid and capture potential co-evolution forces.- b) Predicted AlphaFold2 TP53 structure 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 have low feature signal while ClinVar pathogenic mutations have high D2Deep signal. The VUS features follow the ground truths suggesting a correlation between known and unknown mutations. We consider the top ten 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

Moreover, when only considering the magnitude of D2Deep features, TP53 mutations labeled VUS in ClinVar that are predicted as benign by D2Deep (352 mutations) have a lower feature signal, in accordance with the feature range of confirmed benign

labels taken from ClinVar (104 mutations) Fig.6c. The same is also true for the confirmed ClinVar pathogenic mutations (176 mutations), with the tendency reinforced when we consider confident (score $\geq 50\%$) predictions. Correlating ground truth signals with predictions in this way should help in the interpretation of VUS in the clinical setting, where mutants that connect within the feature space to known driver mutations could be prioritised. Furthermore, the effect of a novel mutation is more confidently recognised by the Neural Network when known pathogenic mutations are present, which likely give similar feature signals to the novel mutation (Supplementary Fig.12).

Discussion

Distinguishing between passenger and driver mutations in cancer is a complex problem that can only be tackled by judicious use and interpretation of data. While rare mutations disrupt protein function, driver cancer mutations enable them, confer a selective advantage and are usually under positive selection. This renders predictors trained on rare mutations unsuitable for the task. We here introduce a somatic-specific approach to create relevant features that capture the effect of a mutation on a protein, in effect combining general protein sequence semantics from language models with specific protein evolutionary information.

Our novel approach models the impact of a mutation throughout the sequence, so better capturing the intricate mechanisms of epistasis whilst only using protein sequence based information. Its power is illustrated on cancer-specific and non-cancer specific test sets, demonstrating that a predictor that utilises only sequence information can in this way consistently outperform state-of-the-art predictors trained with gene, pathway and domain-specific information. As a consequence, it can provide valid predictions for sequences lacking fixed structure (e.g intrinsically disordered proteins) or domain annotations. Most notably, our method offers a quantifiable measure of the mutation's impact on the protein allowing for meaningful comparisons between known and unknown mutations. Variants of Uncertain Significance that continue to pose challenges for scientists and clinicians can subsequently be prioritized according to the latest experimental knowledge. An added advantage of this method is its statistical nature that makes possible to calculate the confidence score for each prediction, enabling better prioritization of novel mutations and setting it apart from other pathogenicity predictors. All predictions and data are available at https://tumorscope.be/d2deep (Supplementary Fig.8), which will be regularly updated with new genes. In the future, we expect to tackle the entirety of cancer-related genes, increasing the length of the input sequences greater than 2200 amino acids long.

There are certainly still pertinent challenges that have not been addressed by this work with the most important one the differentiation of the mutation's impact depending on the phenotype. It is widely recognized that mutations within the same gene, as well as identical variants, can result in varying degrees of disease, and even give rise to distinct disorders. The inclusion of cancer types or other histopathological information will therefore become necessary for even more precise predictions. Additionally,

driver mutations can remain dormant for extended periods and may initiate oncogenesis only combined with other mutations. Multiple mutations on the same gene have not been considered for this work. In the future, our approach can be synergistically integrated with supplementary contextual information, such as cancer type, and harness the power of sequence-only features to address combinations of variants, leading to enhanced predictions that will aid clinicians in identifying various cancer subtypes and ultimately facilitate more effective treatments, even when considering diseases beyond cancer.

Methods and Materials

Training set

We curated 3,093,303 SAVs from the Catalogue of Somatic Mutations in Cancer (COSMIC - Cancer Mutation Census releasev92) [40]. We excluded pseudogenes and RNA genes and kept 3825 mutations that were annotated as Pathogenic or Likely Pathogenic in cancer-related diseases in ClinVar and are recurrent in a Cancer Gene Census oncogene or in a Tumor Suppressor Gene (namely Tiers 1, 2 and 3 in COSMIC as described in Supplementary Fig.10). Their distribution in cancer-related diseases in ClinVar can be seen in Supplementary Fig. 7. We additionally collected 2,657 somatic variants with known molecular consequences from Cancer Genome Interpreter (CGI), excluding the 1,129 mutations overlapping with the COSMIC set. The CGI is an inventory of mutations in cancer genes that are experimentally demonstrated to drive tumor growth or predispose to cancer. To balance the (likely) pathogenic SAVs of each gene with a balanced benign set, we curated the UniProtKB/Swiss-Prot humsavar dataset (release 21st 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) higher than 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 [41]. 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 more than 2200 amino acids, the final training set contained 6,608 mutations, 2,956 deleterious and 3,652 benign, internally balanced within 1,012 genes. The script and data for the above workflow are publicly available (Code and Data Availability).

