

**DEEP CONVOLUTIONAL NEURAL NETWORKS TO PREDICT VARIANT
PATHOGENICITY USING THREE-DIMENSIONAL (3D) PROTEIN STRUCTURES**

PRIORITY APPLICATIONS

[0001] This application claims priority to U.S. Nonprovisional Patent Application No. 17/468,411, titled “Deep Convolutional Neural Networks to Predict Variant Pathogenicity using Three-Dimensional (3D) Protein Structures,” filed September 7, 2021 (Attorney Docket No. ILLM 1037-3/IP-2051A-US) which is a continuation of U.S. Nonprovisional Patent Application No. 17/232,056 titled “Deep Convolutional Neural Networks to Predict Variant Pathogenicity using Three-Dimensional (3D) Protein Structures” filed April 15, 2021 (Attorney Docket No. ILLM 1037-2/IP-2051-US).

[0002] This application also claims priority to U.S. Nonprovisional Patent Application No. 17/703,935 titled “Multi-channel Protein Voxelization To Predict Variant Pathogenicity Using Deep Convolutional Neural Networks,” filed March 24, 2022 (Attorney Docket No. ILLM 1047-2/IP-2142-US), which in turn claims priority to or the benefit of U.S. Provisional Patent Application No. 63/175,495, titled “Multi-channel Protein Voxelization To Predict Variant Pathogenicity Using Deep Convolutional Neural Networks,” filed on April 15, 2021, (Atty. Docket No. ILLM 1047-1/IP-2142-PRV).

[0003] This application also claims priority to U.S. Nonprovisional Patent Application No. 17/703,958 titled “Efficient Voxelization For Deep Learning,” filed March 24, 2022 (Attorney Docket No. ILLM 1048-2/IP-2143-US), which in turn claims priority to or the benefit of U.S. Provisional Patent Application No. 63/175,767, titled “Efficient Voxelization For Deep Learning,” filed on April 16, 2021, (Atty. Docket No. ILLM 1048-1/IP-2143-PRV).

[0004] The priority applications are hereby incorporated by reference for all purposes.

RELATED APPLICATION

[0005] This application is related to PCT Patent Application titled “Artificial Intelligence-based Analysis of Protein Three-Dimensional (3D) Structures,” (Attorney Docket No. ILLM 1037-4/IP-2051-PCT) filed contemporaneously. The related application is hereby incorporated by reference for all purposes.

FIELD OF THE TECHNOLOGY DISCLOSED

[0006] The technology disclosed relates to artificial intelligence type computers and digital data processing systems and corresponding data processing methods and products for emulation of intelligence (*i.e.*, knowledge based systems, reasoning systems, and knowledge acquisition systems); and including systems for reasoning with uncertainty (*e.g.*, fuzzy logic systems),

adaptive systems, machine learning systems, and artificial neural networks. In particular, the technology disclosed relates to using deep convolutional neural networks to analyze multi-channel voxelized data.

INCORPORATIONS

[0007] The following are incorporated by reference for all purposes as if fully set forth herein:

[0008] Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* 50, 1161–1170 (2018);

[0009] Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* 176, 535–548 (2019);

[0010] US Patent Application No. 62/573,144, titled “TRAINING A DEEP PATHOGENICITY CLASSIFIER USING LARGE-SCALE BENIGN TRAINING DATA,” filed October 16, 2017 (Attorney Docket No. ILLM 1000-1/IP-1611-PRV);

[0011] US Patent Application No. 62/573,149, titled “PATHOGENICITY CLASSIFIER BASED ON DEEP CONVOLUTIONAL NEURAL NETWORKS (CNNs),” filed October 16, 2017 (Attorney Docket No. ILLM 1000-2/IP-1612-PRV);

[0012] US Patent Application No. 62/573,153, titled “DEEP SEMI-SUPERVISED LEARNING THAT GENERATES LARGE-SCALE PATHOGENIC TRAINING DATA,” filed October 16, 2017 (Attorney Docket No. ILLM 1000-3/IP-1613-PRV);

[0013] US Patent Application No. 62/582,898, titled “PATHOGENICITY CLASSIFICATION OF GENOMIC DATA USING DEEP CONVOLUTIONAL NEURAL NETWORKS (CNNs),” filed November 7, 2017 (Attorney Docket No. ILLM 1000-4/IP-1618-PRV);

[0014] US Patent Application No. 16/160,903, titled “DEEP LEARNING-BASED TECHNIQUES FOR TRAINING DEEP CONVOLUTIONAL NEURAL NETWORKS,” filed on October 15, 2018 (Attorney Docket No. ILLM 1000-5/IP-1611-US);

[0015] US Patent Application No. 16/160,986, titled “DEEP CONVOLUTIONAL NEURAL NETWORKS FOR VARIANT CLASSIFICATION,” filed on October 15, 2018 (Attorney Docket No. ILLM 1000-6/IP-1612-US);

[0016] US Patent Application No. 16/160,968, titled “SEMI-SUPERVISED LEARNING FOR TRAINING AN ENSEMBLE OF DEEP CONVOLUTIONAL NEURAL NETWORKS,” filed on October 15, 2018 (Attorney Docket No. ILLM 1000-7/IP-1613-US); and

[0017] US Patent Application No. 16/407,149, titled “DEEP LEARNING-BASED TECHNIQUES FOR PRE-TRAINING DEEP CONVOLUTIONAL NEURAL NETWORKS,” filed May 8, 2019 (Attorney Docket No. ILLM 1010-1/IP-1734-US).

BACKGROUND

[0018] The subject matter discussed in this section should not be assumed to be prior art merely as a result of its mention in this section. Similarly, a problem mentioned in this section or associated with the subject matter provided as background should not be assumed to have been previously recognized in the prior art. The subject matter in this section merely represents different approaches, which in and of themselves can also correspond to implementations of the claimed technology.

[0019] Genomics, in the broad sense, also referred to as functional genomics, aims to characterize the function of every genomic element of an organism by using genome-scale assays such as genome sequencing, transcriptome profiling and proteomics. Genomics arose as a data-driven science — it operates by discovering novel properties from explorations of genome-scale data rather than by testing preconceived models and hypotheses. Applications of genomics include finding associations between genotype and phenotype, discovering biomarkers for patient stratification, predicting the function of genes, and charting biochemically active genomic regions such as transcriptional enhancers.

[0020] Genomics data are too large and too complex to be mined solely by visual investigation of pairwise correlations. Instead, analytical tools are required to support the discovery of unanticipated relationships, to derive novel hypotheses and models and to make predictions. Unlike some algorithms, in which assumptions and domain expertise are hard coded, machine learning algorithms are designed to automatically detect patterns in data. Hence, machine learning algorithms are suited to data-driven sciences and, in particular, to genomics. However, the performance of machine learning algorithms can strongly depend on how the data are represented, that is, on how each variable (also called a feature) is computed. For instance, to classify a tumor as malign or benign from a fluorescent microscopy image, a preprocessing algorithm could detect cells, identify the cell type, and generate a list of cell counts for each cell type.

[0021] A machine learning model can take the estimated cell counts, which are examples of handcrafted features, as input features to classify the tumor. A central issue is that classification performance depends heavily on the quality and the relevance of these features. For example, relevant visual features such as cell morphology, distances between cells or localization within

an organ are not captured in cell counts, and this incomplete representation of the data may reduce classification accuracy.

[0022] Deep learning, a subdiscipline of machine learning, addresses this issue by embedding the computation of features into the machine learning model itself to yield end-to-end models. This outcome has been realized through the development of deep neural networks, machine learning models that comprise successive elementary operations, which compute increasingly more complex features by taking the results of preceding operations as input. Deep neural networks are able to improve prediction accuracy by discovering relevant features of high complexity, such as the cell morphology and spatial organization of cells in the above example. The construction and training of deep neural networks have been enabled by the explosion of data, algorithmic advances, and substantial increases in computational capacity, particularly through the use of graphical processing units (GPUs).

[0023] The goal of supervised learning is to obtain a model that takes features as input and returns a prediction for a so-called target variable. An example of a supervised learning problem is one that predicts whether an intron is spliced out or not (the target) given features on the RNA such as the presence or absence of the canonical splice site sequence, the location of the splicing branchpoint or intron length. Training a machine learning model refers to learning its parameters, which commonly involves minimizing a loss function on training data with the aim of making accurate predictions on unseen data.

[0024] For many supervised learning problems in computational biology, the input data can be represented as a table with multiple columns, or features, each of which contains numerical or categorical data that are potentially useful for making predictions. Some input data are naturally represented as features in a table (such as temperature or time), whereas other input data need to be first transformed (such as deoxyribonucleic acid (DNA) sequence into k -mer counts) using a process called feature extraction to fit a tabular representation. For the intron-splicing prediction problem, the presence or absence of the canonical splice site sequence, the location of the splicing branchpoint and the intron length can be preprocessed features collected in a tabular format. Tabular data are standard for a wide range of supervised machine learning models, ranging from simple linear models, such as logistic regression, to more flexible nonlinear models, such as neural networks and many others.

[0025] Logistic regression is a binary classifier, that is, a supervised learning model that predicts a binary target variable. Specifically, logistic regression predicts the probability of the positive class by computing a weighted sum of the input features mapped to the $[0,1]$ interval using the sigmoid function, a type of activation function. The parameters of logistic regression, or other linear classifiers that use different activation functions, are the weights in the weighted

sum. Linear classifiers fail when the classes, for instance, that of an intron spliced out or not, cannot be well discriminated with a weighted sum of input features. To improve predictive performance, new input features can be manually added by transforming or combining existing features in new ways, for example, by taking powers or pairwise products.

[0026] Neural networks use hidden layers to learn these nonlinear feature transformations automatically. Each hidden layer can be thought of as multiple linear models with their output transformed by a nonlinear activation function, such as the sigmoid function or the more popular rectified-linear unit (ReLU). Together, these layers compose the input features into relevant complex patterns, which facilitates the task of distinguishing two classes.

[0027] Deep neural networks use many hidden layers, and a layer is said to be fully-connected when each neuron receives inputs from all neurons of the preceding layer. Neural networks are commonly trained using stochastic gradient descent, an algorithm suited to training models on very large data sets. Implementation of neural networks using modern deep learning frameworks enables rapid prototyping with different architectures and data sets. Fully-connected neural networks can be used for a number of genomics applications, which include predicting the percentage of exons spliced in for a given sequence from sequence features such as the presence of binding motifs of splice factors or sequence conservation; prioritizing potential disease-causing genetic variants; and predicting *cis*-regulatory elements in a given genomic region using features such as chromatin marks, gene expression and evolutionary conservation.

[0028] Local dependencies in spatial and longitudinal data must be considered for effective predictions. For example, shuffling a DNA sequence or the pixels of an image severely disrupts informative patterns. These local dependencies set spatial or longitudinal data apart from tabular data, for which the ordering of the features is arbitrary. Consider the problem of classifying genomic regions as bound versus unbound by a particular transcription factor, in which bound regions are defined as high-confidence binding events in chromatin immunoprecipitation following by sequencing (ChIP-seq) data. Transcription factors bind to DNA by recognizing sequence motifs. A fully-connected layer based on sequence-derived features, such as the number of k -mer instances or the position weight matrix (PWM) matches in the sequence, can be used for this task. As k -mer or PWM instance frequencies are robust to shifting motifs within the sequence, such models could generalize well to sequences with the same motifs located at different positions. However, they would fail to recognize patterns in which transcription factor binding depends on a combination of multiple motifs with well-defined spacing. Furthermore, the number of possible k -mers increases exponentially with k -mer length, which poses both storage and overfitting challenges.

[0029] A convolutional layer is a special form of fully-connected layer in which the same fully-connected layer is applied locally, for example, in a 6 bp window, to all sequence positions. This approach can also be viewed as scanning the sequence using multiple PWMs, for example, for transcription factors GATA1 and TAL1. By using the same model parameters across positions, the total number of parameters is drastically reduced, and the network is able to detect a motif at positions not seen during training. Each convolutional layer scans the sequence with several filters by producing a scalar value at every position, which quantifies the match between the filter and the sequence. As in fully-connected neural networks, a nonlinear activation function (commonly ReLU) is applied at each layer. Next, a pooling operation is applied, which aggregates the activations in contiguous bins across the positional axis, commonly taking the maximal or average activation for each channel. Pooling reduces the effective sequence length and coarsens the signal. The subsequent convolutional layer composes the output of the previous layer and is able to detect whether a GATA1 motif and TAL1 motif were present at some distance range. Finally, the output of the convolutional layers can be used as input to a fully-connected neural network to perform the final prediction task. Hence, different types of neural network layers (*e.g.*, fully-connected layers and convolutional layers) can be combined within a single neural network.

[0030] Convolutional neural networks (CNNs) can predict various molecular phenotypes on the basis of DNA sequence alone. Applications include classifying transcription factor binding sites and predicting molecular phenotypes such as chromatin features, DNA contact maps, DNA methylation, gene expression, translation efficiency, RBP binding, and microRNA (miRNA) targets. In addition to predicting molecular phenotypes from the sequence, convolutional neural networks can be applied to more technical tasks traditionally addressed by handcrafted bioinformatics pipelines. For example, convolutional neural networks can predict the specificity of guide RNA, denoise ChIP-seq, enhance Hi-C data resolution, predict the laboratory of origin from DNA sequences and call genetic variants. Convolutional neural networks have also been employed to model long-range dependencies in the genome. Although interacting regulatory elements may be distantly located on the unfolded linear DNA sequence, these elements are often proximal in the actual 3D chromatin conformation. Hence, modelling molecular phenotypes from the linear DNA sequence, albeit a crude approximation of the chromatin, can be improved by allowing for long-range dependencies and allowing the model to implicitly learn aspects of the 3D organization, such as promoter-enhancer looping. This is achieved by using dilated convolutions, which have a receptive field of up to 32 kb. Dilated convolutions also allow splice sites to be predicted from sequence using a receptive field of 10 kb, thereby enabling the integration of genetic sequence across distances as long as typical human introns (*See*

Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. Cell 176, 535–548 (2019)).

