Computational and Structural Biotechnology Journal 21 (2023) 4743–4758

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Software/web server article

ASCARIS: Positional feature annotation and protein structure-based representation of single amino acid variations

Fatma Cankara [a,b,c], Tunca Doğan [a,d,e,*]

[a] Biological Data Science Laboratory, Dept. of Computer Engineering, Hacettepe University, Ankara, Turkey
[b] Department of Health Informatics, Graduate School of Informatics, METU, Ankara, Turkey
[c] Department of Computational Sciences and Engineering, Koc University, Istanbul, Turkey
[d] Institute of Informatics, Hacettepe University, Ankara, Turkey
[e] Department of Bioinformatics, Graduate School of Health Sciences, Hacettepe University, Ankara, Turkey

ARTICLE INFO

Keywords:
Single amino acid variations
Biomolecular representations
Protein sequence annotations
Protein domains
Bioinformatics tools/models
Variant effect prediction
Machine learning

ABSTRACT

Background: Genomic variations may cause deleterious effects on protein functionality and perturb biological processes. Elucidating the effects of variations is critical for developing novel treatment strategies for diseases of genetic origin. Computational approaches have been aiding the work in this field by modeling and analyzing the mutational landscape. However, new approaches are required, especially for accurate representation and data-centric analysis of sequence variations.

Method: In this study, we propose ASCARIS (Annotation and StruCture-bAsed RepresentatIon of Single amino acid variations), a method for the featurization (i.e., quantitative representation) of single amino acid variations (SAVs), which could be used for a variety of purposes, such as predicting their functional effects or building multi-omics-based integrative models. ASCARIS utilizes the direct and spatial correspondence between the location of the SAV on the sequence/structure and 30 different types of positional feature annotations (e.g., active/lipidation/glycosylation sites; calcium/metal/DNA binding, inter/transmembrane regions, etc.), along with structural features and physicochemical properties. The main novelty of this method lies in constructing reusable numerical representations of SAVs via functional annotations.

Results: We statistically analyzed the relationship between these features and the consequences of variations and found that each carries information in this regard. To investigate potential applications of ASCARIS, we trained variant effect prediction models that utilize our SAV representations as input. We carried out an ablation study and a comparison against the state-of-the-art methods and observed that ASCARIS has a competing and complementary performance against widely-used predictors. ASCARIS can be used alone or in combination with other approaches to represent SAVs from a functional perspective. ASCARIS is available as a programmatic tool at https://github.com/HUBioDataLab/ASCARIS and as a web-service at https://huggingface.co/spaces/HUBioDataLab/ASCARIS.

1. Introduction

Nonsynonymous single nucleotide variations have been associated with diseases [1,2] due to their effects, such as perturbing biological processes and impairing molecular functions of proteins by changing their stability or interactions [2–12]. Interpreting the effect of variations is important for understanding diseases of genetic origin, proposing effective treatment strategies, and developing novel biotechnological products [12]. High-throughput technologies have been producing vast amounts of variation data that awaits interpretation. However, experimental investigation of these variations remains challenging due to extensive resource-centric requirements such as labor and time. For this reason, accurate computational methods are necessary for prioritizing variants to direct experimental analysis and expedite validation. Computational approaches have been used in variation analyses, yet most of the research so far has focused only on predicting the effects of variation. On the other hand, recent developments in artificial intelligence-related technologies have led to a surge of new algorithms and methods to be used in bioinformatics and computational biology [13]. For these algorithms/methods to process biological entities, these

* Corresponding author at: Biological Data Science Laboratory, Dept. of Computer Engineering, Hacettepe University, Ankara, Turkey.
E-mail address: tuncadogan@gmail.com (T. Doğan).

https://doi.org/10.1016/j.csbj.2023.09.017
Received 16 April 2023; Received in revised form 15 September 2023; Accepted 15 September 2023
Available online 17 September 2023
2001-0370/© 2023 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

entities must be numerically represented in a meaningful manner. Therefore, there is a current need for new approaches to yield accurate, comprehensive, and reusable numerical representations (i.e., feature vectors) of sequence variations.

There are numerous computational methods/tools for predicting the effects of variations at the level of genes or proteins [3,14–41]. Most of these methods/tools aim to answer the question of whether a given substitution (or indel) can be deleterious at the molecular level and/or at the whole organism scale, by utilizing the available biological information. These methods differ regarding the employed features and implemented algorithmic techniques (please see Supplementary Information S1 for a review of variant effect prediction methods/tools). One of the current and prevailing issues in this regard, especially associated with machine learning-based approaches, is the interpretability of results [13]. It is crucial for a researcher, the tool's user, to comprehend why the method predicted that specific outcome. Another critical point here is the input data and its featurization. The choice of source/input feature type(s) and the dataset are at least as important as the algorithmic approach used. In this regard, types of features that are unexplored in the framework of variant modeling are potential subjects of investigation, where they can be combined with traditional approaches, with the ultimate aim of constructing new models with performances that are sufficiently high to effectively aid clinical decision-making.

Residue/region-specific annotations of proteins (e.g., nucleotide/DNA binding regions, active sites, modified residues, motifs, domains, etc.) provide crucial information about both their molecular functions and the biological mechanisms in which they are involved. This knowledge is collected, organized, and presented to the user in a standard format in protein-centric resources such as the UniProt database [42]. Especially, the ontology-based and protein-centric versions of these annotations (e.g., Gene Ontology term associations of whole proteins) have been studied within the context of predicting functions [43,44] and phenotypic implications [45] of proteins; however, they are under-explored in the framework of variant modeling, with only a few examples [14,46]. Essentially, there is a correlation between the effect of a SAV and its correspondence with a functional region on the sequence. Hence, residue/region-specific protein annotations hold great potential in terms of enlightening the functional implications of sequence variations.

In this study, we propose ASCARIS (Annotation and StruCture-bAsed RepresentatIon of Single amino acid variations), a new function-centric featurization approach to represent SAVs based on their spatial correspondence with residues/regions of functional importance, such as domains, active sites, binding sites, disulfide bridges, etc. Our hypothesis is simply that mutations that directly correspond to functionally important sites/regions in proteins (or mutations that are proximally located to these functional regions in the 3-D space) are more susceptible to causing deleterious effects, since these localized roles can easily be disrupted by the respective amino acid change. The originality of our work derives from the investigation of positional functional annotations of proteins and their integration with more conventional features (i.e., physicochemical and structural descriptors) for the construction of concise, effective, interpretable, and reusable variant representations.

We incorporated 30 different types of protein sequence annotations from UniProtKB (the full list is provided in Table 1). SAVs rarely correspond directly to functionally annotated positions because of their sparse nature. Due to this, we also accounted for the spatial distance between the SAV and the annotated positions/regions by utilizing the coordinates from the 3-D structure of the protein and incorporated this distance-based information into our variation feature vectors. Furthermore, we included amino acid-specific physicochemical and structural changes caused by these SAVs such as polarity, volume, and accessible surface area, together with the location of mutations in the structure, with the aim of characterizing variations in a more context-dependent manner. To obtain 3-D features, we utilized two different tracks; (i) PDB + homology modeling, and (ii) AlphaFold2 tool's structure predictions. The latter allowed the extension of our variant featurization method to proteins with completely unknown 3-D structures.

Table 1
Types of positional annotations from the UniProt database that are incorporated into the proposed SAV representations.

Annotation Class	Annotation Type	Description
Region	Coiled Coil	Positions of regions of coiled coil within the protein
	Motif	Short sequence motif of biological interest
	Region	Region of interest in the sequence
	Repeat	Positions of repeated sequence motifs or repeated domains
	Zinc Finger	Position(s) and type(s) of zinc fingers within the protein
	Calcium Binding	Position(s) of calcium binding region(s) within the protein
	DNA Binding	Position and type of a DNA-binding domain
	Nucleotide Binding	Nucleotide phosphate binding region
	Intramembrane	Extent of a region located in a membrane without crossing it
	Transmembrane	Extent of a membrane-spanning region
	Topological Domain	Location of non-membrane regions of membrane-spanning proteins
Sites	Active Site	Amino acid(s) directly involved in the activity of an enzyme
	Binding Site	Binding site for any chemical group
	Metal Binding	Binding site for a metal ion
	Site	Any interesting single amino acid site on the sequence
Amino Acid Modification	Cross-link	Residues participating in covalent linkage(s) between proteins
	Disulfide Bond	Cysteine residues participating in disulfide bonds
	Glycosylation	Covalently attached glycan group(s)
	Lipidation	Covalently attached lipid group(s)
	Modified Residue	Modified residues excluding lipids, glycans and protein cross-links
Variants	Natural Variant	Description of a natural variant of the protein
	Mutagenesis	Site which has been experimentally altered by mutagenesis
Secondary Structure	Beta Strand	Beta strand regions within the experimentally determined protein structure
	Helix	Helical regions within the experimentally determined protein structure
	Turn	Turns within the experimentally determined protein structure
Molecule Processing	Peptide	Extent of an active peptide in the mature protein
	Pro-peptide	Part of a protein that is cleaved during maturation or activation
	Signal	Sequence targeting proteins to the secretory pathway
	Initiator methionine	Cleavage of the initiator methionine
	Transit Peptide	Extent of a transit peptide for organelle targeting

ASCARIS is not a variant effect predictor, in particular. Nonetheless, as an example application of the proposed method, we trained machine learning-based classification models (using the 68-dimensional numerical features as input vectors, which are obtained from ASCARIS output data tables -originally composed of 74 columns/dimensions- by removing the 5 meta-data columns and the column representing all available domain annotations) to predict the effects of SAVs as deleterious or neutral. We trained and validated prediction models with more than 100,000 variation data points collected from the UniProtKB [42], ClinVar [47] and PMD [48] databases, and compared the predictive performance with state-of-the-art variant effect predictors. The schematic representation of the study is given in Fig. 1. One of the main