Test sets

DRGN

To assess the construction bias affecting the in-silico predictors, researchers introduced the DRGN set [2], comprising a total of 4,093 variants. Among these variants, 1,809 were identified as deleterious, while 2,284 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 [42]. As for the passenger mutations, they encompassed 63,525 germline variants unrelated to cancer, sourced from Humsavar [32]. To ensure a comprehensive evaluation, the final test set included genes that had at least one positive (deleterious) and one negative (passenger) sample. After the filtering out the sequences with more than 2200 amino acids, the final test set contained 3608 mutations.

Consensus Pathogenic Variants

We measured the performance of the method on important cancer driver genes included in Next Generation Sequencing (NGS) panel of biopsies of haematological and solid tumours from Compermed Guidelines (https://www.compermed.be/en/guidelines). These guidelines aim to harmonise clinical conclusions and provide generic recommendations to all stakeholders, valid independently of the platform used. 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 Pre-Trained model ProtT5-XL

We extracted the amino acid representations learnt by the best-performing protein Large Language Models from [43], called ProtT5-XL. The ProtT5-XL is a self-supervised model trained without using labeled data on the BFD100 [18] and Uniref50 [44] datasets on 2.3 million proteins. It uses the original transformer architecture which consists of an encoder/decoder with objective to predict masked amino acids using the surrounding context. By doing this, the embedding space captures essential biophysical amino acid features, such as charge, polarity, size and hydrophobicity, intrinsic semantics in protein sequences and inter-dependencies among amino acids. The resulting vector of 1024 dimensions (1024-D) was then used as an embedding of the sequence during the subsequent feature design.

Gaussian Mixture Model

The Multiple Sequence Alignments (MSA) for the training and test sets were generated with the mmseq2 algorithm using Uniref100 and PDB70 databases [45] with parameters 1 model and 3 recycles. The resulting alignment was filtered based on the protocol proposed by Hopf et al. [19]. Following this protocol we retained sequences that align to at least 50% of the target sequence, and with at least 70% column occupancy. For the five protein isoforms present in our training set (out of 1,132 total),

the above filtering resulted in less than two aligned sequences in the MSA. For those isoforms, we filtered with a more lenient cut-off of 20% sequence identity with the original protein. Each amino acid in the MSA was then mapped to the 1024-D embedding learned during the Pre-Training phase. We used the Gaussian Mixture fit method from the sklearn library 1.2.2 to estimate the Gaussian Mixture Model parameters for each column of the MSA. Our rationale for this was to attribute each mutation with a measurement of how unusual the observation is based on the distribution of the specific position within that MSA. Since the number of dimensions of the feature space (1024-D) is higher than the average sample size (the number of sequences in the MSA), so possibly leading to overfitting, we reduced the feature dimensions using maxpooling to retain 20 dimensional vectors (20-D) for every amino acid. For each MSA column of the Wild-Type (WT) protein, we fitted a GMM model to the samples of the column. Then we calculated the log-likelihood threshold under which the 1% of the samples are found. For each amino acid of the WT sequence the distance of its log-likelihood to the threshold was calculated (WT-threshold). When a mutant is introduced, the Pre-Trained embeddings of the entire sequence are affected. Thus we computed the distance of Mutant log-likelihood to the threshold (MUT-threshold) for all columns of the MSA. To construct the final feature we calculated the difference between WT-threshold and MUT-threshold for each position of the sequence. This was done to remove the effect of the quality of the fit of the model and only consider the impact of the mutant to the WT. Finally, we concatenated these differences for the entire sequence to be used as features for 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 normalisation 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. 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. 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 continue 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 (numerically more stable than using a plain Sigmoid followed by a BCELoss). The weights of the classifier were initialised using Xavier normalisation 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 aggregating 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.