[0031] Different types of neural network can be characterized by their parameter-sharing schemes. For example, fully-connected layers have no parameter sharing, whereas convolutional layers impose translational invariance by applying the same filters at every position of their input. Recurrent neural networks (RNNs) are an alternative to convolutional neural networks for processing sequential data, such as DNA sequences or time series, that implement a different parameter-sharing scheme. Recurrent neural networks apply the same operation to each sequence element. The operation takes as input the memory of the previous sequence element and the new input. It updates the memory and optionally emits an output, which is either passed on to subsequent layers or is directly used as model predictions. By applying the same model at each sequence element, recurrent neural networks are invariant to the position index in the processed sequence. For example, a recurrent neural network can detect an open reading frame in a DNA sequence regardless of the position in the sequence. This task requires the recognition of a certain series of inputs, such as the start codon followed by an in-frame stop codon.

[0032] The main advantage of recurrent neural networks over convolutional neural networks is that they are, in theory, able to carry over information through infinitely long sequences via memory. Furthermore, recurrent neural networks can naturally process sequences of widely varying length, such as mRNA sequences. However, convolutional neural networks combined with various tricks (such as dilated convolutions) can reach comparable or even better performances than recurrent neural networks on sequence-modelling tasks, such as audio synthesis and machine translation. Recurrent neural networks can aggregate the outputs of convolutional neural networks for predicting single-cell DNA methylation states, RBP binding, transcription factor binding, and DNA accessibility. Moreover, because recurrent neural networks apply a sequential operation, they cannot be easily parallelized and are hence much slower to compute than convolutional neural networks.

[0033] Each human has a unique genetic code, though a large portion of the human genetic code is common for all humans. In some cases, a human genetic code may include an outlier, called a genetic variant, that may be common among individuals of a relatively small group of the human population. For example, a particular human protein may comprise a specific sequence of amino acids, whereas a variant of that protein may differ by one amino acid in the otherwise same specific sequence.

[0034] Genetic variants may be pathogenetic, leading to diseases. Though most of such genetic variants have been depleted from genomes by natural selection, an ability to identify which genetic variants are likely to be pathogenetic can help researchers focus on these genetic

variants to gain an understanding of the corresponding diseases and their diagnostics, treatments, or cures. The clinical interpretation of millions of human genetic variants remains unclear. Some of the most frequent pathogenic variants are single nucleotide missense mutations that change the amino acid of a protein. However, not all missense mutations are pathogenic.

[0035] Models that can predict molecular phenotypes directly from biological sequences can be used as in silico perturbation tools to probe the associations between genetic variation and phenotypic variation and have emerged as new methods for quantitative trait loci identification and variant prioritization. These approaches are of major importance given that the majority of variants identified by genome-wide association studies of complex phenotypes are non-coding, which makes it challenging to estimate their effects and contribution to phenotypes. Moreover, linkage disequilibrium results in blocks of variants being co-inherited, which creates difficulties in pinpointing individual causal variants. Thus, sequence-based deep learning models that can be used as interrogation tools for assessing the impact of such variants offer a promising approach to find potential drivers of complex phenotypes. One example includes predicting the effect of non-coding single-nucleotide variants and short insertions or deletions (indels) indirectly from the difference between two variants in terms of transcription factor binding, chromatin accessibility or gene expression predictions. Another example includes predicting novel splice site creation from sequence or quantitative effects of genetic variants on splicing.

[0036] End-to-end deep learning approaches for variant effect predictions are applied to predict the pathogenicity of missense variants from protein sequence and sequence conservation data (*See* Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* 50, 1161–1170 (2018), referred to herein as “PrimateAI”). PrimateAI uses deep neural networks trained on variants of known pathogenicity with data augmentation using cross-species information. In particular, PrimateAI uses sequences of wild-type and mutant proteins to compare the difference and decide the pathogenicity of mutations using the trained deep neural networks. Such an approach which utilizes the protein sequences for pathogenicity prediction is promising because it can avoid the circularity problem and overfitting to previous knowledge. However, compared to the adequate number of data to train the deep neural networks effectively, the number of clinical data available in ClinVar is relatively small. To overcome this data scarcity, PrimateAI uses common human variants and variants from primates as benign data while simulated variants based on trinucleotide context were used as unlabeled data.

[0037] PrimateAI outperforms prior methods when trained directly upon sequence alignments. PrimateAI learns important protein domains, conserved amino acid positions, and sequence dependencies directly from the training data consisting of about 120,000 human samples. PrimateAI substantially exceeds the performance of other variant pathogenicity

prediction tools in differentiating benign and pathogenic de-novo mutations in candidate developmental disorder genes, and in reproducing prior knowledge in ClinVar. These results suggest that PrimateAI is an important step forward for variant classification tools that may lessen the reliance of clinical reporting on prior knowledge.

[0038] Central to protein biology is the understanding of how structural elements give rise to observed function. The surfeit of protein structural data enables development of computational methods to systematically derive rules governing structural-functional relationships. However, performance of these methods depends critically on the choice of protein structural representation.

[0039] Protein sites are microenvironments within a protein structure, distinguished by their structural or functional role. A site can be defined by a three-dimensional (3D) location and a local neighborhood around this location in which the structure or function exists. Central to rational protein engineering is the understanding of how the structural arrangement of amino acids creates functional characteristics within protein sites. Determination of the structural and functional roles of individual amino acids within a protein provides information to help engineer and alter protein functions. Identifying functionally or structurally important amino acids allows focused engineering efforts such as site-directed mutagenesis for altering targeted protein functional properties. Alternatively, this knowledge can help avoid engineering designs that would abolish a desired function.

[0040] Since it has been established that structure is far more conserved than sequence, the increase in protein structural data provides an opportunity to systematically study the underlying pattern governing the structural-functional relationships using data-driven approaches. A fundamental aspect of any computational protein analysis is how protein structural information is represented. The performance of machine learning methods often depends more on the choice of data representation than the machine learning algorithm employed. Good representations efficiently capture the most critical information while poor representations create a noisy distribution with no underlying patterns.

[0041] The surfeit of protein structures and the recent success of deep learning algorithms provide an opportunity to develop tools for automatically extracting task specific representations of protein structures. Therefore, an opportunity arises to predict variant pathogenicity using multi-channel voxelized representations of 3D protein structures as input to deep neural networks.

BRIEF DESCRIPTION OF THE DRAWINGS

[0042] In the drawings, like reference characters generally refer to like parts throughout the different views. Also, the drawings are not necessarily to scale, with an emphasis instead generally being placed upon illustrating the principles of the technology disclosed. In the following description, various implementations of the technology disclosed are described with reference to the following drawings, in which.

[0043] Figure 1 is a flow diagram that illustrates a process of a system for determining pathogenicity of variants, according to various implementations of the technology disclosed.

[0044] Figure 2 schematically illustrates an example reference amino acid sequence of a protein and an alternative amino acid sequence of the protein, in accordance with one implementation of the technology disclosed.

[0045] Figure 3 illustrates amino acid-wise classification of atoms of amino acids in the reference amino acid sequence of Figure 2, in accordance with one implementation of the technology disclosed.

[0046] Figure 4 illustrates amino acid-wise attribution of 3D atomic coordinates of the alpha-carbon atoms classified in Figure 3 on an amino acid-basis, in accordance with one implementation of the technology disclosed.

[0047] Figure 5 schematically illustrates a process of determining voxel-wise distance values, in accordance with one implementation of the technology disclosed.

[0048] Figure 6 shows an example of twenty-one amino acid-wise distance channels, in accordance with one implementation of the technology disclosed.

[0049] Figure 7 is a schematic diagram of a distance channel tensor, in accordance with one implementation of the technology disclosed.

[0050] Figure 8 shows one-hot encodings of the reference amino acid and the alternative amino acid from Figure 2, in accordance with one implementation of the technology disclosed.

[0051] Figure 9 is a schematic diagram of a voxelized one-hot encoded reference amino acid and a voxelized one-hot encoded variant/alternative amino acid, in accordance with one implementation of the technology disclosed.

[0052] Figure 10 schematically illustrates a concatenation process that voxel-wise concatenates the distance channel tensor of Figure 7 and a reference allele tensor, in accordance with one implementation of the technology disclosed.

[0053] Figure 11 schematically illustrates a concatenation process that voxel-wise concatenates the distance channel tensor of Figure 7, the reference allele tensor of Figure 10, and an alternative allele tensor, in accordance with one implementation of the technology disclosed.

[0054] Figure 12 is a flow diagram that illustrates a process of a system for determining and assigning pan-amino acid conservation frequencies of nearest atoms to voxels (voxelizing), in accordance with one implementation of the technology disclosed.

[0055] Figure 13 illustrates voxels-to-nearest amino acids, in accordance with one implementation of the technology disclosed.

[0056] Figure 14 shows an example multi-sequence alignment of the reference amino acid sequence across a ninety-nine species, in accordance with one implementation of the technology disclosed.

[0057] Figure 15 shows an example of determining a pan-amino acid conservation frequencies sequence for a particular voxel, in accordance with one implementation of the technology disclosed.

[0058] Figure 16 shows respective pan-amino acid conservation frequencies determined for respective voxels using the position frequency logic described in Figure 15, in accordance with one implementation of the technology disclosed.

[0059] Figure 17 illustrates voxelized per-voxel evolutionary profiles, in accordance with one implementation of the technology disclosed.

[0060] Figure 18 depicts example of an evolutionary profiles tensor, in accordance with one implementation of the technology disclosed.

[0061] Figure 19 is a flow diagram that illustrates a process of a system for determining and assigning per-amino acid conservation frequencies of nearest atoms to voxels (voxelizing), in accordance with one implementation of the technology disclosed.

[0062] Figure 20 shows various examples of voxelized annotation channels that are concatenated with the distance channel tensor, in accordance with one implementation of the technology disclosed.

[0063] Figure 21 illustrates different combinations and permutations of input channels that can be provided as inputs to a pathogenicity classifier for pathogenicity determination of a target variant, in accordance with one implementation of the technology disclosed.

[0064] Figure 22 shows different methods of calculating the disclosed distance channels, in accordance with various implementations of the technology disclosed.

[0065] Figure 23 shows different examples of the evolutionary channels, in accordance with various implementations of the technology disclosed.

[0066] Figure 24 shows different examples of the annotations channels, in accordance with various implementations of the technology disclosed.

[0067] Figure 25 shows different examples of the structure confidence channels, in accordance with various implementations of the technology disclosed.

[0068] Figure 26 shows an example processing architecture of the pathogenicity classifier, in accordance with one implementation of the technology disclosed.

[0069] Figure 27 shows an example processing architecture of the pathogenicity classifier, in accordance with one implementation of the technology disclosed.

[0070] Figures 28, 29, 30, and 31 use PrimateAI as a benchmark model to demonstrate the disclosed PrimateAI 3D's classification superiority over PrimateAI.

[0071] Figures 32A and 32B show the disclosed efficient voxelization process, in accordance with various implementations of the technology disclosed.

[0072] Figure 33 depicts how atoms are associated with voxels that contain the atoms, in accordance with one implementation of the technology disclosed.

[0073] Figure 34 shows generating voxel-to-atoms mapping from atom-to-voxels mapping to identify nearest atoms on a voxel-by-voxel basis, in accordance with one implementation of the technology disclosed.

[0074] Figures 35A and 35B illustrate how the disclosed efficient voxelization has a runtime complexity of $O(\#atoms)$ versus the runtime complexity of $O(\#atoms * \#voxels)$ without the use of disclosed efficient voxelization.

[0075] Figure 36 shows an example computer system that can be used to implement the technology disclosed.

DETAILED DESCRIPTION

[0076] The following discussion is presented to enable any person skilled in the art to make and use the technology disclosed and is provided in the context of a particular application and its requirements. Various modifications to the disclosed implementations will be readily apparent to those skilled in the art, and the general principles defined herein may be applied to other implementations and applications without departing from the spirit and scope of the technology disclosed. Thus, the technology disclosed is not intended to be limited to the implementations shown but is to be accorded the widest scope consistent with the principles and features disclosed herein.

[0077] The detailed description of various implementations will be better understood when read in conjunction with the appended drawings. To the extent that the figures illustrate diagrams of the functional blocks of the various implementations, the functional blocks are not necessarily indicative of the division between hardware circuitry. Thus, for example, one or more of the functional blocks (*e.g.*, modules, processors, or memories) may be implemented in a single piece of hardware (*e.g.*, a general purpose signal processor or a block of random access memory, hard disk, or the like) or multiple pieces of hardware. Similarly, the programs may be stand-alone

programs, may be incorporated as subroutines in an operating system, may be functions in an installed software package, and the like. It should be understood that the various implementations are not limited to the arrangements and instrumentality shown in the drawings.

[0078] The processing engines and databases of the figures, designated as modules, can be implemented in hardware or software, and need not be divided up in precisely the same blocks as shown in the figures. Some of the modules can also be implemented on different processors, computers, or servers, or spread among a number of different processors, computers, or servers. In addition, it will be appreciated that some of the modules can be combined, operated in parallel or in a different sequence than that shown in the figures without affecting the functions achieved. The modules in the figures can also be thought of as flowchart steps in a method. A module also need not necessarily have all its code disposed contiguously in memory; some parts of the code can be separated from other parts of the code with code from other modules or other functions disposed in between.