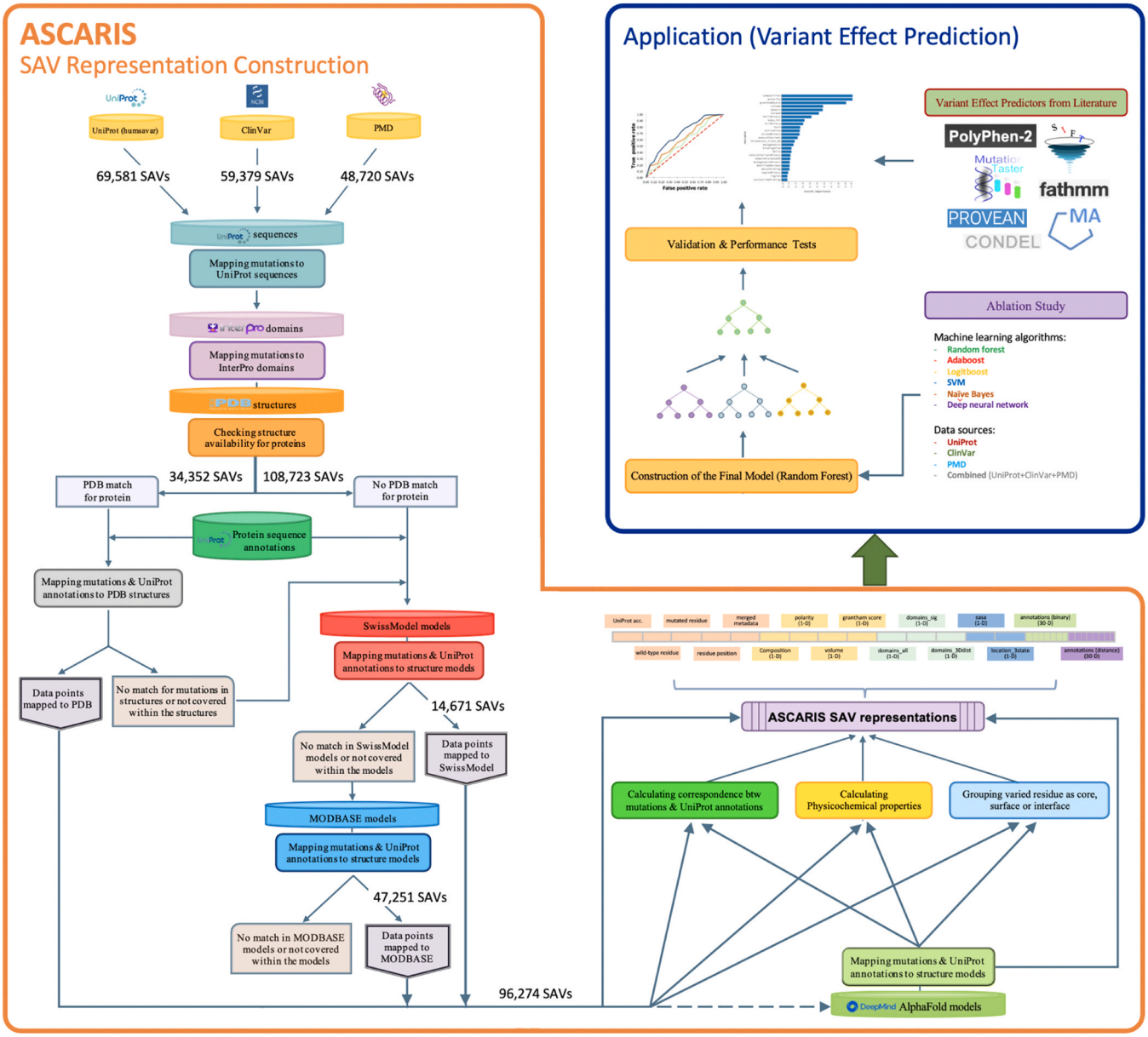


Fig. 1. The overall workflow of the proposed variant featurization/representation method, ASCARIS, together with its application to the problem of variant effect prediction.

advantages of our method is that it practically incorporates structure-related information for variant modeling using fundamental properties and residue/region-based functional features without costly molecular calculations and sequence alignments. A further advantage of our method is that it produces interpretable feature vectors, where each dimension corresponds to a predefined structural or annotation-based property. Our alignment-free featurization approach can be used to represent SAVs as concise numerical vectors to be used in various types of computational approaches, e.g., combining them with traditional conservation-based features under ensemble methods to predict the effects of variations with elevated performance and/or coverage, or as part of multi-omics-based datasets in large-scale integration and modeling of biomedical data.

2. Methods

2.1. Data

The variation datasets were retrieved from three databases, namely UniProt, ClinVar, and Protein Mutant Database (PMD). To be able to use variation data points from different databases together, we grouped each variation into one of the two classes, namely “neutral” and “deleterious”. From UniProt’s (v2019_01) human variation (“humsavar”) dataset (the current link as of 2023: https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/docs/humsavar), we obtained 40,028 polymorphisms and 29,553 disease associated SAVs for 12,519 human protein entries. Here, we labeled polymorphisms as neutral mutations, while variations associated with a disease condition are labeled as deleterious.

The second database, ClinVar, was used to retrieve clinically reported variants. Among the variation data points in the ClinVar database version 1.61 (downloaded from: https://ftp.ncbi.nlm.nih.gov/pub/

clinvar/tab_delimited/), "pathogenic" and "likely pathogenic" variants were considered members of the deleterious class, whereas, "benign" and "likely benign" variants were considered neutral. The rest of the variation data points in ClinVar were discarded since their annotated effect was considered ambiguous. ClinVar variant data points were mapped to UniProt protein sequences using Ensembl transcript IDs and the bioDBnet database [49]. As a result of these filtering and mapping operations, 17,945 benign (i.e., neutral) and 41,434 pathological (i.e., deleterious) mutations were retrieved for 4132 human proteins.

The last data source, Protein Mutant Database (PMD), includes manually curated mutations and their consequences regarding protein's stability, interaction(s) or functional changes, in terms of the severity of the effect. In the PMD SAV dataset (downloaded on February, 2020 from http://pmd.ddbj.nig.ac.jp, which is not accessible as of 2023), [+ +]

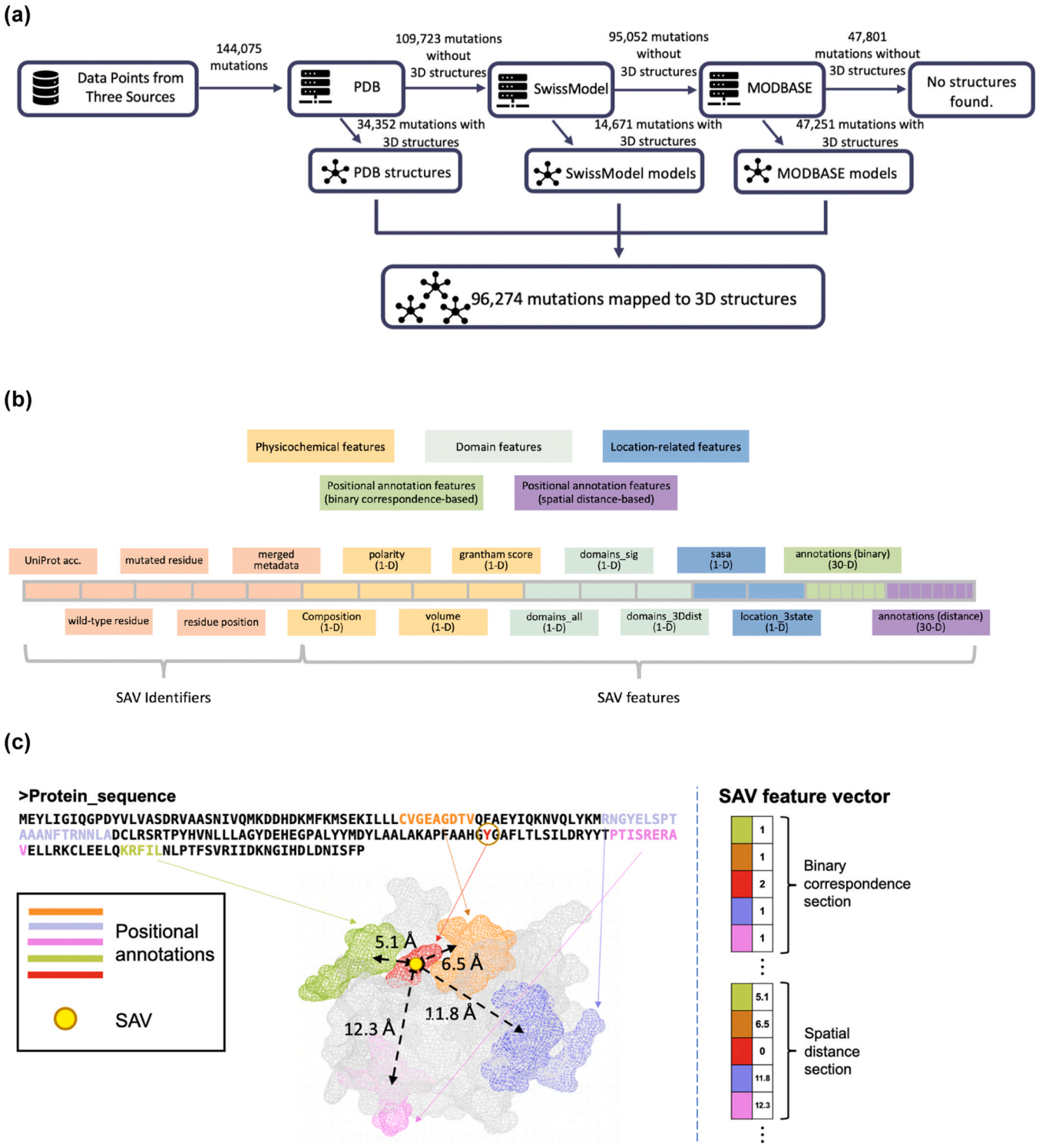


Fig. 2. Dataset construction and featurization steps: (a) the sources of structural information and statistics about the data from each source, (b) the representation of different types of features on SAV representations, (c) mapping of positional sequence annotations and the SAV onto the 3-D protein structure, and part of the SAV feature vector that corresponds to these annotations.

and [+] signs denote an increase in the activity/stability, whereas "[–]" and "[-]" denote a decrease in the activity/stability, "[0]" denotes complete loss of function, and "[=]" denotes no change. An increase or a decrease in the activity, no matter what the magnitude is, has a potential to impair the protein's native state. For this reason, variations with increase and decrease in activity and/or stability, along with the ones with complete loss of function were recorded as deleterious, and no-effect cases were recorded for the neutral class. We collected 15,348 neutral and 33,372 deleterious variations for 3401 distinct proteins (human and model organisms) from PMD.

To prepare the finalized dataset, we removed duplicate data points (Fig. S1a). Since different resources' approaches to variant interpretation change, there were also a few conflicting cases, in terms of the annotated effect of these variations, i.e., labeled as neutral in one source, and deleterious in another (Fig. S1b). Such conflicting data points were removed from our dataset. We also eliminated amino acid changes that resulted in a termination codon as they were highly biased to be deleterious. After these filtering operations, our dataset was composed of 144,075 SAVs (76,951 deleterious and 67,124 neutral) on 15,402 distinct proteins. Finally, we removed SAV data points for which 3-D structural information cannot be obtained (details of this elimination procedure is given below), making the finalized dataset of 96,274 SAVs, of which 52,512 are deleterious and 43,762 are neutral. All relevant data, including input datasets and generated data (e.g., SAV representation vectors), can be accessed on our GitHub repository at https://github.com/HUBioDataLab/ASCARIS.

2.2. Incorporation of structural information

We obtained structure files for the proteins in our dataset from Protein Data Bank (PDB) [50]. For the variation data points for which the corresponding protein's structure has not been solved at all, or that the available structures do not span the region of the protein sequence where the variation is located, homology models from SwissModel [51] and ModBase [52] have been incorporated, in respective order. In the case of availability of multiple PDB structures or models that satisfy the above-mentioned conditions, the structure with the highest resolution or the model that possesses the highest quality score is retained. Variations without any corresponding structure or model were eliminated from the dataset. Fig. 1 and Fig. 2a show the workflow of the structure incorporation process, including the number of variation data points mapped at each step.

As an alternative version of ASCARIS, we utilized the AlphaFold2 tool's [53] protein 3-D structure predictions instead of PDB and homology modeling. AlphaFold2 is a deep learning-based method developed by DeepMind that is capable of predicting monomer protein structures from primary amino acid sequences with high accuracy [53]. We used the AlphaFold version of ASCARIS mainly to assess its performance and compare it against the original version. For this purpose, we downloaded structure models for the reference human proteome (release 2021_03) from AlphaFold-DB (at https://alphafold.ebi.ac.uk/).

2.3. Featurization

ASCARIS representations contain information regarding multiple types of data including protein domains, physicochemical properties, structural location categories (i.e., core, interface or surface) and 30 different types of functional residue- or region-based annotations. This way, each SAV data point is represented by a 74-dimensional feature set including the meta-data columns (e.g., accession of the protein, wild type and mutated residues, position, etc.). Table S1 and Fig. 2b displays the names, descriptions, categories and number of dimensions that correspond to each type of feature on the representation. SAV features vectors that are used in ML-based variant effect prediction models are 68-dimensional and obtained by removing the five meta-data columns and the column representing all available domain annotations from the ASCARIS output data tables. We provided detailed information regarding each feature below.