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 maximized the weighted performances for all our test sets. For pathogenic predictions the overall confidence score for each mutation was set to:

$$\text{overall_confidence}_{pathogenic}(x) = \frac{1}{M}\sum_{i=1}^{M}\log(p_{GMM}(i))$$

and for benign mutations:

$$\text{overall_confidence}_{benign}(x) = 1 - \frac{1}{M}\sum_{i=1}^{M}\log(p_{GMM}(i)) \cdot 1.3$$

where M the number of samples of fitted GMM (equal to number of sequences in the MSA), $\log(\mathrm{p}_{GMM})$ the log-probability of each sample under the current model and x 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. 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 follows:

$$\sum_{i=1}^{n} w_i \cdot p_i / \sum_{i=1}^{n} w_i$$

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.

Data availability

All data used for training and testing the model are available in the public Zenodo repository: https://zenodo.org/record/8200796 (DOI 10.5281/zenodo.8200796).

Code availability

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

Supplementary Figures

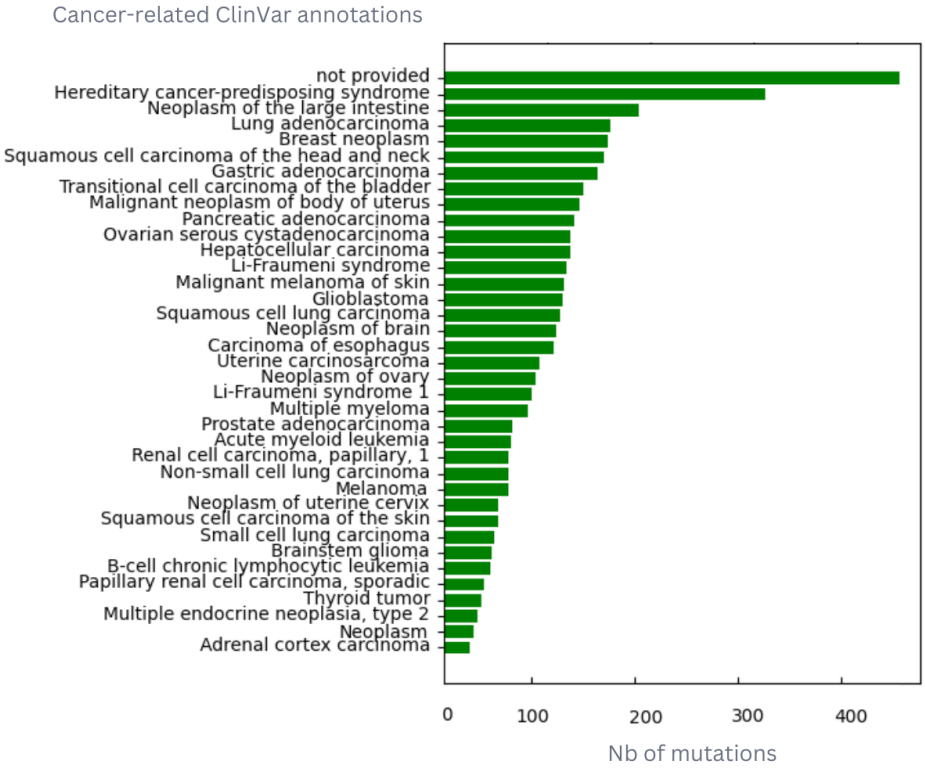

Fig. 7 Distribution of Pathogenic or Likely Pathogenic COSMIC mutations in cancer-related diseases in ClinVar (date of download: June 2023).