Protein Structure-Based Pathogenicity Determination

[0079] Figure 1 is a flow diagram that illustrates a process 100 of a system for determining pathogenicity of variants. At step 102, a sequence accessor 104 of the system accesses reference and alternative amino acid sequences. At 112, a 3D structure generator 114 of the system generates 3D protein structures for a reference amino acid sequence. In some implementations, the 3D protein structures are homology models of human proteins. In one implementation, a so-called SwissModel homology modelling pipeline provides a public repository of predicted human protein structures. In another implementation, a so-called HHpred homology modelling uses a tool called Modeller to predict the structure of a target protein from template structures.

[0080] Proteins are represented by a collection of atoms and their coordinates in 3D space. An amino acid can have a variety of atoms, such as carbon atoms, oxygen (O) atoms, nitrogen (N) atoms, and hydrogen (H) atoms. The atoms can be further classified as side chain atoms and backbone atoms. The backbone carbon atoms can include alpha-carbon (C_{α}) atoms and beta-carbon (C_{β}) atoms.

[0081] At step 122, a coordinate classifier 124 of the system classifies 3D atomic coordinates of the 3D protein structures on an amino acid-basis. In one implementation, the amino acid-wise classification involves attributing the 3D atomic coordinates to the twenty-one amino acid categories (including stop or gap amino acid category). In one example, an amino acid-wise classification of alpha-carbon atoms can respectively list alpha-carbon atoms under each of the twenty-one amino acid categories. In another example, an amino acid-wise classification of beta-

carbon atoms can respectively list beta-carbon atoms under each of the twenty-one amino acid categories.

[0082] In yet another example, an amino acid-wise classification of oxygen atoms can respectively list oxygen atoms under each of the twenty-one amino acid categories. In yet another example, an amino acid-wise classification of nitrogen atoms can respectively list nitrogen atoms under each of the twenty-one amino acid categories. In yet another example, an amino acid-wise classification of hydrogen atoms can respectively list hydrogen atoms under each of the twenty-one amino acid categories.

[0083] A person skilled in the art will appreciate that, in various implementations, the amino acid-wise classification can include a subset of the twenty-one amino acid categories and a subset of the different atomic elements.

[0084] At step 132, a voxel grid generator 134 of the system instantiates a voxel grid. The voxel grid can have any resolution, for example, 3x3x3, 5x5x5, 7x7x7, and so on. Voxels in the voxel grid can be of any size, for example, one angstrom (Å) on each side, two Å on each side, three Å on each side, and so on. One skilled in the art will appreciate that these example dimensions refer to cubic dimensions because voxels are cubes. Also, one skilled in the art will appreciate that these example dimensions are non-limiting, and the voxels can have any cubic dimensions.

[0085] At step 142, a voxel grid centerer 144 of the system centers the voxel grid at the reference amino acid experiencing a target variant at the amino acid level. In one implementation, the voxel grid is centered at an atomic coordinate of a particular atom of the reference amino acid experiencing the target variant, for example, the 3D atomic coordinate of the alpha-carbon atom of the reference amino acid experiencing the target variant.

Distance Channels

[0086] The voxels in the voxel grid can have a plurality of channels (or features). In one implementation, the voxels in the voxel grid have a plurality of distance channels (*e.g.*, twenty-one distance channels for the twenty-one amino acid categories, respectively (including stop or gap amino acid category)). At step 152, a distance channel generator 154 of the system generates amino acid-wise distance channels for the voxels in the voxel grid. The distance channels are independently generated for each of the twenty-one amino acid categories.

[0087] Consider, for example, the Alanine (A) amino acid category. Further consider, for example, that the voxel grid is of size 3x3x3 and has twenty-seven voxels. Then, in one implementation, an Alanine distance channel includes twenty-seven distance values for the twenty-seven voxels in the voxel grid, respectively. The twenty-seven distance values in the

Alanine distance channel are measured from respective centers of the twenty-seven voxels in the voxel grid to respective nearest atoms in the Alanine amino acid category.

[0088] In one example, the Alanine amino acid category includes only alpha-carbon atoms and therefore the nearest atoms are those Alanine alpha-carbon atoms that are most proximate to the twenty-seven voxels in the voxel grid, respectively. In another example, the Alanine amino acid category includes only beta-carbon atoms and therefore the nearest atoms are those Alanine beta-carbon atoms that are most proximate to the twenty-seven voxels in the voxel grid, respectively.

[0089] In yet another example, the Alanine amino acid category includes only oxygen atoms and therefore the nearest atoms are those Alanine oxygen atoms that are most proximate to the twenty-seven voxels in the voxel grid, respectively. In yet another example, the Alanine amino acid category includes only nitrogen atoms and therefore the nearest atoms are those Alanine nitrogen atoms that are most proximate to the twenty-seven voxels in the voxel grid, respectively. In yet another example, the Alanine amino acid category includes only hydrogen atoms and therefore the nearest atoms are those Alanine hydrogen atoms that are most proximate to the twenty-seven voxels in the voxel grid, respectively.

[0090] Like the Alanine distance channel, the distance channel generator 154 generates a distance channel (*i.e.*, a set of voxel-wise distance values) for each of the remaining amino acid categories. In other implementations, the distance channel generator 154 generates distance channels only for a subset of the twenty-one amino acid categories.

[0091] In other implementations, the selection of the nearest atoms is not confined to a particular atom type. That is, within a subject amino acid category, the nearest atom to a particular voxel is selected, irrespective of the atomic element of the nearest atom, and the distance value for the particular voxel calculated for inclusion in the distance channel for the subject amino acid category.

[0092] In yet other implementations, the distance channels are generated on an atomic element-basis. Instead of or in addition to having the distance channels for the amino acid categories, distance values can be generated for atom element categories, irrespective of the amino acids to which the atoms belong. Consider, for example, that the atoms of amino acids in the reference amino acid sequence span seven atomic elements: carbon, oxygen, nitrogen, hydrogen, calcium, iodine, and sulfur. Then, the voxels in the voxel grid are configured to have seven distance channels, such that each of the seven distance channels have twenty-seven voxel wise distance values that specify distances to nearest atoms only within a corresponding atomic element category. In other implementations, distance channels for only a subset of the seven atomic elements can be generated. In yet other implementations, the atomic element categories

and the distance channel generation can be further stratified into variations of a same atomic element, for example, alpha-carbon (C_α) atoms and beta-carbon (C_β) atoms.

[0093] In yet other implementations, the distance channels can be generated on an atom type-basis, for example, distance channels only for side chain atoms and distance channels only for backbone atoms.

[0094] The nearest atoms can be searched within a predefined maximum scan radius from the voxel centers (*e.g.*, six angstrom ($\text{\AA}$)). Also, multiple atoms can be nearest to a same voxel in the voxel grid.

[0095] The distances are calculated between 3D coordinates of the voxel centers and 3D atomic coordinates of the atoms. Also, the distance channels are generated with the voxel grid centered at a same location (*e.g.*, centered at the 3D atomic coordinate of the alpha-carbon atom of the reference amino acid experiencing the target variant).

[0096] The distances can be Euclidean distances. Also, the distances can be parameterized by atom size (or atom influence) (*e.g.*, by using Lennard-Jones potential and/or Van der Waals atom radius of the atom in question). Also, the distance values can be normalized by the maximum scan radius, or by a maximum observed distance value of the furthest nearest atom within a subject amino acid category or a subject atomic element category or a subject atom type category. In some implementations, the distances between the voxels and the atoms are calculated based on polar coordinates of the voxels and the atoms. The polar coordinates are parameterized by angles between the voxels and the atoms. In one implementation, this angle information is used to generate an angle channel for the voxels (*i.e.*, independent of the distance channels). In some implementations, angles between a nearest atom and neighboring atoms (*e.g.*, backbone atoms) can be used as features that are encoded with the voxels.

Reference Allele and Alternative Allele Channels

[0097] The voxels in the voxel grid can also have reference allele and alternative allele channels. At step 162, a one-hot encoder 164 of the system generates a reference one-hot encoding of a reference amino acid in the reference amino acid sequence and an alternative one-hot encoding of an alternative amino acid in an alternative amino acid sequence. The reference amino acid experiences the target variant. The alternative amino acid is the target variant. The reference amino acid and the alternative amino acid are located at a same position respectively in the reference amino acid sequence and the alternative amino acid sequence. The reference amino acid sequence and the alternative amino acid sequence have the same position-wise amino acid composition with one exception. The exception is the position that has the reference amino acid

in the reference amino acid sequence and the alternative amino acid in the alternative amino acid sequence.

[0098] At step 172, a concatenator 174 of the system concatenates the amino acid-wise distance channels and the reference and alternative one-hot encodings. In another implementation, the concatenator 174 concatenates the atomic element-wise distance channels and the reference and alternative one-hot encodings. In yet another implementation, the concatenator 174 concatenates the atomic type-wise distance channels and the reference and alternative one-hot encodings.

[0099] At step 182, runtime logic 184 of the system processes the concatenated amino acid-wise/atomic element-wise/atomic type-wise distance channels and the reference and alternative one-hot encodings through a pathogenicity classifier (pathogenicity determination engine) to determine a pathogenicity of the target variant, which is in turn inferred as a pathogenicity determination of the underlying nucleotide variant that creates the target variant at the amino acid level. The pathogenicity classifier is trained using labelled datasets of benign and pathogenic variants, for example, using the backpropagation algorithm. Additional details about the labelled datasets of benign and pathogenic variants and example architectures and training of the pathogenicity classifier can be found in commonly owned US Patent Application Nos. 16/160,903; 16/160,986; 16/160,968; and 16/407,149.

[00100] Figure 2 schematically illustrates a reference amino acid sequence 202 of a protein 200 and an alternative amino acid sequence 212 of the protein 200. The protein 200 comprises N amino acids. Positions of the amino acids in the protein 200 are labelled 1, 2, 3...N. In the illustrated example, position 16 is the location that experiences an amino acid variant 214 (mutation) caused by an underlying nucleotide variant. For example, for the reference amino acid sequence 202, position 1 has reference amino acid Phenylalanine (F), position 16 has reference amino acid Glycine (G) 204, and position N (e.g., the last amino acid of the sequence 202) has reference amino acid Leucine (L). Though not illustrated for clarity, remaining positions in the reference amino acid sequence 202 contain various amino acids in an order that is specific to the protein 200. The alternative amino acid sequence 212 is the same as the reference amino acid sequence 202 except for the variant 214 at position 16, which contains the alternative amino acid Alanine (A) 214 instead of the reference amino acid Glycine (G) 204.

[00101] Figure 3 illustrates amino acid-wise classification of atoms of amino acids in the reference amino acid sequence 202, also referred to herein as “atom classification 300.” Specific types of amino acids, among the twenty natural amino acids listed in column 302, may repeat in a protein. That is, a particular type of amino acid may occur more than once in a protein. Proteins may also have some undetermined amino acids that are categorized by a twenty-first stop or gap

amino acid category. The right column in Figure 3 contains counts of alpha-carbon (C_α) atoms from different amino acids.

[00102] Specifically, Figure 3 shows amino acid-wise classification of alpha-carbon (C_α) atoms of the amino acids in the reference amino acid sequence 202. Column 308 of Figure 3 lists the total number of alpha-carbon atoms observed for the reference amino acid sequence 202 in each of the twenty-one amino acid categories. For example, column 308 lists eleven alpha-carbon atoms observed for the Alanine (A) amino acid category. Since each amino acid has only one alpha-carbon atom, this means that Alanine occurs 11 times in the reference amino acid sequence 202. In another example, Arginine (R) occurs thirty-five times in the reference amino acid sequence 202. The total number of alpha-carbon atoms across the twenty-one amino acid categories is eight hundred and twenty-eight.

[00103] Figure 4 illustrates amino acid-wise attribution of 3D atomic coordinates of the alpha-carbon atoms of the reference amino acid sequence 202 based on the atom classification 300 in Figure 3. This is referred to herein as “atomic coordinates bucketing 400.” In Figure 4, lists 404-440 tabulate the 3D atomic coordinates of the alpha-carbon atoms bucketed to each of the twenty-one amino acid categories.

[00104] In the illustrated implementation, the bucketing 400 in Figure 4 follows the classification 300 of Figure 3. For example, in Figure 3, the Alanine amino acid category has eleven alpha-carbon atoms, and therefore, in Figure 4, the Alanine amino acid category has eleven 3D atomic coordinates of the corresponding eleven alpha-carbon atoms from Figure 3. This classification-to-bucketing logic flows from Figure 3 to Figure 4 for other amino acid categories too. However, this classification-to-bucketing logic is only for representational purposes, and, in other implementations, the technology disclosed need not perform the classification 300 and the bucketing 400 to locate the voxel-wise nearest atoms, and may perform fewer, additional, or different steps. For example, in some implementations, the technology disclosed can locate the voxel-wise nearest atoms by using a sort and search algorithm that returns the voxel-wise nearest atoms from one or more databases in response to a search query configured to accept query parameters like sort criteria (*e.g.*, amino acid-wise, atomic element-wise, atom type-wise), the predefined maximum scan radius, and the type of distances (*e.g.*, Euclidean, Mahalanobis, normalized, unnormalized). In various implementations of the technology disclosed, a plurality of sort and search algorithms from the current or future technical field can be analogous used by a person skilled in the art to locate the voxel-wise nearest atoms.

[00105] In Figure 4, the 3D atomic coordinates are represented by cartesian coordinates x , y , z , but any type of coordinate system may be used, such as spherical or cylindrical coordinates,

and claimed subject matter is not limited in this respect. In some implementations, one or more databases may include information regarding the 3D atomic coordinates of the alpha-carbon atoms and other atoms of amino acids in proteins. Such databases may be searchable by specific proteins.