2.3.1. Protein domains

Domain region annotations of proteins were retrieved from InterPro [54]. In some cases, multiple domains from the same hierarchy (i.e., a group of domain entries that roughly define the same structure with different levels of specificity) are annotated to the same region of a protein. In such cases, the one at the highest level in the hierarchy (i.e., the most generic one) was selected and other domain annotations were discarded. In the case of multi-domain proteins, only the domain that spans the site of variation is retained. If none of the annotated domains spans the position of variation, the one closest to the variation, in terms of the number of amino acid positions in-between, was kept and the rest were discarded. This way, we retained domain information for 96,131 SAV data points out of 144,075 SAVs in the raw variation dataset. The remaining 47,944 SAV data points have not been associated with any InterPro domains. Out of the 96,131 SAVs, 31,893 of them have distantly located (i.e., out of region) domains, and 64,238 SAVs are found to be within the domain annotated regions. The unique number of retrieved domains was 2401.

Due to the high number of unique domains in our dataset, most of which were only encountered in one or a few SAV data points, we performed a statistical analysis using Fisher's exact test to evaluate their significance considering the separation between neutral and deleterious SAVs. In other words, we observed the change in the frequency of observing deleterious mutations between the cases; (*1*) when the mutation is on the domain of interest, and (*2*) when the mutation is not on the domain of interest. Details and results of this analysis are provided in Supplementary Information S2. Based on the results, we selected the most significant domains by taking the p-values into account (Table S2). We incorporated the domain information into our SAV features using categorical variables, where each domain is encoded by its unique InterPro identifier. Domains are incorporated into two different dimensions of our SAV features (i.e., data tables): first one considers all of the domains (column name: "domains_all"), and the second one only considers the statistically significant domains (column name: "domains_sig"). The choice of using either one of them is left to the user. Throughout this study, we opted for the significant domains while constructing our prediction models.

2.3.2. Physicochemical features

Physicochemical properties are evaluated at the individual amino acid level, considering their property value changes, as the difference between the wild-type amino acid to the mutated one. Differences in three different physicochemical features, i.e., polarity, volume, and composition, together with their consensus, the Grantham Matrix Scores, are calculated for each SAV and incorporated into the four dimensions of the corresponding feature vectors as real values. Amino acid-based volume and polarity values are taken from published data [55–57]. The composition is calculated as the ratio between the atomic weight of non-carbon atoms and the total weight of carbon atoms in the side chain. These values indicate the magnitude of change in terms of physical constraints and provide a measure for the similarity/dissimilarity between the changed amino acid and the original one.

2.3.3. Structural location of the variation

We incorporated the location of the mutated residue on the structure either as core, surface, or interface region since it may provide clues about the possible effect of the mutation [58–61]. We deduced this information from relative solvent accessible surface area (rASA) values. For this, we calculated solvent-accessible surface area (SASA) values for the residues of interest using FreeSASA [62]. In order to classify residues into one of the three groups (i.e., core, interface, and surface), we applied a cut-off accessibility value of 5% to select between core and surface residues [12,63]. According to this, a residue with a rASA value

less than 5% is considered as buried, while a residue with a rASA value greater than or equal to 5% is considered to be located at the surface. In order to differentiate between surface and interface, we directly employed protein-based interface residue information from InteractomeInsider [64]. Here, we selected validated interface residues along with high quality ÉCLAIR interface predictions. When a residue, that was previously labeled as surface, is listed as an interface residue in InteractomeInsider, it is removed from the surface group and placed into the interface group. When a previously core-labeled residue is found in the interface residue list, it is removed from the core residues group and labeled as conflicting. Structural location information is recorded using 2 dimensions of our variation feature vectors: a 1-D categorical variable using core/surface/interface grouping, and a 1-D real-valued variable containing the actual rASA values. 42,976 mutations in our dataset were found to be in the surface region, 12,900 of them in the core region, and 5810 in the interface region. 724 mutations were labeled as conflicting. For the remaining data points, rASA values could not be calculated. Conflicting cases are later merged with those without any SASA values and treated as a fourth category.

2.3.4. Mapping positional sequence annotations

Sequence annotations were retrieved from the UniProt database version v2019_01 for each protein in our dataset. There are 34 different types of positional annotations in UniProt, and we selected 30 of them. The types and descriptions of these positional annotations are given in Table 1. We included the positional annotation data in two different ways. First, we identified the annotated sites/regions that directly correspond to the SAV positions on the sequence. We incorporated this information into our feature vectors by reserving 30-dimensions, where each dimension belongs to a different type of positional annotation. We used a categorical variable to signify the correspondence between an annotation and the SAV of interest, i.e., '2' if the SAV and annotation correspond to each other on the same sequence position, '1' if that type of annotation exists on the protein of interest but the annotation and the SAV are not on the exact same position in the sequence, and '0' if the annotation does not exist for that protein at all (Fig. 2c).

Second, to account for the cases where there is no direct correspondence between the annotation and the SAV, we calculated the spatial distance in-between, using 3-D structural information. Incorporation of this feature is an important element of the ASCARIS framework (and also adds value to its novelty), since it provides our model with the ability to account for the changes that might occur as a result of larger perturbations on the protein. To incorporate this, we identified the spatial location of both the SAV and the annotated residue, using the sequence-based position information and its structural correspondence. Then, we calculated the Euclidean distance between the C-alpha of both residues (i.e., SAV and the functionally annotated site/region) on the 3-D space in the unit of Angstroms (Fig. 2c). If the annotation is region-based instead of site-based, the residue that is in the closest proximity to the SAV residue is taken into account. Some of the proteins have multiple sites/regions annotated with the same type of positional annotation. In such instances, the one closest to the site of variation is retained. The proximity information is incorporated into SAV representations via 30 additional dimensions, each of which corresponds to a different type of annotation, and the value inside is the real-valued spatial distance between the SAV and the corresponding annotated residue. If there is a direct correspondence between a SAV and an annotation, the distance value is recorded as 0. If the annotation type of interest does not exist for that protein or its position is outside the structurally solved regions of the protein, we impute the corresponding cells with the median spatial distance of the respective annotation type in the whole SAV dataset. These annotation type-specific mean distance values are given in Table S3. Here, we did not use the value zero for the imputation in order to distinguish between a missing value and a true 0 distance (i.e., a case where the variation and the annotation correspond to the same position in the sequence).

2.4. Machine learning-based classification of variants

In order to measure the biological relevance of our SAV representations, we trained classification models that use our representations as input and predict the effect of query SAVs as either neutral or deleterious. We evaluated the performance of our models, first, via 5-fold cross-validation, as reported under Section 3.2.2 and second, over independent hold-out test datasets for the comparison with state-of-the-art methods, which is reported under Section 3.2.3. 8 different metrics are used for the evaluation of the performance of prediction models (i.e., accuracy, sensitivity/recall, specificity, negative predictive value - NPV, precision, F1-score, Matthew's correlation coefficient - MCC, and the area under the receiver operating characteristic curve - AUROC). Please see Supplementary Information S3 for detailed information about the metrics.

In this study, we used the random forest (RF) algorithm for the binary classification of variation data points. The random forest algorithm is an extension of decision trees where multiple trees are built, an ensemble of which are used to make a decision [65]. A randomly selected subset of a given size is drawn with replacement from the original data and trees are built with each dataset separately [64,66]. The RF algorithm randomly selects features to be used for splitting at each node, thus being less prone to overfitting.

In the analyses explained in Section 3.2, we used the default hyper-parameters of random forest which can be listed as; the number of trees: 100, the maximum number of decision splits: n-1, and the number of predictors to select at random for each split: $\sqrt{p}$, where n and p represent the number of observations and the number of predictors, respectively. For the hyper-parameter optimization using grid-search, we tested the following values; the number of trees: 50, 150, 300, and 500; the maximum number of decision splits: 3, 81, 2187, 96273; and the number of predictors to select at random for each split: 2, 8, 24, 68. We also employed additional algorithms, such as adaptive boosting (Adaboost) [67], logistic regression (LogitBoost) [68], support vector machines (SVM) [69], and the naive Bayes (NBayes) [70], for algorithmic baseline model comparison.

In this study, Python (v3.7) was employed for data pre-processing and analysis, and feature vector construction jobs. MDS and t-SNE algorithms are implemented using Python's (v3.7) sklearn (v0.21.1) library. MATLAB by MathWorks is employed for classification model development and performance evaluation. As far as we are aware, this is the only available and supported implementation to handle both categorical and real-valued variables in the same feature vector, without any limitation on the number of features.

3. Results and discussion

3.1. Exploration of the dataset and features

In this section, we identified and discussed the relationship between the effects of variations and their structural and annotation-specific properties, to evaluate the biological relevance of incorporating these features. After that, we visualized our variation dataset (based on our representation vectors) on a 2-D space via dimensionality reduction, to observe the distribution of neutral and deleterious variations coming from different sources.

3.1.1. Domain annotation-based evaluation

We examined the relationship between a variant's effect and its location with respect to annotated domain regions. We formed 3 groups for this purpose; "no domain" group signifies SAV data points where the corresponding proteins have no domain annotation in InterPro, whereas the "within domain" and "out of domain" groups signify the variations that are located inside and outside the domain annotated regions on the sequence, respectively. All available domain annotations are used for this analysis. As observed from Fig. 3a, mutations are more likely to have

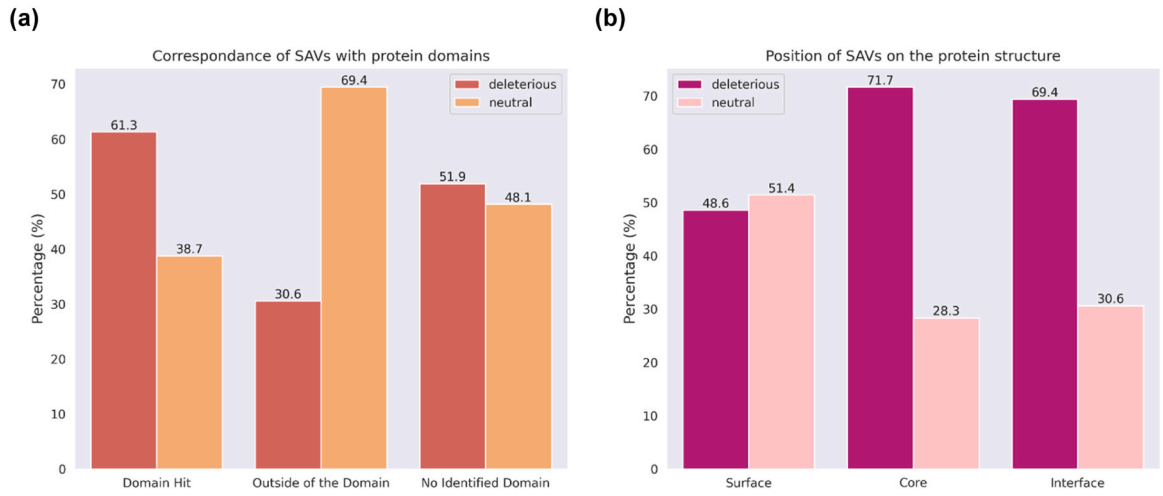


Fig. 3. . (a) Distribution of neutral and deleterious SAV data points according to their domain region correspondence. Mutations found within the domains tend to be more deleterious (61.3%) compared to the ones outside (30.6%), **(b)** distribution of neutral and deleterious SAV data points according to their location on the structure of the protein. Mutations found within the core and interface regions tend to be more deleterious (71.7% and 69.4%, respectively) compared to the ones on the surface (48.6%).

a deleterious effect when they are located within domain annotated regions (61.3%), compared to the variations that remain outside of domain regions (30.6%). We also statistically tested this observation using Fisher's exact test and found the difference in deleteriousness to be statistically significant at a 99% confidence interval (p-value < 0.01). It is expected that a mutation that is within the region of a domain is more likely to cause a deleterious effect on the functionality of the protein, compared to a mutation in a non-domain (probably disordered) region, since domains are the main structural and functional building blocks of proteins [71]. It was also observed that the percentage of deleterious SAVs among no domain regions (51.9%), which is plausible since it is highly probable that these so-called "no domain" proteins are understudied and have domains that are yet to be identified/documented. Therefore, many of these SAVs may actually reside in domain regions.