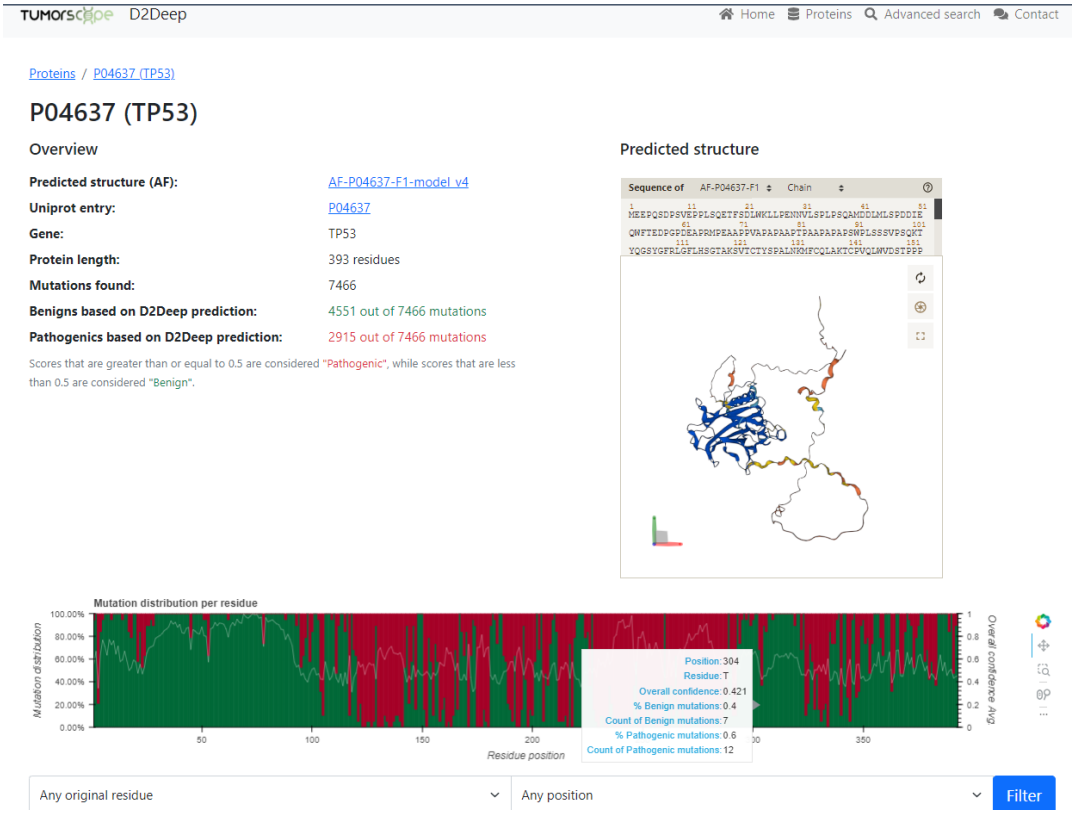

Fig. 8 Screenshot of D2Deep's webserver showing gene view for TP53. An overview of shows the effect of all possible mutations at each position of the gene with the mean confidence score of the predictions. There is also the option of search for specific positions from the option "Advanced search"