[00106] As discussed above, the voxels and the voxel grid are 3D entities. However, for clarity's sake, the drawings depict, and the description discusses the voxels and the voxel grid in a two-dimensional (2D) format. For example, a 3x3x3 voxel grid of twenty-seven voxels is depicted and described herein as a 3x3 2D pixel grid with nine 2D pixels. A person skilled in the art will appreciate that the 2D format is used only for representational purposes and is intended to cover the 3D counterparts (*i.e.*, 2D pixels represent 3D voxels and 2D pixel grid represents 3D voxel grid). Also, the drawings are also not scale. For example, voxels of size two angstrom (Å) are depicted using a single pixel.

Voxel-Wise Distance Calculation

[00107] Figure 5 schematically illustrates a process of determining voxel-wise distance values, also referred to herein as “voxel-wise distance calculation 500.” In the illustrated example, the voxel-wise distance values are calculated only for the Alanine (A) distance channel. However, the same distance calculation logic is executed for each of the twenty-one amino acid categories to generate twenty-one amino acid-wise distance channels and can be further expanded to other atom types like beta-carbon atoms and other atomic elements like oxygen, nitrogen, and hydrogen, as discussed above with respect to Figure 1. In some implementations, the atoms are randomly rotated prior to the distance calculation to make the training of the pathogenicity classifier invariant to atom orientation.

[00108] In Figure 5, a voxel grid 522 has nine voxels 514 identified with indices (1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), and (3, 3). The voxel grid 522 is centered, for example, at the 3D atomic coordinate 532 of the alpha-carbon atom of the Glycine (G) amino acid at position 16 in the reference amino acid sequence 202 because, in the alternative amino acid sequence 212, the position 16 experiences the variant that mutates the Glycine (G) amino acid to the Alanine (A) amino acid, as discussed above with respect to Figure 2. Also, the center of the voxel grid 522 coincides with the center of voxel (2, 2).

[00109] The centered voxel grid 522 is used for the voxel-wise distance calculation for each of the twenty-one amino acid-wise distance channels. Starting, for example, with the Alanine (A) distance channel, distances between the 3D coordinates of respective centers of the nine voxels 514 and the 3D atomic coordinates 402 of the eleven Alanine alpha-carbon atoms are measured to locate a nearest Alanine alpha-carbon atom for each of the nine voxels 514. Then, nine

distance values for nine distances between the nine voxels 514 and the respective nearest Alanine alpha-carbon atoms are used to construct the Alanine distance channel. The resulting Alanine distance channel arranges the nine Alanine distance values in the same order as the nine voxels 514 in the voxel grid 522.

[00110] The above process is executed for each of the twenty-one amino acid categories. For example, the centered voxel grid 522 is similarly used to calculate the Arginine (R) distance channel, such that distances between the 3D coordinates of respective centers of the nine voxels 514 and the 3D atomic coordinates 404 of the thirty-five Arginine alpha-carbon atoms are measured to locate a nearest Arginine alpha-carbon atom for each of the nine voxels 514. Then, nine distance values for nine distances between the nine voxels 514 and the respective nearest Arginine alpha-carbon atoms are used to construct the Arginine distance channel. The resulting Arginine distance channel arranges the nine Arginine distance values in the same order as the nine voxels 514 in the voxel grid 522. The twenty-one amino acid-wise distance channels are voxel-wise encoded to form a distance channel tensor.

[00111] Specifically, in the illustrated example, a distance 512 is between the center of voxel (1, 1) of voxel grid 522 and the nearest alpha-carbon (C_α) atom, which is the $C\alpha^{A5}$ atom in list 402. Accordingly, the value assigned to voxel (1, 1) is the distance 512. In another example, the $C\alpha^{A4}$ atom is the nearest C_α atom to the center of voxel (1, 2). Accordingly, the value assigned to voxel (1, 2) is the distance between the center of voxel (1, 2) and the $C\alpha^{A4}$ atom. In still another example, the $C\alpha^{A6}$ atom is the nearest C_α atom to the center of voxel (2, 1). Accordingly, the value assigned to voxel (2, 1) is the distance between the center of voxel (2, 1) and the $C\alpha^{A6}$ atom. In still another example, the $C\alpha^{A6}$ atom is also the nearest C_α atom to the center of voxels (3, 2) and (3, 3). Accordingly, the value assigned to voxel (3, 2) is the distance between the center of voxel (3, 2) and the $C\alpha^{A6}$ atom and the value assigned to voxel (3, 3) is the distance between the center of voxel (3, 3) and the $C\alpha^{A6}$ atom. In some implementations, the distance values assigned to the voxels 514 may be normalized distances. For example, the distance value assigned to voxel (1, 1) may be the distance 512 divided by a maximum distance 502 (predefined maximum scan radius). In some implementations, the nearest-atom distances may be Euclidean distances and the nearest-atom distances may be normalized by dividing the Euclidean distances with a maximum nearest-atom distance (*e.g.*, such as the maximum distance 502).

[00112] As described above, for amino acids having alpha-carbon atoms, the distances may be nearest-alpha-carbon atom distances from corresponding voxel centers to nearest alpha-carbon atoms of the corresponding amino acids. Additionally, for amino acids having beta-carbon atoms, the distances may be nearest-beta-carbon atom distances from corresponding voxel centers to nearest beta-carbon atoms of the corresponding amino acids. Similarly, for amino

acids having backbone atoms, the distances may be nearest-backbone atom distances from corresponding voxel centers to nearest backbone atoms of the corresponding amino acids. Similarly, for amino acids having sidechain atoms, the distances may be nearest-sidechain atom distances from corresponding voxel centers to nearest sidechain atoms of the corresponding amino acids. In some implementations, the distances additionally/alternatively can include distances to second, third, fourth nearest atoms, and so on.

Amino Acid-Wise Distance Channels

[00113] Figure 6 shows an example of twenty-one amino acid-wise distance channels 600. Each column in Figure 6 corresponds to a respective one of the twenty-one amino acid-wise distance channels 602-642. Each amino acid-wise distance channel comprises a distance value for each of the voxels 514 of the voxel grid 522. For example, the amino acid-wise distance channel 602 for Alanine (A) comprises distance values for respective ones of the voxels 514 of the voxel grid 522. As mentioned above, the voxel grid 522 is 3D grid of volume 3x3x3 and comprises twenty-seven voxels. Likewise, though Figure 6 illustrates the voxels 514 in two dimensions (*e.g.*, nine voxels of a 3x3 grid), each amino acid-wise distance channel may comprise twenty-seven voxel-wise distance values for the 3x3x3 voxel grid.

Directionality Encoding

[00114] In some implementations, the technology disclosed uses a directionality parameter to specify the directionality of the reference amino acids in the reference amino acid sequence 202. In some implementations, the technology disclosed uses the directionality parameter to specify the directionality of the alternative amino acids in the alternative amino acid sequence 212. In some implementations, the technology disclosed uses the directionality parameter to specify the position in the protein 200 that experiences the target variant at the amino acid level.

[00115] As discussed above, all the distance values in the twenty-one amino acid-wise distance channels 602-642 are measured from respective nearest atoms to the voxels 514 in the voxel grid 522. These nearest atoms originate from one of the reference amino acids in the reference amino acid sequence 202. These originating reference amino acids, which contain the nearest atoms, can be classified into two categories: (1) those originating reference amino acids that precede the variant-experiencing reference amino acid 204 in the reference amino acid sequence 202 and (2) those originating reference amino acids that succeed the variant-experiencing reference amino acid 204 in the reference amino acid sequence 202. The originating reference amino acids in the first category can be called preceding reference amino

acids. The originating reference amino acids in the second category can be called succeeding reference amino acids.

[00116] The directionality parameter is applied to those distance values in the twenty-one amino acid-wise distance channels 602-642 that are measured from those nearest atoms that originate from the preceding reference amino acids. In one implementation, the directionality parameter is multiplied with such distance values. The directionality parameter can be any number, such as -1.

[00117] As a result of the application of the directionality parameter, the twenty-one amino acid-wise distance channels 600 include some distance values that indicate to the pathogenicity classifier which end of the protein 200 is the start terminal and which end is the end terminal. This also allows the pathogenicity classifier to reconstruct a protein sequence from the 3D protein structure information supplied by the distance channels and the reference and allele channels.

Distance Channel Tensor

[00118] Figure 7 is a schematic diagram of a distance channel tensor 700. Distance channel tensor 700 is a voxelized representation of the amino acid-wise distance channels 600 from Figure 6. In the distance channel tensor 700, the twenty-one amino acid-wise distance channels 602-642 are concatenated voxel-wise, like RGB channels of a color image. The voxelized dimensionality of the distance channel tensor 700 is $21 \times 3 \times 3 \times 3$ (where 21 denotes the twenty-one amino acid categories and $3 \times 3 \times 3$ denotes the 3D voxel grid with twenty-seven voxels); although Figure 7 is a 2D depiction of dimensionality $21 \times 3 \times 3$.

One-Hot Encodings

[00119] Figure 8 shows one-hot encodings 800 of the reference amino acid 204 and the alternative amino acid 214. In Figure 8, left column is a one-hot encoding 802 of the reference amino acid Glycine (G) 204, with one for the Glycine amino acid category and zeros for all other amino acid categories. In Figure 8, right column is a one-hot encoding 804 of the variant/alternative amino acid Alanine (A) 214, with one for the Alanine amino acid category and zeros for all other amino acid categories.

[00120] Figure 9 is a schematic diagram of a voxelized one-hot encoded reference amino acid 902 and a voxelized one-hot encoded variant/alternative amino acid 912. The voxelized one-hot encoded reference amino acid 902 is a voxelized representation of the one-hot encoding 802 of the reference amino acid Glycine (G) 204 from Figure 8. The voxelized one-hot encoded alternative amino acid 912 is a voxelized representation of the one-hot encoding 804 of the

variant/alternative amino acid Alanine (A) 214 from Figure 8. The voxelized dimensionality of the voxelized one-hot encoded reference amino acid 902 is $21 \times 1 \times 1 \times 1$ (where 21 denotes the twenty-one amino acid categories); although Figure 9 is a 2D depiction of dimensionality $21 \times 1 \times 1$. Similarly, the voxelized dimensionality of the voxelized one-hot encoded alternative amino acid 912 is $21 \times 1 \times 1 \times 1$ (where 21 denotes the twenty-one amino acid categories); although Figure 9 is a 2D depiction of dimensionality $21 \times 1 \times 1$.

Reference Allele Tensor

[00121] Figure 10 schematically illustrates a concatenation process 1000 that voxel-wise concatenates the distance channel tensor 700 of Figure 7 and a reference allele tensor 1004. The reference allele tensor 1004 is a voxel-wise aggregation (repetition/cloning/replication) of the voxelized one-hot encoded reference amino acid 902 from Figure 9. That is, multiple copies of the voxelized one-hot encoded reference amino acid 902 are voxel-wise concatenated according with each other to the spatial arrangement of the voxels 514 in the voxel grid 522, such that the reference allele tensor 1004 has a corresponding copy of the voxelized one-hot encoded reference amino acid 910 for each of the voxels 514 in the voxel grid 522.

[00122] The concatenation process 1000 produces a concatenated tensor 1010. The voxelized dimensionality of the reference allele tensor 1004 is $21 \times 3 \times 3 \times 3$ (where 21 denotes the twenty-one amino acid categories and $3 \times 3 \times 3$ denotes the 3D voxel grid with twenty-seven voxels); although Figure 10 is a 2D depiction of the reference allele tensor 1004 having dimensionality $21 \times 3 \times 3$. The voxelized dimensionality of the concatenated tensor 1010 is $42 \times 3 \times 3 \times 3$; although Figure 10 is a 2D depiction of the concatenated tensor 1010 having dimensionality $42 \times 3 \times 3$.

Alternative Allele Tensor

[00123] Figure 11 schematically illustrates a concatenation process 1100 that voxel-wise concatenates the distance channel tensor 700 of Figure 7, the reference allele tensor 1004 of Figure 10, and an alternative allele tensor 1104. The alternative allele tensor 1104 is a voxel-wise aggregation (repetition/cloning/replication) of the voxelized one-hot encoded alternative amino acid 912 from Figure 9. That is, multiple copies of the voxelized one-hot encoded alternative amino acid 912 are voxel-wise concatenated with each other according to the spatial arrangement of the voxels 514 in the voxel grid 522, such that the alternative allele tensor 1104 has a corresponding copy of the voxelized one-hot encoded alternative amino acid 910 for each of the voxels 514 in the voxel grid 522.

[00124] The concatenation process 1100 produces a concatenated tensor 1110. The voxelized dimensionality of the alternative allele tensor 1104 is $21 \times 3 \times 3 \times 3$ (where 21 denotes the twenty-

one amino acid categories and $3 \times 3 \times 3$ denotes the 3D voxel grid with twenty-seven voxels); although Figure 11 is a 2D depiction of the alternative allele tensor 1104 having dimensionality $21 \times 3 \times 3$. The voxelized dimensionality of the concatenated tensor 1110 is $63 \times 3 \times 3 \times 3$; although Figure 11 is a 2D depiction of the concatenated tensor 1110 having dimensionality $63 \times 3 \times 3$.

[00125] In some implementations, the runtime logic 184 processes the concatenated tensor 1110 through the pathogenicity classifier to determine a pathogenicity of the variant/alternative amino acid Alanine (A) 214, which is in turn inferred as a pathogenicity determination of the underlying nucleotide variant that creates the variant/alternative amino acid Alanine (A) 214.