3.1.2. Physicochemical property-based evaluation

We analyzed changes in physicochemical descriptor values between the wild-type amino acid and the mutated one, in terms of polarity, volume, composition, and the Grantham score, which represents the consensus of the former three. Since these physicochemical properties are given as relative values (i.e., changes occurred due to the variation with respect to the wild-type amino acid in that position), their evaluation should be made accordingly. Here, SAVs with large physicochemical values indicate significant property changes and, thus, are expected to be deleterious, on the other hand, we expect to observe neutral variations with a higher ratio in the cases with insignificant property value changes. To test this, we applied statistical testing to each property independently.

First, we drew histograms of the value distributions (Fig. S2). Then, for each physicochemical property, we labeled each data point with conditions as "significant change" or "non-significant change" via thresholding using a cut-off value determined with respect to the whole value distribution. Thresholds were chosen to leave approximately the same number of data points per group (i.e., significant and non-significant change). Since the change in polarity, volume, and composition can also be negative, two thresholds were selected for each property type. Data points with polarity values higher than 1.6 or lower than -1.6 are considered for the "significant change" group, while data points with polarity values between -1.6 and 1.6 are considered for the 'non-significant change' group. For the volume property, the threshold values are set to -38 and 38. For the composition property, the thresholds are -0.52 and 0.52. Since the calculation of the Grantham score comprises the summation of weighted and squared values of individual property differences, the minimum value is 0, and as a result, only a positive threshold is set, which is 81 (i.e., the median of the distribution). The data points in each group are also divided into two conditions as deleterious and neutral SAVs. Counts for each group-condition combination (e.g., deleterious mutations in the significant polarity change group) are obtained and used in Fisher's exact test to calculate p-values of associations between the magnitude of physicochemical change and the variant effect. The results, considering all four types of physicochemical properties, were found to be statistically significant at the 99% confidence interval, with p-values of 0, 0, 1.5×10^{-146}, and 0 for polarity, volume, composition, and the Grantham score, respectively. These results indicate that physicochemical properties can be considered good indicators of the functional effects of SAVs.

3.1.3. Structural location-based evaluation

Here, we aimed to observe if the location of the mutation on the protein structure contains information regarding its effect. For this, we grouped variations' positions on the sequence as core, interface, or surface, according to the structural information and relative solvent accessible surface area measures (please refer to Section 2.3.3 for more information). As observed in Fig. 3b, mutations found in the core and interface regions have a higher deleteriousness rate (i.e., 71.7% and 69.4%, respectively) compared to the ones in the surface regions (i.e., 48.6%). We also tested this observation using Fisher's exact test (taking into account that the number of deleterious mutations is higher than neutrals in the overall source dataset) and found the relationship between a mutation's structural location as core, interface or surface is significantly related to being deleterious at 99% confidence interval with p-values of 0, 2.9×10^{-141} and 4.6×10^{-127}, respectively. These results were expected since core regions are critical in terms of the stability of the protein and mutations may have a destabilizing effect leading to structural changes [59,60], whereas interface regions are important because they play roles in protein-protein interactions and a mutation at these regions may prevent the formation of a protein complex or a transient interaction, causing a deleterious effect [59,61]. Thus, our results are in correlation with the literature.

3.1.4. Positional sequence annotation-based evaluation

Mutations in critical sites/regions in proteins (e.g., active sites, DNA binding regions, etc.) are generally more disruptive compared to the

ones found in other regions. Information related to these important functional sites/regions can be incorporated into models using protein sequence annotations such as the positional annotations provided by UniProt [72]. However, only a small number of proteins are associated with a certain type of positional annotation (Fig. 4a). One of the possible reasons is that some of these annotation categories are protein family/class-specific, such as the active sites of enzymes, and are expected to be annotated to enzymes only. The second reason is the fact that annotations of proteins are incomplete, and further experimental and computational analyses are required to increase coverage. Nevertheless, proteins of only 201 SAV data points (out of 96,274) have no positional annotation at all in UniProtKB. On average, an annotation category is associated with 24% of our dataset. Individual rates are shown in Fig. 4a for each type of annotation.

With the aim of evaluating the effects of mutations in functional sites/regions, we extracted the percentage of deleterious and neutral mutations in our dataset that corresponded to each of the 30 different positional annotation categories explained in Table 1. The results are displayed in Fig. 4b, which indicates a prevalence of either being deleterious or neutral for most of the categories. Here, variations that coincide with 6 annotation categories (i.e., peptide, glycosylation, coiled-coil, propeptide, signal peptide, and transit peptide) have higher rates of neutral mutations; whereas, rates of deleterious mutations are higher for 23 categories. For the category called "natural variation sites", the rates for neutral and deleterious SAVs are equal to each, since UniProt lists both disrupting and benign variations together under this category.

The results are as expected for the 23 categories that have higher rates of deleterious mutations since, for example, a mutation in the active site of an enzyme is highly likely to disrupt the enzymatic function, and thus, have an overall deleterious effect. Considering those 6 categories where the rate of neutrals is higher, we can infer that these sites/regions are not strongly related to the function of the mature protein or SAV does not tstructural properties. For example, coiled coils are dynamic and flexible regions, and they differ in terms of both length and variability, from being almost invariant to being hypervariable [73]. The flexible nature of coiled coils may explain why the percentage of neutrality is higher in variations coinciding with these regions. Other such categories, e.g., propeptides, signal peptides, and transit peptides, function as recognition sites and for targeting proteins, which are cleaved during the maturation of the protein. SAVs usually cause a partial decrease in the efficiency of the recognition and targeting processes since the patterns themselves are variable. Thus, severe impacts are rarely observed on the overall protein function [74].

In our dataset, 67% of the SAV positions coincide with at least one positional annotation (excluding the "natural variant" category annotations), which means that we cannot utilize this information for 33% of the data points, to model and predict the effect. On the other hand, even if a variation does not directly correspond to an important site/region, those that are located proximally to critical positions still tend to disrupt the intended function. For example, a mutation in the same pocket as the active site of an enzyme, where the mutated amino acid has significantly different properties from the wild-type, may have a deleterious effect on the enzymatic function. In order to take these cases into account, we utilized spatial distances between the SAV and the positionally annotated sites/regions on the 3-D structure of the protein. To observe whether the distance data contain information relevant to the effects of variations, we calculated the average spatial distances between variations and each of the 30 positional annotation categories, considering the cases where SAVs and annotations are on the same protein, but do not coincide with each other on the same residue. We plotted the curves for both neutral and deleterious SAV data points in Fig. 4c. Here, it is observed that, for most of the annotation categories, average distances are lower for deleterious variations compared to neutral ones, which indicates that the spatial distance data carry information that can be utilized for modeling and predicting the effects of SAVs.

3.1.5. Visualization of the variation space

Visualizing high dimensional data in reduced dimensions provides a means for exploring the distribution of different properties of the samples in the dataset. One obvious visualization in our case would be to compare the neutral and deleterious SAVs. Another one would be a comparison between SAVs from different source databases. For this, we conducted a dimensionality reduction analysis using both multidimensional scaling (MDS) [75] and t-distributed stochastic neighbor embedding (t-SNE) [76] on our SAV dataset and visualized the results on a 2-dimensional space (Fig. S3). Results generally indicated that it is not possible to separate neutral and deleterious variations from each other at reduced dimensions, indicating the requirement for more sophisticated methods (e.g., machine learning algorithms) to process the data and capture the non-linear relationships between data points. Details of the dimensionality reduction and visualization analyses can be found in Supplementary Information S4.

3.2. Predicting the Effects of Variations via ASCARIS

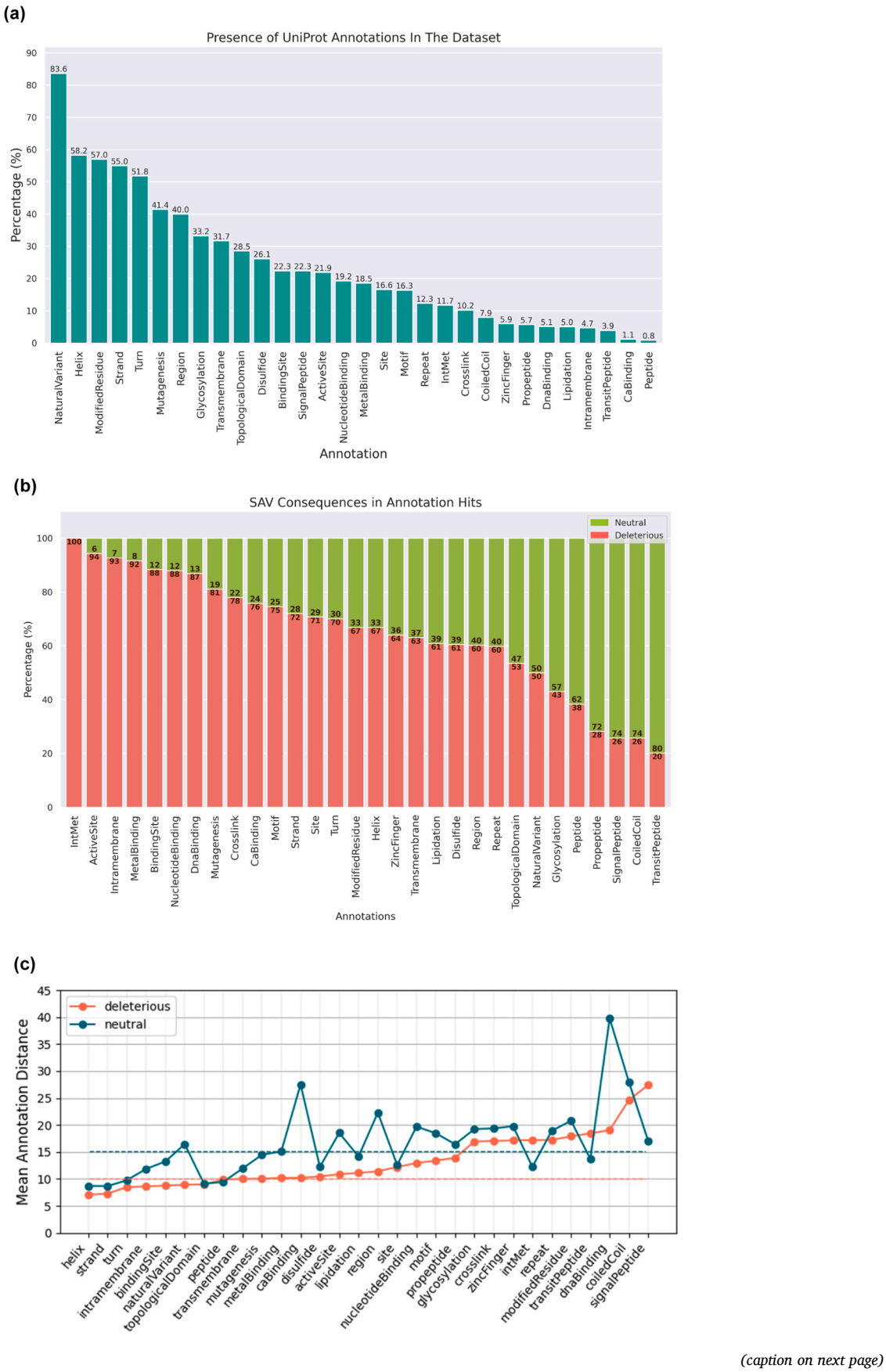
In this section, we evaluate the discriminative power of our SAV feature vectors in terms of separating deleterious mutations from neutral ones, in the framework of machine learning-based modeling, considering the PDB and AlphaFold versions of ASCARIS independently. For this, we first carried out an investigative (ablation) analysis to observe the effect of features, datasets, and algorithms. Afterward, we trained and optimized a predictive model using our multi-source variation dataset. We then compared the performance of our model against the state-of-the-art variant effect predictors on different benchmark datasets. Finally, we conducted a use-case analysis on 2 SAVs, the effects of which were correctly predicted by ASCARIS.