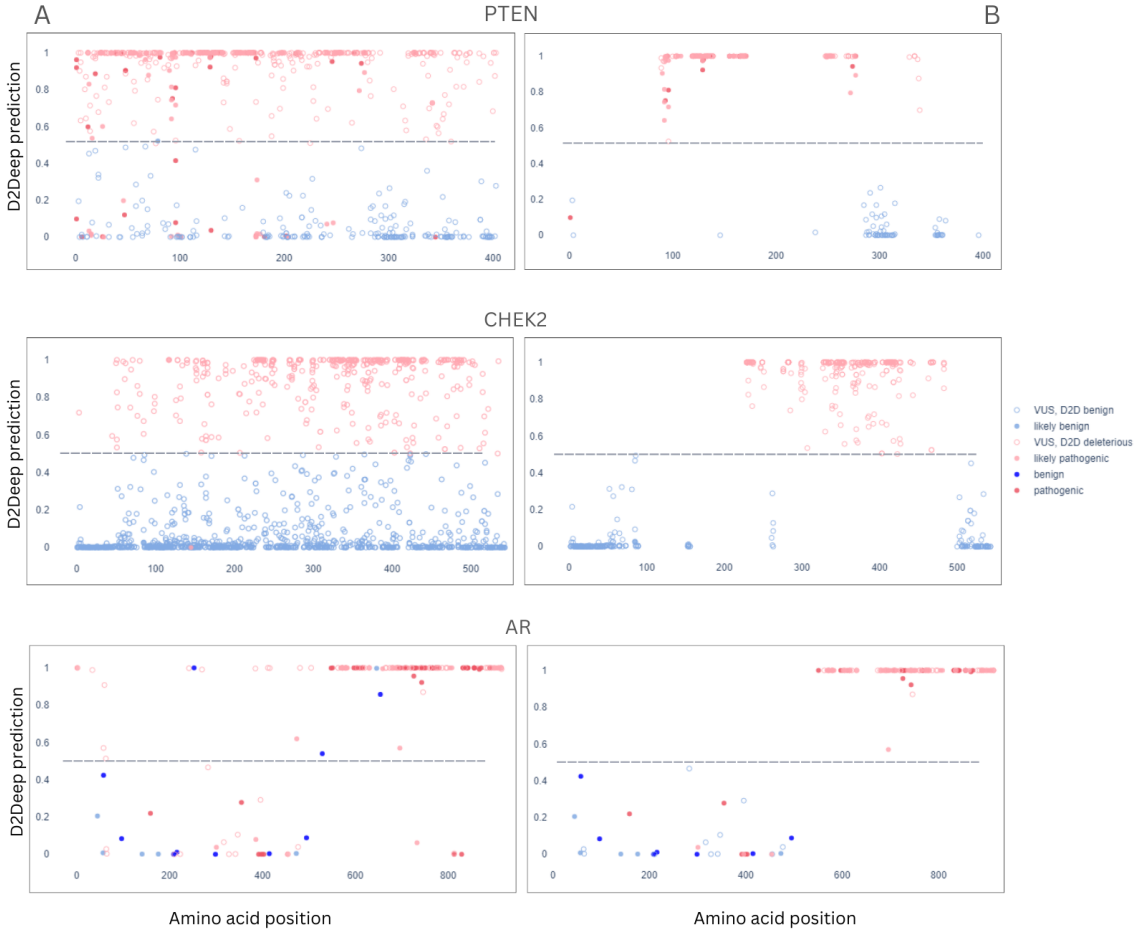

Fig. 9 D2Deep results for Clinvar annotations on 3 cancer driver genes (PTEN, CHEK2 and AR) when A: all D2Deeps' predictions are shown - B: high confidence predictions for Clinvar annotations are shown

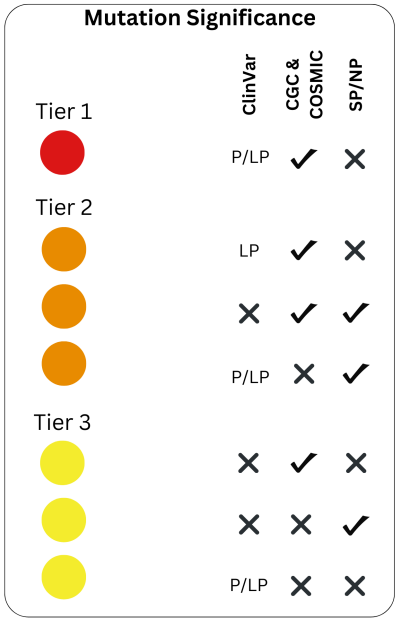

Fig. 10 Evidence criteria: 1) Clinvar: classified as pathogenic in Clinvar cancer-related diseases, 2) CGC & COSMIC recurrent missense or small inframe indel in a Cancer Gene Census (CGC) oncogene or Loss-of-function (LOF) frameshift, nonsense, large indel) in a CGC Tumour Suppressor Gene (TSG), 3) SP/NP: Evidence of positive selection as determined by dN/dS algorithm. Reference: https://cancer.sanger.ac.uk/cmc/help.

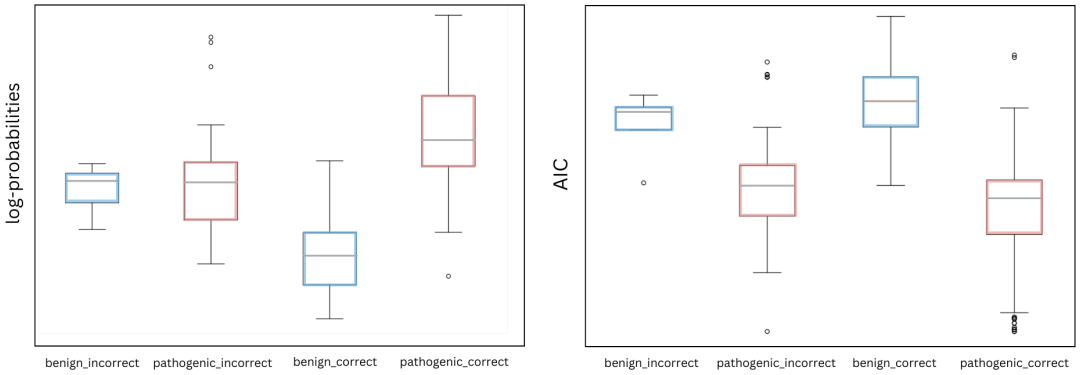

Fig. 11 Samples at MSA positions with benign mutations predicted correctly, have smaller GMM log-probability average, while for pathogenic mutations the opposite is observed. Similar conclusions are extracted by the Akaike information criterion (AIC) plot estimating the prediction error - lower values indicate more secure predictions.