Evolutionary Conservation Channels

[00126] Predicting the functional consequences of variants relies at least in part on the assumption that crucial amino acids for protein families are conserved through evolution due to negative selection (*i.e.*, amino acid changes at these sites were deleterious in the past), and that mutations at these sites have an increased likelihood of being pathogenic (causing disease) in humans. In general, homologous sequences of a target protein are collected and aligned, and a metric of conservation is computed based on the weighted frequencies of different amino acids observed in the target position in the alignment.

[00127] Accordingly, the technology disclosed concatenates the distance channel tensor 700, the reference allele tensor 1004, and the alternative allele tensor 1104 with evolutionary channels. One example of the evolutionary channels is pan-amino acid conservation frequencies. Another example of the evolutionary channels is per-amino acid conservation frequencies.

[00128] In some implementations, the evolutionary channels are constructed using position weight matrices (PWMs). In other implementations, the evolutionary channels are constructed using position specific frequency matrices (PSFMs). In yet other implementations, the evolutionary channels are constructed using computational tools like SIFT, PolyPhen, and PANTHER-PSEC. In yet other implementations, the evolutionary channels are preservation channels based on evolutionary preservation. Preservation is related to conservation, as it also reflects the effect of negative selection that has acted to prevent evolutionary change at a given site in a protein.

Pan-Amino Acid Evolutionary Profiles

[00129] Figure 12 is a flow diagram that illustrates a process 1200 of a system for determining and assigning pan-amino acid conservation frequencies of nearest atoms to voxels (voxelizing), in accordance with one implementation of the technology disclosed. Figures 12, 13, 14, 15, 16, 17, and 18 are discussed in tandem.

[00130] At step 1202, a similar sequence finder 1204 of the system retrieves amino acid sequences that are similar (homologous) to the reference amino acid sequence 202. The similar amino acid sequences can be selected from multiple species like primates, mammals, and vertebrates.

[00131] At step 1212, an aligner 1214 of the system position-wise aligns the reference amino acid sequence 202 with the similar amino acid sequences, *i.e.*, the aligner 1214 performs a multi-sequence alignment. Figure 14 shows an example multi-sequence alignment 1400 of the reference amino acid sequence 202 across a ninety-nine species. In some implementations, the multi-sequence alignment 1400 can be partitioned, for example, to generate a first position frequency matrix 1402 for primates, a second position frequency matrix 1412 for mammals, and a third position frequency matrix 1422 for primates. In other implementations, a single position frequency matrix is generated across the ninety-nine species.

[00132] At step 1222, a pan-amino acid conservation frequency calculator 1224 of the system uses the multi-sequence alignment to determine pan-amino acid conservation frequencies of the reference amino acids in the reference amino acid sequence 202.

[00133] At step 1232, a nearest atom finder 1234 of the system finds nearest atoms to the voxels 514 in the voxel grid 522. In some implementations, the search for the voxel-wise nearest atoms may not be confined to any particular amino acid category or atom type. That is, the voxel-wise nearest atoms can be selected across the amino acid categories and the amino acid types, as long as they are the most proximate atoms to the respective voxel centers. In other implementations, the search for the voxel-wise nearest atoms may be confined to only a particular atom category, such as only to a particular atomic element like oxygen, nitrogen, and hydrogen, or only to alpha-carbon atoms, or only to beta-carbon atoms, or only to sidechain atoms, or only to backbone atoms.

[00134] At step 1242, an amino acid selector 1244 of the system selects those reference amino acids in the reference amino acid sequence 202 that contain the nearest atoms identified at the step 1232. Such reference amino acids can be called nearest reference amino acids. Figure 13 shows an example of locating nearest atoms 1302 to the voxels 514 in the voxel grid 522 and respectively mapping nearest reference amino acids 1312 that contain the nearest atoms 1302 to the voxels 514 in the voxel grid 522. This is identified in Figure 13 as “voxels-to-nearest amino acids mapping 1300.”

[00135] At step 1252, a voxelizer 1254 of the system voxelizes pan-amino acid conservation frequencies of the nearest reference amino acids. Figure 15 shows an example of determining a pan-amino acid conservation frequencies sequence for the first voxel (1, 1) in the voxel grid 522, also referred to herein as “per-voxel evolutionary profile determination 1500.”

[00136] Turning to Figure 13, the nearest reference amino acid that was mapped to the first voxel (1, 1) is Aspartic acid (D) amino acid at position 15 in the reference amino acid sequence 202. Then, the multi-sequence alignment of the reference amino acid sequence 202 with, for example, ninety-nine homologous amino acid sequences of the ninety-nine species is analyzed at position 15. Such a position-specific and cross-species analysis reveals how many instances of amino acids from each of the twenty-one amino acid categories are found at position 15 across the hundred aligned amino acid sequences (*i.e.*, the reference amino acid sequence 202 plus the ninety-nine homologous amino acid sequences).

[00137] In the example illustrated in Figure 15, the Aspartic acid (D) amino acid is found at position 15 in ninety-six out of the hundred aligned amino acid sequences. So, the Aspartic acid amino acid category 1504 is assigned a pan-amino acid conservation frequency of 0.96. Similarly, in the illustrated example, the Valine (V) acid amino acid is found at position 15 in four out of the hundred aligned amino acid sequences. So, the Valine acid amino acid category 1514 is assigned a pan-amino acid conservation frequency of 0.04. Since no instances of amino acids from other amino acid categories are detected at position 15, the remaining amino acid categories are assigned a pan-amino acid conservation frequency of zero. This way, each of the twenty-one amino acid categories is assigned a respective pan-amino acid conservation frequency, which can be encoded in the pan-amino acid conservation frequencies sequence 1502 for the first voxel (1, 1).

[00138] Figure 16 shows respective pan-amino acid conservation frequencies 1612-1692 determined for respective ones of the voxels 514 in the voxel grid 522 using the position frequency logic described in Figure 15, also referred to herein as “voxels-to- evolutionary profiles mapping 1600.”

[00139] Per-voxel evolutionary profiles 1602 are then used by the voxelizer 1254 to generate voxelized per-voxel evolutionary profiles 1700, illustrated in Figure 17. Often, each of the voxels 514 in the voxel grid 522 has a different pan-amino acid conservation frequencies sequence and therefore a different voxelized per-voxel evolutionary profile because the voxels are regularly mapped to different nearest atoms and therefore to different nearest reference amino acids. Of course, when two or more voxels have a same nearest atom and thereby a same nearest reference amino acid, a same pan-amino acid conservation frequencies sequence and a same voxelized per-voxel evolutionary profile is assigned to each of the two or more voxels.

[00140] Figure 18 depicts example of an evolutionary profiles tensor 1800 in which the voxelized per-voxel evolutionary profiles 1700 are voxel-wise concatenated with each other according to the spatial arrangement of the voxels 514 in the voxel grid 522. The voxelized dimensionality of the evolutionary profiles tensor 1800 is 21x3x3x3 (where 21 denotes the

twenty-one amino acid categories and $3 \times 3 \times 3$ denotes the 3D voxel grid with twenty-seven voxels); although Figure 18 is a 2D depiction of the evolutionary profiles tensor 1800 having dimensionality $21 \times 3 \times 3$.

[00141] At step 1262, the concatenator 174 voxel-wise concatenates the evolutionary profiles tensor 1800 with the distance channel tensor 700. In some implementations, the evolutionary profiles tensor 1800 is voxel-wise concatenated with the concatenator tensor 1110 to generate a further concatenated tensor of dimensionality $84 \times 3 \times 3 \times 3$ (not shown).

[00142] At step 1272, the runtime logic 184 processes the further concatenated tensor of dimensionality $84 \times 3 \times 3 \times 3$ through the pathogenicity classifier to determine the pathogenicity of the target variant, which is in turn inferred as a pathogenicity determination of the underlying nucleotide variant that creates the target variant at the amino acid level.

Per-Amino Acid Evolutionary Profiles

[00143] Figure 19 is a flow diagram that illustrates a process 1900 of a system for determining and assigning per-amino acid conservation frequencies of nearest atoms to voxels (voxelizing). In Figure 19, the steps 1202 and 1212 are the same as Figure 12.

[00144] At step 1922, a per-amino acid conservation frequency calculator 1924 of the system uses the multi-sequence alignment to determine per-amino acid conservation frequencies of the reference amino acids in the reference amino acid sequence 202.

[00145] At step 1932, a nearest atom finder 1934 of the system finds, for each of the voxels 514 in the voxel grid 522, twenty-one nearest atoms across each of the twenty-one amino acid categories. Each of the twenty-one nearest atoms is different from each other because they are selected from different amino acid categories. This leads to the selection of twenty-one unique nearest reference amino acids for a particular voxel, which in turn leads to generation of twenty-one unique position frequency matrices for the particular voxel, and which in turn leads to determination of twenty-one unique per-amino acid conservation frequencies for the particular voxel.

[00146] At step 1942, an amino acid selector 1944 of the system selects, for each of the voxels 514 in the voxel grid 522, twenty-one reference amino acids in the reference amino acid sequence 202 that contain the twenty-one nearest atoms identified at the step 1932. Such reference amino acids can be called nearest reference amino acids.

[00147] At step 1952, a voxelizer 1954 of the system voxelizes per-amino acid conservation frequencies of the twenty-one nearest reference amino acids identified for the particular voxel at the step 1942. The twenty-one nearest reference amino acids are necessarily located at twenty-one different positions in the reference amino acid sequence 202 because they correspond to

different underlying nearest atoms. Accordingly, for the particular voxel, twenty-one position frequency matrices can be generated for the twenty-one nearest reference amino acids. The twenty-one position frequency matrices can be generated across multiple species whose homologous amino acid sequences are position-wise aligned with the reference amino acid sequence 202, as discussed above with respect to Figures 12 to 15.

[00148] Then, using the twenty-one position frequency matrices, twenty-one position-specific conservation scores can be calculated for the twenty-one nearest reference amino acids identified for the particular voxel. These twenty-one position-specific conservation scores form the per-amino acid conservation frequencies for the particular voxel, similar to the pan-amino acid conservation frequencies sequence 1502 in Figure 12; except the sequence 1502 has many zero entries, whereas each element (feature) in a per-amino acid conservation frequencies sequence has a value (*e.g.*, a floating point number) because the twenty-one nearest reference amino acids across the twenty-one amino acid categories necessarily have different positions that yield different position frequency matrices and thereby different per-amino acid conservation frequencies.

[00149] The above process is executed for each of the voxels 514 in the voxel grid 522, and the resulting voxel-wise per-amino acid conservation frequencies voxelized, tensorized, concatenated, and processed for pathogenicity determination similar to the pan-amino acid conservation frequencies discussed with respect to Figures 12 to 18.

Annotation Channels

[00150] Figure 20 shows various examples of voxelized annotation channels 2000 that are concatenated with the distance channel tensor 700. In some implementations, the voxelized annotation channels are one-hot indicators for different protein annotations, for example whether an amino acid (residue) is part of a transmembrane region, a signal peptide, an active site, or any other binding site, or whether the residue is subject to posttranslational modifications, PathRatio (*See* Pei P, Zhang A: A Topological Measurement for Weighted Protein Interaction Network. *CSB* 2005, 268–278.), *etc.* Additional examples of the annotation channels can be found below in the Particular Implementations section and in the Claims.

[00151] The voxelized annotation channels are arranged voxel-wise such that the voxels can have a same annotation sequence like the voxelized reference allele and alternative allele sequences (*e.g.*, annotation channels 2002, 2004, 2006), or the voxels can have respective annotation sequences like the voxelized per-voxel evolutionary profiles 1700 (*e.g.*, annotation channels 2012, 2014, 2016 (as indicated by different colors)).

[00152] The annotation channels are voxelized, tensorized, concatenated, and processed for pathogenicity determination similar to the pan-amino acid conservation frequencies discussed with respect to Figures 12 to 18.

Structural Confidence Channels

[00153] The technology disclosed can also concatenate various voxelized structural confidence channels with the distance channel tensor 700. Some examples of the structure confidence channels include GMQE score (provided by SwissModel); B-factor; temperature factor column of homology models (indicates how well a residue satisfies (physical) constraints in the protein structure); normalized number of aligning template proteins for the residue nearest to the center of a voxel (alignments provided by HHpred, *e.g.*, voxel is nearest to a residue at which 3 of 6 template structures align, signifying that the feature has value $3/6=0.5$; minimum, maximum, and mean TM-scores; and predicted TM-scores of the template protein structures that align to the residue that is nearest to a voxel (continuing the example above, assume the 3 template structure has TM-scores 0.5, 0.5 and 1.5, then the minimum is 0.5, the mean is $2/3$, and the maximum is 1.5). The TM-scores can be provided per protein template by HHpred. Additional examples of the structural confidence channels can be found below in the Particular Implementations section and in the Claims.

[00154] The voxelized structural confidence channels are arranged voxel-wise such that the voxels can have a same structural confidence sequence like the voxelized reference allele and alternative allele sequences, or the voxels can have respective structural confidence sequences like the voxelized per-voxel evolutionary profiles 1700.

[00155] The structural confidence channels are voxelized, tensorized, concatenated, and processed for pathogenicity determination similar to the pan-amino acid conservation frequencies discussed with respect to Figures 12 to 18.