3.2.1. Ablation study to investigate features, data sources, and classifiers

We evaluated the predictive power of different feature types in our ASCARIS-PDB and ASCARIS-AlphaFold SAV representations by generating feature vectors with different feature combinations and training/validating a random forest classification prediction model with each version (using our combined SAV dataset of 96,274 data points). We then compared the performance results and calculated feature rankings to evaluate their importance. We measured the performance of all models via a 5-fold cross-validation analysis by keeping the data points on each validation split the same between models, and using the metrics: AUROC, accuracy, precision, recall, F1-score, and finally MCC, which is considered the main metric due to the slight imbalance between neutral and deleterious SAVs in some of the test datasets. It is important to note that our training data is relatively balanced, therefore, training of our model was not affected by such an issue.

Our first set of models (i.e., p1 and a1) only contained domain annotation information, which is incorporated in the form of identifiers of InterPro domains where the SAV of interest resides in the protein. This model incorporates domain annotations, considering 307 domains that were found to be statistically significant in terms of separating neutral and deleterious SAVs from each other in Fisher's exact test analysis (please see Methods subsection 2.3.1 and Supplementary Information S2). The feature vectors of this model are composed of a single dimensional variable that contains 308 categories (i.e., 307 significant domains displayed in Table S2 and an additional category to accommodate the rest of the domains and the "no domain hit" cases). Our second set of models (i.e., p2 and a2) incorporates physicochemical properties by generating 4-dimensional feature vectors containing real-valued polarity, composition, volume, and Grantham scores. Our third set of models (i.e., p3 and a3) is composed of two dimensions related to the location of the SAV on the protein sequence: (1) the solvent accessible surface area value, and (2) its categorization as "core", "surface" or "interface".

Performance comparison between the first 3 PDB-based models (Table 2a) indicated that physicochemical features are notable indicators for variant effect prediction together with domains (MCC: 0.26



(caption on next page)

Fig. 4. . (a) The coverage of each annotation category on the proteins in our SAV dataset (e.g., nearly 23% of proteins in our dataset have at least one active site annotation in UniProtKB), **(b)** the rates of deleterious vs neutral variations for each annotation category, which are calculated considering the SAV data points in our dataset that coincide with a positional annotation on the same residue, **(c)** Mean spatial distances between annotated residues and mutated residues in the structure, calculated independently for each annotation category, considering the cases in which the SAV and annotated residue do not coincide on the same residue. In the case of the existence of multiple annotations on a protein, only the annotation that is located spatially closest to the mutation of interest is taken into account. Dashed lines indicate overall averages calculated by taking the mean of all distances. For the calculation of Euclidean distances between residues, the coordinates of Cα of the respective amino acids are extracted from PDB models.

and 0.24, respectively). These findings are in accordance with the literature, as previous studies also highlighted the importance of physicochemical features for variant effect prediction [77] and the usefulness of domain annotations in modeling functions [71] and ligand interactions [78] of proteins. As our fourth set of models, we integrated features of the first 3 models and trained a new one with these integrated features, which resulted in a significant performance increase (MCC: 0.42 and 0.48 for p4 and a4, respectively), indicating their complementarity. In the fifth set of models, we utilized positional feature/sequence annotations in terms of one-to-one correspondence between the annotated positions and SAVs on the sequence. This binary 30-dimensional feature vector resulted in a high performance (MCC: 0.56 and 0.54 for p5 and a5, respectively) which points out to the effectiveness of this approach. This also happens to be the main contribution of our study to the literature. Furthermore, adding 30 more dimensions corresponding to the spatial distances between the annotated residues and the SAV of interest (as our 6th set of models) further increased the performance (MCC:0.59 and 0.61 for p6 and a6, respectively). The 7th set of models (i.e., p7 and a7) measures the predictive performance when the binary positional annotations are added to the p4/a4 models. This addition results in a tremendous increase in all of the metrics (MCC: 0.60 for both p7 and a7) and shows the importance of positional annotations. In our 8th and final set of models, we incorporated all features in 68-dimensions, which displayed the best performance in terms of all metrics (MCC: 0.61 and 0.63 for p8 and a8, respectively). Based on the fact that the maximum predictive performance has been achieved by the model that incorporates all types of features, we decided to construct our finalized variant effect prediction model using all features.

With the aim of selecting the classification algorithm, we performed a cross-validation based analysis to compare the performance of random forest classifier (RF), Adaptive boosting (Adaboost), adaptive logistic regression (LogitBoost), support vector machine (SVM), naive Bayes (NBayes) and fully connected feed forward deep neural networks (FFNN). Widely accepted and default hyperparameter values were used for all classifiers, such as the maximum number of decision splits: n-1, and the number of predictors to select at random for each split: $\sqrt{n}$ (n represents the number of predictors/features), the number of ensemble learning cycles: 100, minimum number of leaf node observations: 1 for all ensemble-based methods, learning rate: 1 for Adaboost and LogitBoost, kernel function: linear and kernel scale: 1 for SVM, and distribution type for categorical and real values predictors: multivariate multinomial and Gaussian, respectively. Since the performance of deep neural network models are highly dependent on the selected hyperparameters, we carried out a Bayesian search-based optimization run and found the best parameters as: number of hidden layers: 3, layer sizes: [143 64 32], and activation function: sigmoid (default optimization algorithm: Levenberg–Marquardt). The results of the 5-fold cross-validation are given in Table 2b, which indicates that RF is the most successful classifier. With the observation of these results, we decided to base our predictor on the RF algorithm.

Finally, we performed an analysis to observe the performance of the models trained on SAV datasets from individual data sources; i.e., ClinVar, UniProt, and PMD, to observe which source yields a higher generalization power to the model, and evaluate whether combining data from 3 different SAV resources under one model, or using only one of the data sources, is the better approach for training our final predictive model. We prepared one test dataset composed of 7694 SAVs (8% of the whole combined dataset) by taking a nearly equal number of data points from each data source, and used it as a hold-out test dataset to calculate the performance of all models. Three training datasets, each composed of SAV data points from an individual data resource and containing 20,876 SAVs (the size of the smallest individual SAV dataset), have been prepared. SAVs in the test dataset have already been excluded from these training datasets. Three predictive models were trained with these datasets. For all of these models, the finalized full set of features were incorporated into SAV representation vectors. Table 2c shows the performance results of these three individual data resource models, along with the model that utilized the combined training dataset. According to MCC scores, the combined dataset model provided the best performances with 0.61 and 0.63 for PDB and AlphaFold versions of ASCARIS, respectively (Table 2c). In terms of individual-dataset models, the UniProt dataset led to the highest performances (MCC: 0.43 and 0.42) followed by ClinVar (MCC: 0.34 and 0.40) and PMD (MCC: 0.13 and 0.13). This could be due to the way SAVs are classified in these databases (i.e., UniProt and ClinVar focus on the reported effect in terms of associations with diseases, whereas PMD focuses on effects related to protein's structural stability). Since the performance of the model using the combined dataset is the best, we based our predictive method on this training dataset and used it further in this study.

An interesting observation here is that ASCARIS-AlphaFold generally performs slightly better compared to the PDB version. We believe the reason is not related to the quality of the models but their coverage, since AlphaFold provides full structural coverage over the entire protein sequences, whereas in the PDB version, less than half of the SAVs and positional annotations could be mapped to regions covered by a PDB model, and the rest were resolved using SwissModel or MODBASE models (Fig. 2a).

3.2.2. *Training, validation and evaluation of the finalized model*

We built our finalized variant effect predictor model using all types of variables in feature vectors (with significant domains, omitting the "all domains" dimension), the merged training dataset from all three sources, and the RF algorithm. In these analyses, we did not apply a feature selection procedure since the ratio between the number of feature vector dimensions (i.e., 68-D) and size of our dataset (i.e., ~100,000 samples) was acceptable. Due to this, we allowed the ML model to automatically select the relevant features from the constructed vector.

We employed a grid-search-based hyper-parameter optimization test via 5-fold cross-validation. The evaluated hyper-parameter types and their respective values are explained in Section 2.4. The selected values at the end of the optimization process are; number of trees: 300, maximum number of decision splits: 96273 (size of the whole dataset - 1), and number of predictors to select at random for each split: 8. The detailed results of hyper-parameter optimization tests can be found in Table S4. We measured the final performance of our model on a hold-out test dataset, which corresponds to 10% of the data points in our original dataset. According to the results of this analysis, our ASCARIS-PDB and ASCARIS-AlphaFold models perform with; AUROC: 0.89 and 0.90, accuracy: 0.81 and 0.82, recall: 0.85 and 0.88, precision: 0.81 and 0.81, F1-score: 0.83 and 0.84, and MCC: 0.62 and 0.64, respectively.

To explain/interpret our models, we calculated feature importance values. Feature importance ranking of the optimized models shows that "significant domains", solvent accessible surface area and physicochemical features are the most critically important determinants of the decision process (Fig. S4). Positional annotations vary in their

Table 2
Performance results of the ablation study for ASCARIS-PDB and ASCARIS-AlphaFold where different; **(a)** combinations of features are tested, **(b)** classification algorithms are evaluated, and **(c)** training data resources are analyzed, by benchmarking on the same hold-out test dataset.