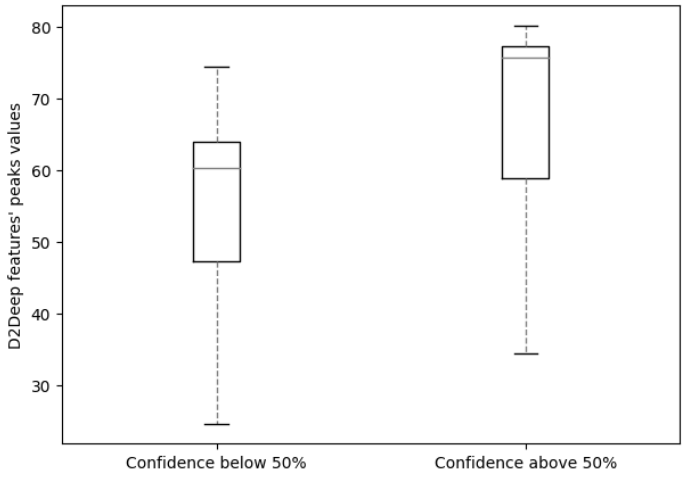

Fig. 12 TP53 VUS mutations with confidence score above 0.5 (50%) have bigger peaks than those below 0.5, suggesting that mutated sequences with stronger feature signal have higher confidence of being correctly predicted. As peaks were considered the 10 maximum values of the feature signal used by the predictor.)

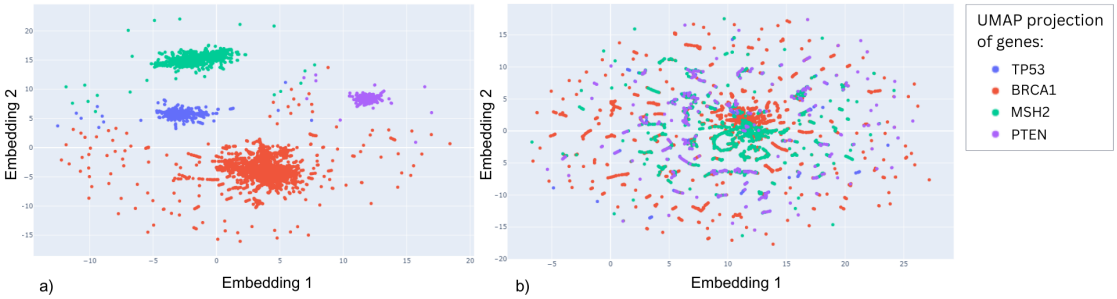

Fig. 13 a) We projected in 2D the LLM-based features for the mutated sequences for 4 genes using UMAP dimensionality reduction algorithm. Overall, the generic grammar rules learnt by the model cluster by gene and prevail over the changed induced by the mutation, even though there are regions where amino acid changes in different genes are represented by very similar vectors . b) The UMAP projection of mutated sequences using both Evolutionary Information and LLM features. The mutated sequences are distributed throughout the 2D embedding space

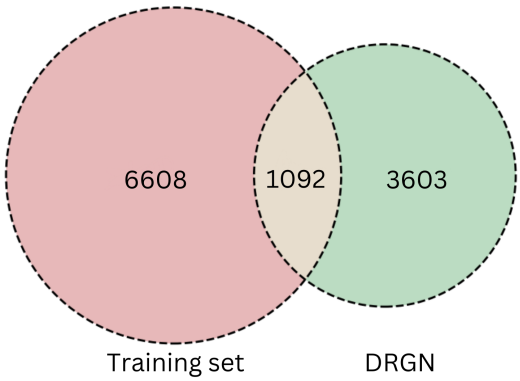

Fig. 14 Number of DRGN variants present on Training set.