Pathogenicity Classifier

[00156] Figure 21 illustrates different combinations and permutations of input channels that can be provided as inputs 2102 to a pathogenicity classifier 2108 for a pathogenicity determination 2106 of a target variant. One of the inputs 2102 can be distance channels 2104 generated by a distance channels generator 2272. Figure 22 shows different methods of calculating the distance channels 2104. In one implementation, the distance channels 2104 are generated based on distances 2202 between voxel centers and atoms across a plurality of atomic elements irrespective of amino acids. In some implementations, the distances 2202 are normalized by a maximum scan radius to generate normalized distances 2202a. In another

implementation, the distance channels 2104 are generated based on distances 2212 between voxel centers and alpha-carbon atoms on an amino acid-basis. In some implementations, the distances 2212 are normalized by the maximum scan radius to generate normalized distances 2212a. In yet another implementation, the distance channels 2104 are generated based on distances 2222 between voxel centers and beta-carbon atoms on an amino acid-basis. In some implementations, the distances 2222 are normalized by the maximum scan radius to generate normalized distances 2222a. In yet another implementation, the distance channels 2104 are generated based on distances 2232 between voxel centers and side chain atoms on an amino acid-basis. In some implementations, the distances 2232 are normalized by the maximum scan radius to generate normalized distances 2232a. In yet another implementation, the distance channels 2104 are generated based on distances 2242 between voxel centers and backbone atoms on an amino acid-basis. In some implementations, the distances 2242 are normalized by the maximum scan radius to generate normalized distances 2242a. In yet another implementation, the distance channels 2104 are generated based on distances 2252 (one feature) between voxel centers and the respective nearest atoms irrespective of atom type and amino acid type. In yet another implementation, the distance channels 2104 are generated based on distances 2262 (one feature) between voxel centers and atoms from non-standard amino acids. In some implementations, the distances between the voxels and the atoms are calculated based on polar coordinates of the voxels and the atoms. The polar coordinates are parameterized by angles between the voxels and the atoms. In one implementation, this angle information is used to generate an angle channel for the voxels (*i.e.*, independent of the distance channels). In some implementations, angles between a nearest atom and neighboring atoms (*e.g.*, backbone atoms) can be used as features that are encoded with the voxels.

[00157] Another one of the inputs 2102 can be a feature 2114 indicating missing atoms within a specified radius.

[00158] Another one of the inputs 2102 can be one-hot encoding 2124 of the reference amino acid. Another one of the inputs 2102 can be one-hot encoding 2134 of the variant/alternative amino acid.

[00159] Another one of the inputs 2102 can be evolutionary channels 2144 generated by an evolutionary profiles generator 2372, shown in Figure 23. In one implementation, the evolutionary channels 2144 can be generated based on pan-amino acid conservation frequencies 2302. In another implementation, the evolutionary channels 2144 can be generated based on pan-amino acid conservation frequencies 2312.

[00160] Another one of the inputs 2102 can be a feature 2154 indicating missing residue or missing evolutionary profile.

[00161] Another one of the inputs 2102 can be annotations channels 2164 generated by an annotations generator 2472, shown in Figure 24. In one implementation, the annotations channels 2164 can be generated based on molecular processing annotations 2402. In another implementation, the annotations channels 2164 can be generated based on regions annotations 2412. In yet another implementation, the annotations channels 2164 can be generated based on sites annotations 2422. In yet another implementation, the annotations channels 2164 can be generated based on Amino acid modifications annotations 2432. In yet another implementation, the annotations channels 2164 can be generated based on secondary structure annotations 2442. In yet another implementation, the annotations channels 2164 can be generated based on experimental information annotations 2452.

[00162] Another one of the inputs 2102 can be structure confidence channels 2174 generated by a structure confidence generator 2572, shown in Figure 25. In one implementation, the structure confidence 2174 can be generated based on global model quality estimations (GMQEs) 2502. In another implementation, the structure confidence 2174 can be generated based on qualitative model energy analysis (QMEAN) scores 2512. In yet another implementation, the structure confidence 2174 can be generated based on temperature factors 2522. In yet another implementation, the structure confidence 2174 can be generated based on template modeling scores 2542. Examples of the template modeling scores 2542 include minimum template modeling scores 2542a, mean template modeling scores 2542b, and maximum template modeling scores 2542c.

[00163] A person skilled in the art will appreciate that any permutation and combination of the input channels can be concatenated into an input for processing through the pathogenicity classifier 2108 for the pathogenicity determination 2106 of the target variant. In some implementations, only a subset of the input channels may be concatenated. The input channels can be concatenated in any order. In one implementation, the input channels can be concatenated into a single tensor by a tensor generator (input encoder) 2110. This single tensor can then be provided as input to the pathogenicity classifier 2108 for the pathogenicity determination 2106 of the target variant.

[00164] In one implementation, the pathogenicity classifier 2108 uses convolutional neural networks (CNNs) with a plurality of convolution layers. In another implementation, the pathogenicity classifier 2108 uses recurrent neural networks (RNNs) such as a long short-term memory networks (LSTMs), bi-directional LSTMs (Bi-LSTMs), and gated recurrent units (GRU)s. In yet another implementation, the pathogenicity classifier 2108 uses both the CNNs and the RNNs. In yet another implementation, the pathogenicity classifier 2108 uses graph-convolutional neural networks that model dependencies in graph-structured data. In yet another

implementation, the pathogenicity classifier 2108 uses variational autoencoders (VAEs). In yet another implementation, the pathogenicity classifier 2108 uses generative adversarial networks (GANs). In yet another implementation, the pathogenicity classifier 2108 can also be a language model based, for example, on self-attention such as the one implemented by Transformers and BERTs.

[00165] In yet other implementations, the pathogenicity classifier 2108 can use 1D convolutions, 2D convolutions, 3D convolutions, 4D convolutions, 5D convolutions, dilated or atrous convolutions, transpose convolutions, depthwise separable convolutions, pointwise convolutions, 1 x 1 convolutions, group convolutions, flattened convolutions, spatial and cross-channel convolutions, shuffled grouped convolutions, spatial separable convolutions, and deconvolutions. It can use one or more loss functions such as logistic regression/log loss, multi-class cross-entropy/softmax loss, binary cross-entropy loss, mean-squared error loss, L1 loss, L2 loss, smooth L1 loss, and Huber loss. It can use any parallelism, efficiency, and compression schemes such TFRecords, compressed encoding (*e.g.*, PNG), sharding, parallel calls for map transformation, batching, prefetching, model parallelism, data parallelism, and synchronous/asynchronous stochastic gradient descent (SGD). It can include upsampling layers, downsampling layers, recurrent connections, gates and gated memory units (like an LSTM or GRU), residual blocks, residual connections, highway connections, skip connections, peephole connections, activation functions (*e.g.*, non-linear transformation functions like rectifying linear unit (ReLU), leaky ReLU, exponential liner unit (ELU), sigmoid and hyperbolic tangent (tanh)), batch normalization layers, regularization layers, dropout, pooling layers (*e.g.*, max or average pooling), global average pooling layers, attention mechanisms, and gaussian error linear unit.

[00166] The pathogenicity classifier 2108 is trained using backpropagation-based gradient update techniques. Example gradient descent techniques that can be used for training the pathogenicity classifier 2108 include stochastic gradient descent, batch gradient descent, and mini-batch gradient descent. Some examples of gradient descent optimization algorithms that can be used to train the pathogenicity classifier 2108 are Momentum, Nesterov accelerated gradient, Adagrad, Adadelta, RMSprop, Adam, AdaMax, Nadam, and AMSGrad. In other implementations, the pathogenicity classifier 2108 can be trained by unsupervised learning, semi-supervised learning, self-learning, reinforcement learning, multitask learning, multimodal learning, transfer learning, knowledge distillation, and so on.

[00167] Figure 26 shows an example processing architecture 2600 of the pathogenicity classifier 2108, in accordance with one implementation of the technology disclosed. The processing architecture 2600 includes a cascade of processing modules 2606, 2610, 2614, 2618, 2622, 2626, 2630, 2634, 2638, and 2642 each of which can include 1D convolutions (1x1x1

CONV), 3D convolutions (3x3x3 CONV), ReLU non-linearity, and batch normalization (BN). Other examples of the processing modules include fully-connected (FC) layers, a dropout layer, a flattening layer, and a final softmax layer that produces exponentially normalized scores for the target variant belonging to a benign class and a pathogenic class. In Figure 26, “64” denotes a number of convolution filters applied by a particular processing module. In Figure 26, the size of an input voxel 2602 is 15x15x15x8. Figure 26 also shows respective volumetric dimensionalities of the intermediate inputs 2604, 2608, 2612, 2616, 2620, 2624, 2628, 2632, 2636, and 2640 generated by the processing architecture 2600.

[00168] Figure 27 shows an example processing architecture 2700 of the pathogenicity classifier 2108, in accordance with one implementation of the technology disclosed. The processing architecture 2700 includes a cascade of processing modules 2708, 2714, 2720, 2726, 2732, 2738, 2744, 2750, 2756, 2762, 2768, 2774, and 2780 such as 1D convolutions (CONV 1D), 3D convolutions (CONV 3D), ReLU non-linearity, and batch normalization (BN). Other examples of the processing modules include fully-connected (dense) layers, a dropout layer, a flattening layer, and a final softmax layer that produces exponentially normalized scores for the target variant belonging to a benign class and a pathogenic class. In Figure 27, “64” and “32” denote a number of convolution filters applied by a particular processing module. In Figure 27, the size of an input voxel 2704 supplied by an input layer 2702 is 7x7x7x108. Figure 27 also shows respective volumetric dimensionalities of the intermediate inputs 2710, 2716, 2722, 2728, 2734, 2740, 2746, 2752, 2758, 2764, 2770, 2776, and 2782 and the resulting intermediate outputs 2706, 2712, 2718, 2724, 2730, 2736, 2742, 2748, 2754, 2760, 2766, 2772, 2778, and 2784 generated by the processing architecture 2700.

[00169] A person skilled in the art will appreciate that other current and future artificial intelligence, machine learning, and deep learning models, datasets, and training techniques can be incorporated in the disclosed variant pathogenicity classifier without deviating from the spirit of the technology disclosed.

Performance Results as Objective Indicia of Inventiveness and Non-Obviousness

[00170] The variant pathogenicity classifier disclosed herein makes pathogenicity predictions based on 3D protein structures and is referred to as “PrimateAI 3D.” “Primate AI” is a commonly owned and previously disclosed variant pathogenicity classifier that makes pathogenicity predictions based protein sequences. Additional details about PrimateAI can be found in commonly owned US Patent Application Nos. 16/160,903; 16/160,986; 16/160,968; and 16/407,149 and in Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* 50, 1161–1170 (2018).

[00171] Figures 28, 29, 30, and 31 use PrimateAI as a benchmark model to demonstrate PrimateAI 3D's classification superiority over PrimateAI. The performance results in Figures 28, 29, 30, and 31 are generated on the classification task of accurately distinguishing benign variants from pathogenic variants across a plurality of validation sets. PrimateAI 3D is trained on training sets that are different from the plurality of validation sets. PrimateAI 3D is trained on common human variants and variants from primates used as benign dataset while simulated variants based on trinucleotide context used as unlabeled or pseudo-pathogenic dataset.

[00172] New developmental delay disorder (new DDD) is one example of a validation set used to compare the classification accuracy of Primate AI 3D against Primate AI. The new DDD validation set labels variants from individuals with DDD as pathogenic and labels the same variants from healthy relatives of the individuals with the DDD as benign. A similar labelling scheme is used with an autism spectrum disorder (ASD) validation set shown in Figure 31.

[00173] BRCA1 is another example of a validation set used to compare the classification accuracy of Primate AI 3D against Primate AI. The BRCA1 validation set labels synthetically generated reference amino acid sequences simulating proteins of the BRCA1 gene as benign variants and labels synthetically altered allele amino acid sequences simulating proteins of the BRCA1 gene as pathogenic variants. A similar labelling scheme is used with different validation sets of the TP53 gene, TP53S3 gene and its variants, and other genes and their variants shown in Figure 31.

[00174] Figure 28 identifies performance of the benchmark PrimateAI model with horizontal bars (labeled as "PAI") and performance of the disclosed PrimateAI 3D model with horizontal bars (labeled as "ens10_7x7x7x2_hhpred_evo+alt"). Horizontal bars labeled as "ens10_7x7x7x2_hhpred_evo+alt_paisum" depict pathogenicity predictions derived by combining respective pathogenicity predictions of the disclosed PrimateAI 3D model and the benchmark PrimateAI model. In the legend, "ens10" denotes an ensemble of ten PrimateAI 3D models, each trained with a different seed training dataset and randomly initialized with different weights and biases. Also, "7x7x7x2" depicts the size of the voxel grid used to encode the input channels during the training of the ensemble of ten PrimateAI 3D models. For a given variant, the ensemble of ten PrimateAI 3D models respectively generates ten pathogenicity predictions, which are subsequently combined (*e.g.*, by averaging) to generate a final pathogenicity prediction for the given variant. This logic analogous applies to ensembles of different group sizes.

[00175] Also, in Figure 28, the y-axis has the different validation sets and the x-axis has p-values. Greater p-values, *i.e.*, longer horizontal bars denote greater accuracy in differentiating benign variants from pathogenic variants. As demonstrated by the p-values in Figure 28,

PrimateAI 3D outperforms PrimateAI across most of the validation sets (only exception being the tp53s3_A549 validation set). That is, the horizontal bars for PrimateAI 3D (labeled as “ens10_7x7x7x2_hhpred_evo+alt”) are consistently longer than the horizontal bars for PrimateAI (labeled as “PAI”).

[00176] Also, in Figure 28, a “mean” category along the y-axis calculates the mean of the p-values determined for each of the validation sets. In the mean category as well, PrimateAI 3D outperforms PrimateAI.

[00177] In Figure 29, PrimateAI is represented by horizontal bars (labeled as “PAI”), an ensemble of twenty PrimateAI 3D models trained with a voxel grid of size 3x3x3 is represented by horizontal bars labeled as “ns20_3x3x3x2_evo+alt”, an ensemble of ten PrimateAI 3D models trained with a voxel grid of size 7x7x7 is represented by horizontal bars labeled as “ensl0_7x7x7x2_evo+alt”, an ensemble of twenty PrimateAI 3D models trained with a voxel grid of size 7x7x7 is represented by horizontal bars labeled as “ens20_7x7x7x2_evo+alt”, and an ensemble of twenty PrimateAI 3D models trained with a voxel grid of size 17x17x17 is represented by horizontal bars labeled as “ens20_17x17x17x2_evo+alt”.