(a)

	Model #	Content of the feature vector					Performance scores					
		Domains	Physico-chemical features	Location in the structure	Positional Annotations (binary)	Positional Annotations (distance)	AUROC	Accuracy	Recall	Precision	F1-score	MCC
ASCARIS-PDB	p1	+					0.68	0.63	**0.90**	0.61	0.73	0.25
	p2		+				0.68	0.63	0.70	0.65	0.68	0.26
	p3			+			0.61	0.58	0.81	0.59	0.68	0.14
	p4	+	+	+			0.78	0.71	0.75	0.73	0.74	0.42
	p5				+		0.86	0.78	0.82	0.79	0.80	0.56
	p6				+	+	0.88	0.80	0.83	0.81	0.82	0.59
	p7	+	+	+	+		0.88	0.80	0.85	0.80	0.82	0.60
	p8	+	+	+	+	+	**0.89**	**0.81**	0.85	**0.81**	**0.83**	**0.61**
ASCARIS-AlphaFold	a1	+					0.68	**0.90**	0.61	0.72	0.63	0.25
	a2		+				0.68	0.70	0.65	0.68	0.63	0.26
	a3			+			0.68	0.85	0.63	0.72	0.65	0.29
	a4	+	+	+			0.81	0.84	0.73	0.78	0.74	0.48
	a5				+		0.85	0.82	0.77	0.80	0.77	0.54
	a6				+	+	**0.89**	0.85	0.80	0.83	0.81	0.61
	a7	+	+	+	+		0.88	0.86	0.79	0.83	0.80	0.60
	a8	+	+	+	+	+	**0.89**	0.87	**0.81**	**0.84**	**0.82**	**0.63**

(b)

	Classification algorithm	Performance results					
		AUROC	Accuracy	Recall	Precision	F1-score	MCC
ASCARIS-PDB	Random forest	**0.90**	0.86	**0.81**	**0.84**	**0.82**	**0.63**
	Adaboost	0.84	0.80	0.78	0.79	0.76	0.52
	Logitboost	0.87	0.82	0.79	0.81	0.79	0.57
	SVM	0.62	0.31	0.69	0.43	0.55	0.16
	Naive Bayes	0.75	**0.90**	0.62	0.73	0.64	0.28
	Deep neural network (FFNN)	0.84	0.81	0.80	0.80	0.78	0.56
ASCARIS-AlphaFold	Random forest	**0.90**	**0.88**	**0.81**	**0.84**	**0.82**	**0.63**
	Adaboost	0.85	0.84	0.77	0.80	0.78	0.55
	Logitboost	0.87	0.85	0.79	0.82	0.79	0.58
	SVM	0.57	0.53	0.60	0.56	0.55	0.11
	Naive Bayes	0.57	0.83	0.68	0.75	0.69	0.38
	Deep neural network (FFNN)	0.85	**0.88**	0.71	0.78	0.74	0.47

(c)

	Training dataset source	Performance results					
		AUROC	Accuracy	Recall	Precision	F1-score	MCC
ASCARIS-PDB	UniProt	0.8	0.59	**0.86**	0.7	0.69	0.43
	ClinVar	0.77	0.85	0.71	0.78	0.7	0.34
	PMD	0.64	**0.92**	0.64	0.75	0.63	0.13
	Combined (UniProt + ClinVar + PMD)	**0.89**	0.86	0.84	**0.85**	**0.82**	**0.61**
ASCARIS-AlphaFold	UniProt	0.78	0.56	**0.87**	0.68	0.68	0.42
	ClinVar	0.79	0.82	0.75	0.78	0.72	0.4
	PMD	0.67	**0.94**	0.64	0.76	0.63	0.13
	Combined (UniProt + ClinVar + PMD)	**0.9**	0.89	0.84	**0.86**	**0.83**	**0.63**

importance ranking; however, it is important to note that these annotations are scarce, thus, the missing information may cause some of the annotation categories to rank lower. Among them, the spatial distance to previously reported mutagenesis sites is the most critical. We did not discard any features at the end of this analysis due to the fact that the importance values of features vary depending on the dataset used; consequently, features at the bottom of the importance table may become essential on a different dataset.

It is also important to note that variations recorded in natural variant and mutagenesis variables in our feature vectors do not correspond to the SAV data points in our variant effect prediction model training/validation/test datasets. As a result, there is no data/information leak from training to test.

Mainly due to the simplicity of our feature vectors, the proposed model had convenient run times, such that, training of the full model with the optimal parameters took 91 s, running the whole hyper-parameter optimization analysis with 5-fold cross-validation and grid search (64 hyperparameter values sets * 5 folds = 320 training/validation runs) required 7 h, and predicting the effects of 10,000 SAVs in the hold-out test dataset (with the pre-trained model) took 3 s on an 8-core 2.3 MHz Intel i9 CPU with 16 GB memory.

3.2.3. Performance comparison with other methods

In order to compare our model to the state-of-the-art methods, we performed five different benchmark analyses against widely-used variant effect predictors (VEP). The first four benchmarks had similar characteristics; therefore, we detailed the analysis and the results of the first one below, and placed the second, third and the fourth in Supplementary Information. The fifth and final benchmark is relatively new and involves deep mutational scanning data together with numerous widely-used VEPs.

In the study by Schwarz et al., authors compared the performance of their method, MutationTaster2 [29], to that of SIFT [24], PROVEAN [21] and two different versions of PolyPhen-2 [16] on multiple benchmark datasets [29]. Here, we used the main benchmark test dataset from Schwarz et al., that contains 2600 variation data points from ClinVar and the 1000 Genomes project [79]. We created feature vectors for the variation data points in the *MutationTaster* dataset. To yield a fair comparison, we filtered our training dataset by first, removing the hold-out test data points, and second, by removing all SAV data points that entered our source databases (i.e., UniProt, ClinVar and PMD) at and after the year 2014, so that our training data would be temporally consistent with the training datasets used in Schwarz et al. We re-trained our RF model using default hyper-parameter values. The results are displayed in Table 3, where two versions of ASCARIS (i.e., PDB and AlphaFold) are shown together with methods from the literature. ASCARIS-AlphaFold was among the top three in terms of accuracy and F1-score (after MutationTaster2) and the best in terms of precision and specificity. The predictive performances of all competing methods on this test dataset are considerably high, which decreases the capacity for discriminating competing methods. To address this issue, we constructed a challenging sub-set by selecting SAV data points that are correctly predicted by half or less of the competing methods (i.e., < 4 methods), including ours (for fair comparison). Then we used this challenging sub-set, which is composed of 167 SAVs (74 neutral and 93 deleterious), as our hold-out test set and calculated the performance metrics of all methods. These challenging SAVs are provided in Table S5. According to performance results in Table 3, our method (ASCARIS-AlphaFold) was the best performer, followed by MutationTester2, which indicates that our approach performs well on challenging/difficult cases. The reason behind observing low performance values here is deliberately selecting mostly inaccurately predicted data points in this analysis. To observe whether our method produces complementary results to others, we analyzed prediction similarities/intersections among methods. Fig. S5 displays the number of intersecting predictions on this challenging dataset, in terms of neutral SAVs and deleterious SAVs via Venn diagrams (in panels a and b, respectively) [80]. As shown, our method has the highest number of distinct predictions in this benchmark, especially for neutral SAVs. Intersections among the state-of-the-art methods are much higher compared to intersections between our method and the state-of-the-art methods, indicating the value of the proposed approach in terms of complementing the widely-used alignment and/or structure-based methods. This can be attributed to our annotation-based featurization approach, as it is marginally different from the widely-used state-of-the-art methods included in this analysis. Due to the elevated performance of the AlphaFold version of ASCARIS compared to the PDB version, we continued the remaining benchmarking analysis only with ASCARIS-AlphaFold.

The benchmark dataset number 2, 3 and 4, namely predictSNP, VariBench and SwissVar, were retrieved from Grimm et al. [81]. Information about these datasets is provided in Supplementary Information S5. Also, performance results obtained on these datasets are given in Table S6. ASCARIS displayed considerably high performances on these benchmarks, as well (Table S6), and showed that it is on par with the methods that were specifically designed for high performance variant effect prediction. This indicates our approach has the ability to successfully represent SAVs regarding their functional consequences. The input benchmark datasets and ASCARIS feature vectors generated for each dataset can be accessed in our GitHub repository at https://github.com/HUBioDataLab/ASCARIS.

In our fifth benchmark analysis, we employed a study by Livesey & Marsh [82]. In this work, the authors assessed the performance of 46 VEPs using data from 31 previously published deep mutational scanning (DMS) experiments. DMS experiments allow the quantification of the functional impact of a high number of mutations, usually covering all possible single amino acid substitutions for the selected positions on the sequence (in some cases, across the entire protein), in one experiment. Therefore, they generate valuable data for variant prioritization and allow the direct identification of pathogenic variants on a large scale. Variant effect datasets that are produced by the DMS framework can also be used to benchmark and assess the performance of VEPs. For this purpose, the authors formed a dataset by selecting a subset of pathogenic missense mutations from ClinVar [47] and benign mutations from

Table 3
Variant effect prediction performance comparison on the MutationTaster dataset on the full dataset and its challenging sub-set. The best performances are shown in bold font for each metric.

Method name*	Performance on the challenging sub-set						Performance on the full dataset					
	NPV	Specificity	Recall	Precision	F1-score	Accuracy	NPV	Specificity	Recall	Precision	F1-score	Accuracy
PPH2_div	0.10	0.09	0.34	0.32	0.33	0.23	0.85	0.83	0.91	0.89	0.90	0.88
PPH2_var	0.20	0.26	0.16	0.21	0.18	0.20	0.80	0.89	0.87	0.93	0.89	0.87
MT	0.33	0.30	**0.53**	0.49	0.51	0.43	**0.90**	0.87	**0.94**	0.92	**0.93**	**0.91**
SIFT	0.22	0.26	0.26	0.30	0.28	0.26	0.82	0.85	0.88	0.90	0.89	0.87
PROVEAN	0.21	0.27	0.18	0.24	0.21	0.22	0.80	0.87	0.86	0.91	0.89	0.86
ASCARIS_pdb	0.46	0.53	0.40	0.47	0.43	0.46	0.62	0.82	0.58	0.80	0.67	0.69
ASCARIS_alphafold	**0.56**	**0.85**	0.47	**0.80**	**0.59**	**0.64**	0.76	**0.95**	0.82	**0.96**	0.88	0.87

* NPV: Negative predictive value, PPH2_div & PPH2_var: PolyPhen-2 w/ HumDiv & HumVar classifiers, MT: MutationTaster2.

gnomAD [83] for particular genes, details of which can be found in the original article [82]. We aimed to compare ASCARIS with 46 different VEPs on the abovementioned DMS data. To achieve this, we first removed all data points that are present in both our dataset and the DMS test set from our training, and then examined the date of the data from the original study to ensure temporal consistency and a fair comparison. We then trained ASCARIS models with our PDB and AlphaFold vectors, and calculated the model performances on the test dataset (i.e., BRCA1: 834, CALM1: 30, and P53: 375 SAVs) using the AUROC metric as employed in the original study. The performance of our model can be seen as blue bars in different panels of Fig. 5. ASCARIS ranked among the top 10% for CALM1 (AUROC: 0.909), the top 19% for BRCA1 (AUROC: 0.919), and the top 26% for P53 variations (AUROC: 0.872) excluding the actual DMS results (red bars), which is highly satisfactory considering the large number of competing methods and the high rate of rank changeability between different datasets. Even though ASCARIS was not particularly designed as a VEP, it could not be consistently beaten by any method in this analysis, the majority of which were highly optimized VEPs with high computational complexity. This, again, indicates that ASCARIS have the power of expressing the functional characteristics of SAVs.

3.2.4. Use-case analyses

To evaluate a few examples where the proposed method could successfully detect the variant effect where the other predictors failed, we selected 3 example SAV data points (2 deleterious/disrupting and 1 neutral), from benchmark 1 and examined the corresponding SAV representation vectors together with relevant information from the literature.