Acknowledgments. Not applicable

Declarations

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

- Conflict of interest/Competing interests: Not applicable
- Ethics approval Not applicable
- Consent to participate Not applicable
- Consent for publication Not applicable
- Availability of data and materials: All data used for training and testing the model are available in the public Zenodo repository: https://zenodo.org/record/8200796 (DOI 10.5281/zenodo.8200796).
- Code availability : All code required to run D2Deep is available via our public github repository: https://github.com/KonstantinaT/D2Deep/
- Authors' 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.

References

[1] Won DG, Kim DW, Woo J, Lee K. 3Cnet: pathogenicity prediction of human variants using multitask learning with evolutionary constraints;37(24):4626–4634. https://doi.org/10.1093/bioinformatics/btab529.

[2] Raimondi D, Passemiers A, Fariselli P, Moreau Y. Current cancer driver variant predictors learn to recognize driver genes instead of functional variants;19(1):3. https://doi.org/10.1186/s12915-020-00930-0.

[3] Jin SC, Homsy J, Zaidi S, Lu Q, Morton S, DePalma SR, et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands;49(11):1593–1601. Number: 11 Publisher: Nature Publishing Group. https://doi.org/10.1038/ng.3970.

[4] Agarwal SK, Beth Kester M, Debelenko LV, Heppner C, Emmert-Buck MR, Skarulis MC, et al. Germline Mutations of the MEN1 Gene in Familial Multiple Endocrine Neoplasia Type 1 and Related States;6(7):1169–1175. https://doi.org/10.1093/hmg/6.7.1169.

[5] Sévenet N, Sheridan E, Amram D, Schneider P, Handgretinger R, Delattre O. Constitutional mutations of the hSNF5/INI1 gene predispose to a variety of cancers;65(5):1342–1348. https://doi.org/10.1086/302639.

[6] Lamlum H, Al Tassan N, Jaeger E, Frayling I, Sieber O, Reza FB, et al. Germline APC variants in patients with multiple colorectal adenomas, with evidence for the particular importance of E1317Q;9(15):2215–2221. https://doi.org/10.1093/oxfordjournals.hmg.a018912.

[7] Venkitaraman AR. Cancer Susceptibility and the Functions of BRCA1 and BRCA2;108(2):171–182. https://doi.org/10.1016/S0092-8674(02)00615-3.

[8] Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense;381(6664):eadg7492. https://doi.org/10.1126/science.adg7492.

[9] Roy DM, Walsh LA, Chan TA. Driver mutations of cancer epigenomes;5(4):265–296. https://doi.org/10.1007/s13238-014-0031-6.

[10] Greaves M, Maley CC. CLONAL EVOLUTION IN CANCER;481(7381):306–313. https://doi.org/10.1038/nature10762.

[11] Aaltonen LA, Abascal F, Abeshouse A, Aburatani H, Adams DJ, Agrawal N, et al. Pan-cancer analysis of whole genomes;578(7793):82–93. Number: 7793 Publisher: Nature Publishing Group. https://doi.org/10.1038/s41586-020-1969-6.

[12] Chen H, Li J, Wang Y, Ng PKS, Tsang YH, Shaw KR, et al. Comprehensive assessment of computational algorithms in predicting cancer driver mutations;21(1):43. https://doi.org/10.1186/s13059-020-01954-z.

[13] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021;49:D480–D489. https://doi.org/10.1093/nar/gkaa1100.

[14] Steinegger M, Söding J. Clustering huge protein sequence sets in linear time;9(1):2542. Number: 1 Publisher: Nature Publishing Group. https://doi.org/10.1038/s41467-018-04964-5.

[15] Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning;33(21):3387–3395. https://doi.org/10.1093/bioinformatics/btx431.

[16] Schelling M, Hopf TA, Rost B. Evolutionary couplings and sequence variation effect predict protein binding sites;86(10):1064–1074. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.25585. https://doi.org/10.1002/prot.25585.

[17] Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations;117(3):1496–1503. Publisher: Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1914677117.

[18] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold;596(7873):583–589. Number: 7873 Publisher: Nature Publishing Group. https://doi.org/10.1038/s41586-021-03819-2.

[19] Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation;35(2):128–135. Number: 2 Publisher: Nature Publishing Group. https://doi.org/10.1038/nbt.3769.