[00178] Also, in Figure 29, the y-axis has the different validation sets and the x-axis has p-values. As before, greater p-values, *i.e.*, longer horizontal bars denote greater accuracy in differentiating benign variants from pathogenic variants. As demonstrated by the p-values in Figure 20, different configurations of PrimateAI 3D outperform PrimateAI across most of the validation sets. That is, the horizontal bars representing ensembles of multiple PrimateAI 3D models (labeled as “ns20_3x3x3x2_evo+alt”, “ensl0_7x7x7x2_evo+alt”, “ens20_7x7x7x2_evo+alt” and “ens20_17x17x17x2_evo+alt”) are mostly longer than the horizontal bars for PrimateAI (labeled as “PAI”).

[00179] Also, in Figure 29, a “mean” category along the y-axis calculates the mean of the p-values determined for each of the validation sets. In the mean category as well, the different configurations of PrimateAI 3D outperform PrimateAI.

[00180] In Figure 30, the vertical bars with dark shades represent PrimateAI (“PrimateAI (v1)”), and the vertical bars with light shades represent PrimateAI 3D (“PrimateAI 3D”). In Figure 30, the y-axis has p-values, and the x-axis has the different validation sets. In Figure 30, without exceptions, PrimateAI 3D consistently outperforms PrimateAI across all of the validation sets. That is, the vertical bars for PrimateAI 3D are always longer than the vertical bars for PrimateAI.

[00181] Figure 31 identifies performance of the benchmark PrimateAI model with vertical bars labeled as “PAI-v1_plain” and performance of the disclosed PrimateAI 3D model with vertical bars labeled as “PAI-3D-orig_plain”. Vertical bars labeled as “PAI-3D-orig_paisum”

depict pathogenicity predictions derived by combining respective pathogenicity predictions of the disclosed PrimateAI 3D model and the benchmark PrimateAI model. In Figure 31, the y-axis has p-values, and the x-axis has the different validation sets.

[00182] As demonstrated by the p-values in Figure 31, PrimateAI 3D outperforms PrimateAI across most of the validation sets (only exception being the tp53s3_A549_p53NULL_Nutlin-3 validation set). That is, the vertical bars for PrimateAI 3D are consistently longer than the vertical bars for PrimateAI.

[00183] Also, in Figure 31, a separate “mean” chart calculates the mean of the p-values determined for each of the validation sets. In the mean chart as well, PrimateAI 3D outperforms PrimateAI.

[00184] The mean statistics may be biased by outliers. To address this, a separate “method ranks” chart is also depicted in Figure 31. Higher rank denotes poorer classification accuracy. In the method ranks chart as well, PrimateAI 3D outperforms PrimateAI by having more counts of lower ranks 1 and 2 versus Primate AI having all 3s.

[00185] In Figures 28 to 31, it is also evident that combining PrimateAI 3D with PrimateAI produces superior classification accuracy. That is, a protein can be fed as an amino acid sequence to PrimateAI to generate a first output, and the same protein can be fed as a 3D, voxelized protein structure to PrimateAI 3D to generate a second output, and the first and second outputs can be combined or analyzed in aggregate to produce a final pathogenicity prediction for a variant experienced by the protein.

Efficient Voxelization

[00186] Figure 32 is a flowchart illustrating an efficient voxelization process 3200 that efficiently identifies nearest atoms on a voxel-by-voxel basis.

[00187] The discussion now revisits the distance channels. As discussed above, the reference amino acid sequence 202 can contain different types of atoms, such as alpha-carbon atoms, beta-carbon atoms, oxygen atoms, nitrogen atoms, hydrogen atoms, and so on. Accordingly, as discussed above, the distance channels can be arranged by nearest alpha-carbon atoms, nearest beta-carbon atoms, nearest oxygen atoms, nearest nitrogen atoms, nearest hydrogen atoms, and so on. For example, in Figure 6, each of the nine voxels 514 has twenty-one amino acid-wise distance channels for nearest alpha-carbon atoms. Figure 6 can be further expanded for each of the nine voxels 514 to also have twenty-one amino acid-wise distance channels for nearest beta-carbon atoms, and for each of the nine voxels 514 to also have a nearest generic atom distance channel for a nearest atom irrespective of the type of the atom and the type of the amino acid. This way, each of the nine voxels 514 can have forty-three distance channels.

[00188] The discussion now turns to the number of distance calculations required to identify the nearest atoms on a voxel-by-voxel basis for inclusion in the distance channels. Consider the example in Figure 3 that depicts a total of eight hundred and twenty-eight alpha-carbon atoms distributed across the twenty-one amino acid categories. To calculate the amino acid-wise distance channels 602-642 in Figure 6, *i.e.*, to determine the one hundred and eighty-nine distance values, distances are measured from each of the nine voxels 514 to each of the eight hundred and twenty-eight alpha-carbon atoms, resulting in $9 * 828 = 7,452$ distance calculations. In the 3D case of twenty-seven voxels, this results in $27 * 828 = 22,356$ distance calculations. When the eight hundred and twenty-eight beta-carbon atoms are also included, this number increases to $27 * 1656 = 44,712$ distance calculations.

[00189] This means that the runtime complexity of identifying the nearest atoms on a voxel-by-voxel basis for a single protein voxelization is $O(\#atoms * \#voxels)$, as illustrated by Figure 35A. Furthermore, the runtime complexity for a single protein voxelization increases to $O(\#atoms * \#voxels * \#attributes)$ when the distance channels are calculated across a variety of attributes (*e.g.*, different features or channels per voxel like annotation channels and structural confidence channels).

[00190] Consequently, the distance calculations can become the most compute-consuming part of the voxelization process, taking valuable compute resources away from critical runtime tasks like model training and model inference. Consider, for example, the case of model training with a training dataset of 7,000 proteins. Generating distance channels for a plurality of voxels across a plurality of amino acids, atoms, and attributes can involve more than 100 voxelizations per protein, resulting in about 800,000 voxelizations in a single training iteration (epoch). A training run of 20-40 epochs, with rotation of atomic coordinates in each epoch, can result in as many as 32 million voxelizations.

[00191] In addition to the high compute cost, the size of the data for 32 million voxelizations is too big to fit in main memory (*e.g.*, >20TB for a 15x15x15 voxel grid). Considering repeated training runs for parameter optimization and ensemble learning, the memory footprint of the voxelization process gets too big to be stored on disk, making the voxelization process a part of the model training and not a precomputation step.

[00192] The technology disclosed provides an efficient voxelization process that achieves up to ~100x speedup over the runtime complexity of $O(\#atoms * \#voxels)$. The disclosed efficient voxelization process reduces the runtime complexity for a single protein voxelization to $O(\#atoms)$. In the case of different features or channels per voxel, the disclosed efficient voxelization process reduces the runtime complexity for a single protein voxelization to $O(\#atoms * \#attributes)$. As a result, the voxelization process becomes as fast as model training,

shifting the computational bottleneck from voxelization back to computing neural network weights on processors such as GPUs, ASICs, TPUs, FPGAs, CGRAs, *etc.*

[00193] In some implementations of the disclosed efficient voxelization process involving large voxel grids, the runtime complexity for a single protein voxelization is $O(\#atoms + voxels)$ and $O(\#atoms * \#attributes + voxels)$ for the case of different features or channels per voxel. The “+ voxels” complexity is observed when the number of atoms is minuscule compared to the number of voxels, for example, when there is one atom in a 100x100x100 voxel grid (*i.e.*, one million voxels per atom). In such a scenario, the runtime is dominated by the overhead of the huge number of voxels, for example, for allocating the memory for one million voxels, initialization one million voxels to zero, *etc.*

[00194] The discussion now turns to details of the disclosed efficient voxelization process. Figures 32A, 32B, 33, 34, and 35B are discussed in tandem.

[00195] Starting with Figure 32A, at step 3202, each atom (*e.g.*, each of the 828 alpha-carbon atoms and each of the 828 beta-carbon atoms) is associated with a voxel that contains the atom (*e.g.*, one of the nine voxels 514). The term “contains” refers to the 3D atomic coordinates of the atom being located in the voxel. The voxel that contains the atom is also referred to herein as “the atom-containing voxel.”

[00196] Figures 32B and 33 describe how a voxel that contains a particular atom is selected. Figure 33 uses 2D atomic coordinates as representative of 3D atomic coordinates. Note that the voxel grid 522 is regularly spaced with each of the voxels 514 having a same step size (*e.g.*, 1 angstrom (Å) or 2 Å).

[00197] Also, in Figure 33, the voxel grid 522 has indices [0, 1, 2] along a first dimension (*e.g.*, x-axis) and indices [0, 1, 2] along a second dimension (*e.g.*, y-axis). Also, in Figure 33, the respective voxels 514 in the voxel grid 522 are identified by voxel indices [Voxel 0, Voxel 1, ..., Voxel 8] and by voxel center indices [(1, 1), (1, 2), ..., (3, 3)].

[00198] Also, in Figure 33, center coordinates of the voxel centers along the first dimension, *i.e.*, first dimension voxel coordinates, are identified. Also, in Figure 33, center coordinates of the voxel centers along the second dimension, *i.e.*, second dimension voxel coordinates, are identified.

[00199] First, at step 3202a (Step 1 in Figure 33), 3D atomic coordinates (1.7456, 2.14323) of the particular atom are quantized to generated quantized 3D atomic coordinates (1.7, 2.1). The quantization can be achieved by rounding or truncation of bits.

[00200] Then, at step 3202b (Step 2 in Figure 33), voxel coordinates (or voxel centers or voxel center coordinates) of the voxels 514 are assigned to the quantized 3D atomic coordinates on a dimension-basis. For the first dimension, the quantized atomic coordinate 1.7 is assigned to

Voxel 1 because it covers first dimension voxel coordinates ranging from 1 to 2 and is centered at 1.5 in the first dimension. Note that Voxel 1 has index 1 along the first dimension, in contrast to having index 0 along the second dimension.

[00201] For the second dimension, starting from Voxel 1, the voxel grid 522 is traversed along the second dimension. This results in the quantized atomic coordinate 2.5 being assigned to Voxel 7 because it covers second dimension voxel coordinates ranging from 2 to 3 and is centered at 2.5 in the second dimension. Note that Voxel 7 has index 2 along the second dimension, in contrast to having index 1 along the first dimension.

[00202] Then, at step 3202c (Step 3 in Figure 33), dimension indices corresponding to the assigned voxel coordinates are selected. That is, for Voxel 1, index 1 is selected along the first dimension, and, for Voxel 7, index 2 is selected along the second dimension. A person skilled in the art will appreciate that the above steps can be analogously executed for a third dimension to select a dimension index along the third dimension.

[00203] Then, at step 3202d (Step 4 in Figure 33), an accumulated sum is generated based on position-wise weighting the selected dimension indices by powers of a radix. The general idea behind positional numbering systems is that a numeric value is represented through increasing powers of the radix (or base), for example, binary is base two, ternary is base three, octal is base eight, and hexadecimal is base sixteen. This is often referred to as a weighted numbering system because each position is weighted by a power of the radix. The set of valid numerals for a positional numbering system is equal in size to the radix of that system. For example, there are ten digits in the decimal system, zero through nine, and three digits in the ternary system, zero, one, and two. The largest valid number in a radix system is one smaller than the radix (so eight is not a valid numerical in any radix system smaller than nine). Any decimal integer can be expressed exactly in any other integral base system, and *vice-versa*.

[00204] Returning to the example in Figure 33, the selected dimension indices 1 and 2 are converted to a single integer by position-wise multiplying them with respective powers of base three and summing the results of the position-wise multiplications. Base three is selected here because the 3D atomic coordinates have three dimensions (although Figure 33 shows only 2D atomic coordinates along two dimensions for simplicity's sake).

[00205] Since index 2 is positioned at the rightmost bit (*i.e.*, the least significant bit), it is multiplied by three to the power of zero to yield two. Since index 1 is positioned at the second rightmost bit (*i.e.*, the second least significant bit), it is multiplied by three to the power of one to yield three. This results in the accumulated sum being five.

[00206] Then, at step 3202e (Step 5 in Figure 33), based on the accumulated sum, a voxel index of the voxel containing the particular atom is selected. That is, the accumulated sum is interpreted as the voxel index of the voxel containing the particular atom.

[00207] At step 3212, after each atom is associated with the atom-containing voxel, each atom is further associated with one or more voxels that are in a neighborhood of the atom-containing voxel, also referred to herein as “neighborhood voxels.” The neighborhood voxels can be selected based on being within a predefined radius of the atom-containing voxel (*e.g.*, 5 angstrom (Å)). In other implementations, the neighborhood voxels can be selected based on being contiguously adjacent to the atom-containing voxel (*e.g.*, top, bottom, right, left adjacent voxels). The resulting association that associates each atom with the atom-containing voxel and the neighborhood voxels is encoded in an atom-to-voxels mapping 3402, also referred to herein as element-to-cells mapping. In one example, a first alpha-carbon atom is associated with a first subset of voxels 3404 that includes an atom-containing voxel and neighborhood voxels for the first alpha-carbon atom. In another example, a second alpha-carbon atom is associated with a second subset of voxels 3406 that includes an atom-containing voxel and neighborhood voxels for the second alpha-carbon atom.