A SAV of the human Arylsulfatase B protein (ARSB_HUMAN) is selected as the first example (gene name: ASB or ARSB, protein UniProt accession: P15848, variation: C405Y, consequence: deleterious). ASB (or ARSB) is an enzyme (*N*-acetylgalactosamine-4-sulfatase) that removes the 4-sulfate group from chondroitin-4-sulfate (C4S) and regulates its degradation [84]. Mutations in the N-acetylgalactosamine-4-sulfatase gene cause reduced enzyme activity, and ASB deficiency is reported to be the cause of Mucopolysaccharidosis type VI (MPS VI; Maroteaux-Lamy syndrome) which is a lysosomal storage disorder [85]. Disease onset and rate of progression is variable for this disorder depending on the mutation, thus assessing mutational characteristics is critical. C405Y mutation changes cysteine in the 405th position to tyrosine and disrupts the disulfide bond between the residues 405 and 447. Theoretically, such inability to form a disulfide bridge that was present in the native form will cause the destabilization of the protein. According to a study by [86], where the authors analyzed mutations from 105 patients with Mucopolysaccharidosis Type VI, they reported that the C405Y mutation causes a slowly progressing disease with a late onset [86]. They suggest that this effect might be observed due to the destabilization introduced by the breaking of the disulfide bond. ASB's active site involves at least 10 residues, and mutations around this region are expected to be highly critical. It is possible to observe from the visualization of the structure of the ASB protein that the mutation is located in a relatively close proximity with the active site (i.e., minimum spatial distance: 11.44 A), which may explain the effect suggested by the authors (Fig. S6a). It is also reported that this mutation is within the annotated region of the superfamily "Alkaline-phosphatase-like, core domain superfamily" [InterPro ID: IPR017850], which is found to be heavily associated with deleterious/disrupting mutations (deleterious and neutral occurrences in our dataset for this domain are 722 and 112, respectively). To sum up, it is possible to state that the information included in our mutation feature vector including; (i) the direct correspondence of this mutation with a disulfide bond, (ii) close proximity to the active site of the protein, and (iii) the domain/family region in which the mutation occurred, may contribute to the correct classification of this mutation as deleterious/disrupting by our method. The remaining use-cases are provided in Supplementary Information S6.

3.3. SAV representation construction tool

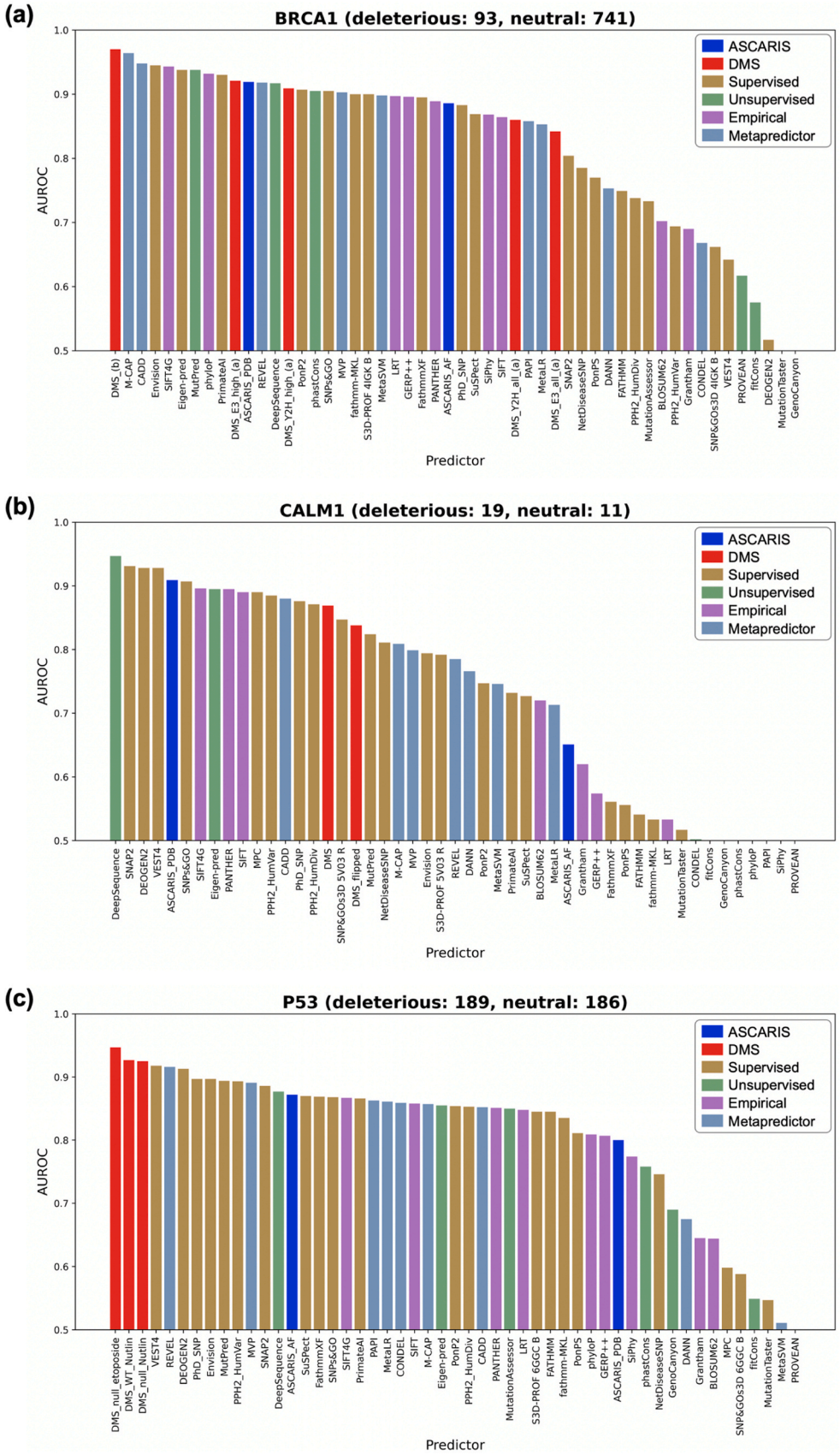
We developed an open access tool for ASCARIS and shared it in two different platforms: 1) as a web-service with graphical user interface at https://huggingface.co/spaces/HUBioDataLab/ASCARIS, and 2) as a command line tool at https://github.com/HUBioDataLab/ASCARIS that can be run locally. The ASCARIS tool generates 74-dimensional representations for the given SAV data points. The input to the tool is a file composed of one or more SAV data points (one in each line), composed of UniProt accession of the protein containing the SAV of interest, one letter notation of the wild type residue of interest, position of the mutated residue, and one letter notation of the mutated residue of interest in different columns, in tab delimited format. Another input option is directly entering the SAV data point (including the same four different types of information this type separated by the "-" character) to the provided window in the user interface of the web-service or inside the Python one-liner that executes ASCARIS (more information can be found in the readme file of the Github repo). The output is again a tab delimited file containing the representation of each input SAV on a different row. Each row contains meta-data related to the SAV (5 dimensions), "all domains" column (1 dimension) which was not used in our VEP analyses, along with the actual 68-dimensions of numerical/categorical features. The detailed explanation of the output file is provided in Table S1. More information can be found at https://github.com/HUBioDataLab/ASCARIS together with all datasets, instructions, and dependencies.

4. Conclusion

In this study, we developed a methodology, ASCARIS, to quantitatively represent single amino acid variations in proteins in terms of their spatial organization with positional sequence features, to reflect their functional characteristics. Our representations also include structure-derived information regarding physicochemical changes caused by the amino acid change, its location on the protein structure and its domain correspondence, constituting a 74-dimensional representation that can be utilized in any statistical data model to represent SAVs in a function-centric way. Possible applications can be predicting the effect of variations, omics-based modeling of cells or patients for precision medicine, designing new proteins, and many more.

As an application of our method, we trained ML models to predict consequences of single amino acid variations on protein functionality and compared their performance against well-known predictors from the literature. During this application, our initial idea was that, when used in the modeling alone, these representations could not compete with alignment-based variant effect predictors as it quantitatively describes SAVs from a limited perspective. However, we expected that it could produce complementary results as its point of view is different from existing methods. Results indicated that our method actually produces complementary results to conventional variant effect predictors. Moreover, it performs quite well in challenging cases where these methods mostly fail. Another interesting observation was that the AlphaFold version of ASCARIS scored a higher performance compared to the PDB version of the method, indicating the benefit of having high quality structure predictions with almost complete coverage on sequences.

One of the advantages of our method is being practical as it utilizes documented annotations instead of trying to detect conservation via sequence alignments. Also, since the annotations are curated, the noise in data is expected to be low. Another advantage of our method is the interpretability of results when it is used in ML-based modeling, given that each dimension on our feature vectors has a known molecular/structural/functional correspondence. One limitation of ASCARIS is that the curated feature annotations of proteins are far from being complete, which means that we are dealing with missing information during modeling. Our method's representation power will further increase with



(caption on next page)

Fig. 5. Performance results (AUROC) of ASCARIS and 46 different VEPs from the literature, on the deep mutational scanning data. AUROC scores indicate the success of methods in distinguishing deleterious (pathogenic) SAVs from the neutral (benign) ones on the; **(a)** BRCA1, **(b)** CALM1, and **(c)** P53 genes. The number of test SAVs for each class are shown in plot titles. ASCARIS scores are shown in blue colored bars (DMS: deep mutational scanning, Supervised: machine learning -ML- models that learns from labeled samples, Unsupervised: ML models that learns from evolutionary conservation and multiple sequence alignments, Empirical: different from ML models, these are VEPs that rely on empirical calculations, Metapredictor: mostly supervised ML models that utilize the results of other VEPs as its input).

the addition of new protein feature annotations in databases such as UniProt. As future work, we plan to construct ensemble-based variant representations by integrating successful structure and alignment-based approaches with our method using multi-modal deep learning. We also plan to incorporate ASCARIS representations in large-scale biomedical knowledge graphs [87] as variant feature vectors for integrative modeling of heterogeneous biomedical data via deep graph learning. We hope that these comprehensive SAV representations will be effectively utilized for data-centric modeling in various areas of biomedicine and biotechnology.

Funding

No funding was received for this work.

CRediT authorship contribution statement

Considering the submission id: CSBJ-D-23–00487R1 and title "ASCARIS: Positional Feature Annotation and Protein Structure-Based Representation of Single Amino Acid Variations", we, the authors of the study, hereby state that both authors have seen and approved the final version of the manuscript being submitted. We, the authors, warrant that the article is our original work, hasn't received prior publication and isn't under consideration for publication elsewhere.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

ASCARIS is available as an open access programmatic tool at https://github.com/HUBioDataLab/ASCARIS. Users can access all input files and the constructed variant representations through this repository. The instructions for constructing representation vectors for completely new variation datasets are also provided. ASCARIS is also available as an open access web-service as a web-service with graphical user interface at https://huggingface.co/spaces/HUBioDataLab/ASCARIS which can easily be used for generating representations for a given SAV dataset without any programming.

Acknowledgement

The authors thank Dr. Nurcan Tuncbag (Koc University, Turkey) for valuable discussions and insight regarding protein structure analysis in the context of single amino acid variations.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.09.017.