[20] Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. Epistasis as the primary factor in molecular evolution;490(7421):535–538. Number: 7421 Publisher: Nature Publishing Group. https://doi.org/10.1038/nature11510.

[21] Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1;33(1):268–280. https://doi.org/10.1093/molbev/msv211.

[22] Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2;11(1):5918. Number: 1 Publisher: Nature Publishing Group. https://doi.org/10.1038/s41467-020-19669-x.

[23] Weinreich DM, Lan Y, Wylie CS, Heckendorn RB. Should evolutionary geneticists worry about higher-order epistasis?;23(6):700–707. https://doi.org/10.1016/j.gde.2013.10.007.

[24] Domingo J, Diss G, Lehner B. Pairwise and higher order genetic interactions during the evolution of a tRNA;558(7708):117–121. https://doi.org/10.1038/s41586-018-0170-7.

[25] Echave J, Wilke CO. Biophysical models of protein evolution: Understanding the patterns of evolutionary sequence divergence;46:85. Publisher: NIH Public Access. https://doi.org/10.1146/annurev-biophys-070816-033819.

[26] Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function;31(13):3812–3814. https://doi.org/10.1093/nar/gkg509.

[27] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations;7(4):248–249. https://doi.org/10.1038/nmeth0410-248.

[28] Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models;34(1):57–65. https://doi.org/10.1002/humu.22225.

[29] Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data;599(7883):91–95. Number: 7883 Publisher: Nature Publishing Group. https://doi.org/10.1038/s41586-021-04043-8.

[30] Wittmann BJ, Yue Y, Arnold FH. Informed training set design enables efficient machine learning-assisted directed protein evolution;12(11):1026–1045.e7. https://doi.org/10.1016/j.cels.2021.07.008.

[31] Hsu C, Nisonoff H, Fannjiang C, Listgarten J. Learning protein fitness models from evolutionary and assay-labeled data;40(7):1114–1122. https://doi.org/10.1038/s41587-021-01146-5.

[32] The UniProt Consortium. UniProt: the universal protein knowledgebase;45:D158–D169. https://doi.org/10.1093/nar/gkw1099.

[33] Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments;28(3):405–420. https://doi.org/10.1002/(sici)1097-0134(199707)28:3⟨405::aid-prot10⟩3.0.co;2-l.

[34] Raimondi D, Tanyalcin I, Ferté J, Gazzo A, Orlando G, Lenaerts T, et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins;45:W201–W206. https://doi.org/10.1093/nar/gkx390.

[35] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2;Chapter 7:Unit7.20. https://doi.org/10.1002/0471142905.hg0720s76.

[36] Dunham AS, Beltrao P. Exploring amino acid functions in a deep mutational landscape;17(7):e10305. https://doi.org/10.15252/msb.202110305.

[37] Garziera M, Cecchin E, Canzonieri V, Sorio R, Giorda G, Scalone S, et al. Identification of Novel Somatic TP53 Mutations in Patients with High-Grade Serous Ovarian Cancer (HGSOC) Using Next-Generation Sequencing (NGS);19(5):1510. https://doi.org/10.3390/ijms19051510.

[38] Saha G, Singh R, Mandal A, Das S, Chattopadhyay E, Panja P, 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;26(1):59. https://doi.org/10.1186/s10020-020-00183-1.

[39] Wang Y, Goh KY, Chen Z, Lee WX, Choy SM, Fong JX, et al. A Novel TP53 Gene Mutation Sustains Non-Small Cell Lung Cancer through Mitophagy;11(22):3587. https://doi.org/10.3390/cells11223587.

[40] Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC);Chapter 10:Unit 10.11. https://doi.org/10.1002/0471142905.hg1011s57.

[41] Zhou W, Chen T, Chong Z, Rohrdanz MA, Melott JM, Wakefield C, et al. TransVar: a multilevel variant annotator for precision genomics;12(11):1002–1003. Number: 11 Publisher: Nature Publishing Group. https://doi.org/10.1038/nmeth.3622.

[42] Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations;10(1):25. https://doi.org/10.1186/s13073-018-0531-8.

[43] Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning;44(10):7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381.

[44] Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches;31(6):926–932. https://doi.org/10.1093/bioinformatics/btu739.

[45] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank;28(1):235–242. https://doi.org/10.1093/nar/28.1.235.