References

[1] Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest 2008;118:1590–605.
[2] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 2009;106:9362–7.
[3] Fariselli P, Martelli PL, Savojardo C, Casadio R. INPS: predicting the impact of non-synonymous variations on protein stability from sequence. Bioinformatics 2015;31:2816–21.
[4] Datta A, Mazumder MH, Chowdhury AS, Hasan MA. Functional and structural consequences of damaging single nucleotide polymorphisms in human prostate cancer predisposition gene RNASEL. Biomed Res Int 2015;2015:271458.
[5] Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJ, Shishkin AA, Hatan M, Carrasco-Alfonso MJ, Mayer D, Luckey CJ, Patsopoulos NA, De Jager PL, Kuchroo VK, Epstein CB, Daly MJ, Hafler DA, Bernstein BE. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 2015;518:337–43.
[6] Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 1999;22:239–47.
[7] Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, Gerstein M. Role of non-coding sequence variants in cancer. Nat Rev Genet 2016;17:93–108.
[8] Presnyak V, Alhusaini N, Chen Y-H, Martin S, Morris N, Kline N, Olson S, Weinberg D, Baker KE, Graveley BR, Coller J. Codon optimality is a major determinant of mRNA stability. Cell 2015;160:1111–24.
[9] Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet 2011;12:683–91.
[10] Supek F, Minana B, Valcarcel J, Gabaldon T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. Cell 2014;156:1324–35.
[11] Zwart MP, Schenk MF, Hwang S, Koopmanschap B, de Lange N, van de Pol L, Nga TTT, Szendro IG, Krug J, de Visser J. Unraveling the causes of adaptive benefits of synonymous mutations in TEM-1 beta-lactamase. Hered (Edinb) 2018;121:406–21.
[12] C. Dincer, T. Kaya, O. Keskin, A. Gursoy, N. Tuncbag, 3D spatial organization and network-guided comparison of mutation profiles in Glioblastoma reveals similarities across patients.
[13] Unsal S, Atas H, Albayrak M, Turhan K, Acar AC, Doğan T. Learning functional properties of proteins with language models. Nat Mach Intell 2022;4:227–45.
[14] Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 2009;30:1237–44.
[15] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.
[16] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet, Chapter 2013;7. Unit7.20.
[17] Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 2007;35:3823–35.
[18] Bromberg Y, Yachdav G, Rost B. SNAP predicts effect of mutations on protein function. Bioinformatics 2008;24:2397–8.
[19] Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res 2009;69:6660–7.
[20] Chennen K, Weber T, Lornage X, Kress A, Böhm J, Thompson J, Laporte J, Poch O. MISTIC: A prediction tool to reveal disease-relevant deleterious missense variants. PLoS One 2020;15:e0236962.
[21] Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 2015;31:2745–7.
[22] Clifford RJ, Edmonson MN, Nguyen C, Buetow KH. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. Bioinformatics 2004;20:1006–14.
[23] Kaminker JS, Zhang Y, Watanabe C, Zhang Z. CanPredict: a computational tool for predicting cancer-associated missense mutations. Nucleic Acids Res 2007;35:W595–8.
[24] Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812–4.
[25] Pandurangan AP, Blundell TL. Prediction of impacts of mutations on protein structure and interactions: SDM, a statistical approach, and mCSM, using machine learning. Protein Sci 2020;29:247–57.
[26] Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 2014;30:335–42.
[27] Quan L, Lv Q, Zhang Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics 2016;32:2936–46.
[28] Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 2019;47:D886–94.
[29] Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 2014;11:361–2.

[30] Tavtigian SV, Byrnes GB, Goldgar DE, Thomas A. Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications. Hum Mutat 2008;29:1342–54.
[31] Topham CM, Srinivasan N, Blundell TL. Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Eng 1997;10:7–21.
[32] Worth CL, Preissner R, Blundell TL. SDM–a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res 2011;39: W215–22.
[33] Yang Y, Chen B, Tan G, Vihinen M, Shen B. Structure-based prediction of the effects of a missense variant on protein stability. Amino Acids 2013;44:847–55.
[34] Yue P, Moult J. Identification and Analysis of Deleterious Human SNPs. J Mol Biol 2006;356:1263–74.
[35] Konig E, Rainer J, Domingues FS. Computational assessment of feature combinations for pathogenic variant prediction. Mol Genet Genom Med 2016;4: 431–46.
[36] Tan KP, Kanitkar TR, Kwoh CK, Madhusudhan MS. Packpred: predicting the functional effect of missense mutations. Front Mol Biosci 2021;8:646288.
[37] Pei J, Kinch LN, Otwinowski Z, Grishin NV. Mutation severity spectrum of rare alleles in the human genome is predictive of disease type. PLoS Comput Biol 2020; 16:e1007775.
[38] Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated? J Mol Biol 2019;431:2197–212.
[39] Marquet C, Heinzinger M, Olenyi T, Dallago C, Erckert K, Bernhofer M, Nechaev D, Rost B. Embeddings from protein language models predict conservation and variant effects. Hum Genet 2022;141:1629–47.
[40] Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function. bioRxiv 2021. 2021.2007.2009.450648.
[41] Brandes N, Goldman G, Wang CH, Ye CJ, Ntranos V. Genome-wide prediction of disease variants with a deep protein language model. bioRxiv 2022;2022. 2008.2025.505311.
[42] UniProt C. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–15.
[43] Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics 2020;36:422–9.
[44] Rifaioglu AS, Dogan T, Sarac OS, Ersahin T, Saidi R, Atalay MV, Martin MJ, Cetin-Atalay R. Large-scale automated function prediction of protein sequences and an experimental case study validation on PTEN transcript variants. Proteins 2018;86: 135–51.
[45] Dogan T. HPO2GO: prediction of human phenotype ontology term associations for proteins using cross ontology annotation co-occurrences. PeerJ 2018;6:e5298.
[46] Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. BMC Genom 2013;14(Suppl 3):S6.
[47] Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipatla Z, Malheiro A, McDaniel K, Ovetsky M, Riley G, Zhou G, Holmes JB, Kattman BL, Maglott DR. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 2018;46:D1062–7.
[48] Kawabata T, Ota M, Nishikawa K. The Protein Mutant Database. Nucleic Acids Res 1999;27:355–7.
[49] Mudunuri U, Che A, Yi M, Stephens RM. bioDBnet: the biological database network. Bioinformatics 2009;25:555–6.
[50] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–42.
[51] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 2018; 46:W296–303.
[52] Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, Kim SJ, Khuri N, Spill YG, Weinkam P, Hammel M, Tainer JA, Nilges M, Sali A. ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 2014;42:D336–46.
[53] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.
[54] Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang H-Y, El-Gebali S, Fraser MI, Gough J, Haft DR, Huang H, Letunic I, Lopez R, Luciani A, Madeira F, Marchler-Bauer A, Mi H, Natale DA, Necci M, Nuka G, Orengo C, Pandurangan AP, Paysan-Lafosse T, Pesseat S, Potter SC, Qureshi MA, Rawlings ND, Redaschi N, Richardson LJ, Rivoire C, Salazar GA, Sangrador-Vegas A, Sigrist CJA, Sillitoe I, Sutton GG, Thanki N, Thomas PD, Tosatto SCE, Yong S-Y, Finn RD. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res 2019;47:D351–60.
[55] Aboderin AA. An empirical hydrophobicity scale for α-amino-acids and some of its applications. Int J Biochem 1971;2:537–44.
[56] Goldsack DE, Chalifoux RC. Contribution of the free energy of mixing of hydrophobic side chains to the stability of the tertiary structure of proteins. J Theor Biol 1973;39:645–51.
[57] Grantham R. Amino acid difference formula to help explain protein evolution. Science 1974;185:862–4.
[58] Capriotti E, Ozturk K, Carter H. Integrating molecular networks with genetic variant interpretation for precision medicine. Wiley Interdiscip Rev Syst Biol Med 2019;11:e1443.
[59] Engin HB, Hofree M, Carter H. Identifying mutation specific cancer pathways using a structurally resolved protein interaction network. Pac Symp Biocomput 2015: 84–95.
[60] Guo HH, Choe J, Loeb LA. Protein tolerance to random amino acid change. Proc Natl Acad Sci USA 2004;101:9205–10.
[61] Nishi H, Tyagi M, Teng S, Shoemaker BA, Hashimoto K, Alexov E, Wuchty S, Panchenko AR. Cancer missense mutations alter binding properties of proteins and their interaction networks. PLoS One 2013;8:e66273.
[62] Mitternacht S. FreeSASA: An open source C library for solvent accessible surface area calculations. F1000Res 2016;5:189.
[63] Momen-Roknabadi A, Sadeghi M, Pezeshk H, Marashi S-A. Impact of residue accessible surface area on the prediction of protein secondary structures. BMC Bioinforma 2008;9:357.
[64] Meyer MJ, Beltrán JF, Liang S, Fragoza R, Rumack A, Liang J, Wei X, Yu H. Interactome INSIDER: a structural interactome browser for genomic studies. Nat Methods 2018;15:107–14.
[65] Breiman L. Mach Learn 2001;45:261–77.
[66] Dasgupta A, Sun YV, König IR, Bailey-Wilson JE, Malley JD. Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. Genet Epidemiol 2011;35(Suppl 1): S5–11.
[67] Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J Comput Syst Sci 1997;55:119–39.
[68] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). Ann Stat 2000;28.
[69] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, 2000.
[70] Hastie T, Friedman J, Tibshirani R. The Elements of Statistical Learning. Springer Series in Statistics,; 2001.
[71] Doğan T, MacDougall A, Saidi R, Poggioli D, Bateman A, O'Donovan C, Martin MJ. UniProt-DAAC: domain architecture alignment and classification, a new method for automatic functional annotation in UniProtKB. Bioinformatics 2016;32: 2264–71.
[72] McGarvey PB, Nightingale A, Luo J, Huang H, Martin MJ, Wu C, UniProt C. UniProt genomic mapping for deciphering functional effects of missense variants. Hum Mutat 2019;40:694–705.
[73] Truebestein L, Leonard TA. Coiled-coils: The long and short of it. BioEssays 2016; 38:903–16.
[74] Holbrook K, Subramanian C, Chotewutmontri P, Reddick LE, Wright S, Zhang H, Moncrief L, Bruce BD. Functional Analysis of Semi-conserved Transit Peptide Motifs and Mechanistic Implications in Precursor Targeting and Recognition. Mol Plant 2016;9:1286–301.
[75] Cox MAA, Cox TF. Multidimensional Scaling. Handb Data Vis 2008:315–47.
[76] van der M, Geoffrey Hinton L. Visualizing Data using t-SNE. J Mach Learn Res 2008;9:2579–605.
[77] Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res 2005; 15:978–86.
[78] Doğan T, Akhan Güzelcan E, Baumann M, Koyas A, Atas H, Baxendale IR, Martin M, Cetin-Atalay R. Protein domain-based prediction of drug/ compound–target interactions and experimental validation on LIM kinases. PLOS Comput Biol 2021;17:e1009171.
[79] Consortium GP, Auton A, Brooks L, Durbin R, Garrison E, Kang H. A global reference for human genetic variation. Nature 2015;526:68–74.
[80] Heberle H, Meirelles GV, da Silva FR, Telles GP, Minghim R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinforma 2015; 16:169.
[81] Grimm DG, Azencott C-A, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, Duncan LE, Borgwardt KM. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat 2015;36:513–23.
[82] Livesey BJ, Marsh JA. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Mol Syst Biol 2020;16:e9380.
[83] Karczewski K, Francioli L. The genome aggregation database (gnomAD). MacArthur Lab 2017:1–10.
[84] Sharma G, Burke J, Bhattacharyya S, Sharma N, Katyal S, Park RL, Tobacman J. Reduced Arylsulfatase B activity in leukocytes from cystic fibrosis patients. Pedia Pulmonol 2013;48:236–44.
[85] Bhattacharyya S, Tobacman JK. Arylsulfatase B regulates colonic epithelial cell migration by effects on MMP9 expression and RhoA activation. Clin Exp Metastas– 2009;26:535–45.
[86] Karageorgos L, Brooks DA, Pollard A, Melville EL, Hein LK, Clements PR, Ketteridge D, Swiedler SJ, Beck M, Giugliani R, Harmatz P, Wraith JE, Guffon N, Leao Teles E, Sa Miranda MC, Hopwood JJ. Mutational analysis of 105 mucopolysaccharidosis type VI patients. Hum Mutat 2007;28:897–903.
[87] Doğan T, Atas H, Joshi V, Atakan A, Rifaioglu AS, Nalbat E, Nightingale A, Saidi R, Volynkin V, Zellner H, Cetin-Atalay R, Martin M, Atalay V. CROssBAR: comprehensive resource of biomedical relations with knowledge graph representations. Nucleic Acids Res 2021;49:e96.