[00208] Note that no distance calculations are made to determine the atom-containing voxel and the neighborhood voxels. The atom-containing voxel is selected by virtue of the spatial arrangement of the voxels that allows assignment of quantized 3D atomic coordinates to corresponding regularly spaced voxel centers in the voxel grid (without using any distance calculations). Also, the neighborhood voxels are selected by virtue of being spatially contiguous to the atom-containing voxel in the voxel grid (again without using any distance calculations).

[00209] At step 3222, each voxel is mapped to atoms to which it was associated at steps 3202 and 3212. In one implementation, this mapping is encoded in a voxel-to-atoms mapping 3412, which is generated based on the atom-to-voxels mapping 3402 (*e.g.*, by applying a voxel-based sorting key on the atom-to-voxels mapping 3402). The voxel-to-atoms mapping 3412 is also referred to herein as “cell-to-elements mapping.” In one example, a first voxel is mapped to a first subset of alpha-carbon atoms 3414 that includes alpha-carbon atoms associated with the first voxel at steps 3202 and 3212. In another example, a second voxel is mapped to a second subset of alpha-carbon atoms 3416 that includes alpha-carbon atoms associated with the second voxel at steps 3202 and 3212.

[00210] At step 3232, for each voxel, distances are calculated between the voxel and atoms mapped to the voxel at step 3222. Step 3232 has a runtime complexity of $O(\#atoms)$ because distance to a particular atom is measured only once from a respective voxel to which the particular atom is uniquely mapped in the voxel-to-atoms mapping 3412. This is true when no

neighboring voxels are considered. Without neighbors, the constant factor that is implied in the big-O notation is 1. With neighbors, the big-O notation is equal to the number of neighbors + 1 since the number of neighbors is constant for each voxel, and therefore the runtime complexity of $O(\#atoms)$ remains true. In contrast, in Figure 35A, distances to a particular atom are redundantly measured as many times as the number of voxels (*e.g.*, 27 distances for a particular atom due to 27 voxels).

[00211] In Figure 35B, based on the voxel-to-atoms mapping 3412, each voxel is mapped to a respective subset of the 828 atoms (not including distance calculations to neighborhood voxels), as illustrated by respective ovals for respective voxels. The respective subsets are largely non-overlapping, with some exceptions. Insignificant overlap exists due to some instances when multiple atoms are mapped to a same voxel, as indicated in Figure 35B by the prime symbol “'” and the yellow overlap between the ovals. This minimal overlap has an additive effect on the runtime complexity of $O(\#atoms)$ and not a multiplicative effect. This overlap is a result of considering neighboring voxels, after determining the voxel that contains the atom. Without neighboring voxels, there can be no overlap, because an atom is only associated with one voxel. Considering neighbors, however, each neighbor could potentially be associated with the same atom (as long as there is no other atom of the same amino acid that is closer).

[00212] At step 3242, for each voxel, based on the distances calculated at step 3232, a nearest atom to the voxel is identified. In one implementation, this identification is encoded in a voxel-to-nearest atom mapping 3422, also referred to herein as “cell-to-nearest element mapping.” In one example, the first voxel is mapped to a second alpha-carbon atom as its nearest alpha-carbon atom 3424. In another example, the second voxel is mapped to a thirty-first alpha-carbon atom as its nearest alpha-carbon atom 3426.

[00213] Furthermore, as the voxel-wise distances are calculated using the technique discussed above, the atom-type and amino acid-type categorization of the atoms and the corresponding distance values are stored to generate categorized distance channels.

[00214] Once the distances to nearest atoms are identified using the technique discussed above, these distances can be encoded in the distance channels for voxelization and subsequent processing by the pathogenicity classifier 2108.

Computer System

[00215] Figure 36 shows an example computer system 3600 that can be used to implement the technology disclosed. Computer system 3600 includes at least one central processing unit (CPU) 3672 that communicates with a number of peripheral devices via bus subsystem 3655. These peripheral devices can include a storage subsystem 3610 including, for example, memory

devices and a file storage subsystem 3636, user interface input devices 3638, user interface output devices 3676, and a network interface subsystem 3674. The input and output devices allow user interaction with computer system 3600. Network interface subsystem 3674 provides an interface to outside networks, including an interface to corresponding interface devices in other computer systems.

[00216] In one implementation, the pathogenicity classifier 2108 is communicably linked to the storage subsystem 3610 and the user interface input devices 3638.

[00217] User interface input devices 3638 can include a keyboard; pointing devices such as a mouse, trackball, touchpad, or graphics tablet; a scanner; a touch screen incorporated into the display; audio input devices such as voice recognition systems and microphones; and other types of input devices. In general, use of the term “input device” is intended to include all possible types of devices and ways to input information into computer system 3600.

[00218] User interface output devices 3676 can include a display subsystem, a printer, a fax machine, or non-visual displays such as audio output devices. The display subsystem can include an LED display, a cathode ray tube (CRT), a flat-panel device such as a liquid crystal display (LCD), a projection device, or some other mechanism for creating a visible image. The display subsystem can also provide a non-visual display such as audio output devices. In general, use of the term “output device” is intended to include all possible types of devices and ways to output information from computer system 3600 to the user or to another machine or computer system.

[00219] Storage subsystem 3610 stores programming and data constructs that provide the functionality of some or all of the modules and methods described herein. These software modules are generally executed by processors 3678.

[00220] Processors 3678 can be graphics processing units (GPUs), field-programmable gate arrays (FPGAs), application-specific integrated circuits (ASICs), and/or coarse-grained reconfigurable architectures (CGRAs). Processors 3678 can be hosted by a deep learning cloud platform such as Google Cloud Platform™, Xilinx™, and Cirrascale™. Examples of processors 3678 include Google’s Tensor Processing Unit (TPU)™, rackmount solutions like GX4 Rackmount Series™, GX36 Rackmount Series™, NVIDIA DGX-1™, Microsoft’ Stratix V FPGA™, Graphcore’s Intelligent Processor Unit (IPU)™, Qualcomm’s Zeroth Platform™ with Snapdragon processors™, NVIDIA’s Volta™, NVIDIA’s DRIVE PX™, NVIDIA’s JETSON TX1/TX2 MODULE™, Intel’s Nirvana™, Movidius VPU™, Fujitsu DPI™, ARM’s DynamicIQ™, IBM TrueNorth™, Lambda GPU Server with Testa V100s™, and others.

[00221] Memory subsystem 3622 used in the storage subsystem 3610 can include a number of memories including a main random access memory (RAM) 3632 for storage of instructions and data during program execution and a read only memory (ROM) 3634 in which fixed instructions

are stored. A file storage subsystem 3636 can provide persistent storage for program and data files, and can include a hard disk drive, a floppy disk drive along with associated removable media, a CD-ROM drive, an optical drive, or removable media cartridges. The modules implementing the functionality of certain implementations can be stored by file storage subsystem 3636 in the storage subsystem 3610, or in other machines accessible by the processor.

[00222] Bus subsystem 3655 provides a mechanism for letting the various components and subsystems of computer system 3600 communicate with each other as intended. Although bus subsystem 3655 is shown schematically as a single bus, alternative implementations of the bus subsystem can use multiple busses.

[00223] Computer system 3600 itself can be of varying types including a personal computer, a portable computer, a workstation, a computer terminal, a network computer, a television, a mainframe, a server farm, a widely-distributed set of loosely networked computers, or any other data processing system or user device. Due to the ever-changing nature of computers and networks, the description of computer system 3600 depicted in Figure 36 is intended only as a specific example for purposes of illustrating the preferred implementations of the present invention. Many other configurations of computer system 3600 are possible having more or less components than the computer system depicted in Figure 36.

Particular Implementations 1

[00224] The following implementations can be practiced as a system, method, or article of manufacture. One or more features of an implementation can be combined with the base implementation. Implementations that are not mutually exclusive are taught to be combinable. One or more features of an implementation can be combined with other implementations. This disclosure periodically reminds the user of these options. Omission from some implementations of recitations that repeat these options should not be taken as limiting the combinations taught in the preceding sections – these recitations are hereby incorporated forward by reference into each of the following implementations.

[00225] Though the technology disclosed uses 3D data as input, in other implementations, it can analogously use 1D data, 2D data (*e.g.*, pixels and 2D atomic coordinates), 4D data, 5D data, and so on.

[00226] In some implementations, a system comprises memory storing amino acid-wise distance channels for a plurality of amino acids in a protein. Each of the amino acid-wise distance channels has voxel-wise distance values for voxels in a plurality of voxels. The voxel-wise distance values specify distances from corresponding voxels in the plurality of voxels to atoms of corresponding amino acids in the plurality of amino acids. The system further

comprises a pathogenicity determination engine configured to process a tensor that includes the amino acid-wise distance channels and an alternative allele of the protein expressed by a variant. The pathogenicity determination engine can also be configured to determine a pathogenicity of the variant based at least in part on the tensor.

[00227] In some implementations, the system further comprises a distance channels generator that centers a voxel grid of the voxels on an alpha-carbon atom of respective residues of the amino acids. The distance channels generator can center the voxel grid on an alpha-carbon atom of a residue of a particular amino acid that positioned at a variant amino acid in the protein.

[00228] The system can be configured to encode, in the tensor, a directionality of the amino acids and a position of the particular amino acid by multiplying, with a directionality parameter, voxel-wise distance values for those amino acids that precede the particular amino acid. The distances can be nearest-atom distances from corresponding voxel centers in the voxel grid to nearest atoms of the corresponding amino acids. In some implementations, the nearest-atom distances can be Euclidean distances. The nearest-atom distances can be normalized by dividing the Euclidean distances with a maximum nearest-atom distance. The amino acids can have alpha-carbon atoms and, in some implementations, the distances can be nearest-alpha-carbon atom distances from the corresponding voxel centers to nearest alpha-carbon atoms of the corresponding amino acids. The amino acids can have beta-carbon atoms and, in some implementations, the distances can be nearest-beta-carbon atom distances from the corresponding voxel centers to nearest beta-carbon atoms of the corresponding amino acids. The amino acids can have backbone atoms and, in some implementations, the distances can be nearest-backbone atom distances from the corresponding voxel centers to nearest backbone atoms of the corresponding amino acids. The amino acids have side chain atoms and, in some implementations, the distances can be nearest-sidechain atom distances from the corresponding voxel centers to nearest sidechain atoms of the corresponding amino acids.

[00229] The system can further be configured to encode, in the tensor, a nearest atom channel that specifies a distance from each voxel to a nearest atom. The nearest atom can be selected irrespective of the amino acids and atomic elements of the amino acids. In some implementations, the distance is a Euclidean distance. The distance can be normalized by dividing the Euclidean distance with a maximum distance. The amino acids can include non-standard amino acids. The tensor can include an absentee atom channel that specifies atoms not found within a predefined radius of a voxel center, and the absentee atom channel can be one-hot encoded. In some implementations, the tensor can further include a one-hot encoding of the alternative allele that is voxel-wise encoded to each of the amino acid-wise distance channels. The tensor can further include a reference allele of the protein. In some implementations, the

tensor can further include a one-hot encoding of the reference allele that is voxel-wise encoded to each of the amino acid-wise distance channels. The tensor can further include evolutionary profiles that specify conservation levels of the amino acids across a plurality of species.

[00230] The system can further comprise an evolutionary profiles generator that, for each of the voxels, selects a nearest atom across the amino acids and the atom categories, selects a pan-amino acid conservation frequencies sequence for a residue of an amino acid that includes the nearest atom, and makes the pan-amino acid conservation frequencies sequence available as one of the evolutionary profiles. The pan-amino acid conservation frequencies sequence can be configured for a particular position of the residue as observed in the plurality of species. The pan-amino acid conservation frequencies sequence can specify whether there is a missing conservation frequency for a particular amino acid. In some implementations, the evolutionary profiles generator, for each of the voxels, can select respective nearest atoms in respective ones of the amino acids, can select respective per-amino acid conservation frequencies for respective residues of the amino acids that include the nearest atoms, and can make the per-amino acid conservation frequencies available as one of the evolutionary profiles. The per-amino acid conservation frequencies can be configured for a particular position of the residues as observed in the plurality of species. The per-amino acid conservation frequencies can specify whether there is a missing conservation frequency for a particular amino acid.

[00231] In some implementations of the system, the tensor can further include annotation channels for the amino acids. The annotation channels can be one-hot encoded in the tensor. The annotation channels can be molecular processing annotations that include initiator methionine, signal, transit peptide, propeptide, chain, and peptide. The annotation channels can be regions annotations that include topological domain, transmembrane, intramembrane, domain, repeat, calcium binding, zinc finger, deoxyribonucleic acid (DNA) binding, nucleotide binding, region, coiled coil, motif, and compositional bias. The annotation channels can be sites annotations that include active site, metal binding, binding site, and site. The annotation channels can be amino acid modifications annotations that include non-standard residue, modified residue, lipidation, glycosylation, disulfide bond, and cross-link. The annotation channels can be secondary structure annotations that include helix, turn, and beta strand. The annotation channels can be experimental information annotations that include mutagenesis, sequence uncertainty, sequence conflict, non-adjacent residues, and non-terminal residue.

[00232] In some implementations of the system, the tensor further includes structure confidence channels for the amino acids that specify quality of respective structures of the amino acids. The structure confidence channels can be global model quality estimations (GMQEs). The structure confidence channels can include qualitative model energy analysis (QMEAN) scores.

The structure confidence channels can be temperature factors that specify a degree to which the residues satisfy physical constraints of respective protein structures. The structure confidence channels can be template structures alignments that specify a degree to which residues of atoms nearest to the voxels have aligned template structures. The structure confidence channels can be template modeling scores of the aligned template structures. The structure confidence channels can be a minimum one of the template modeling scores, a mean of the template modeling scores, and a maximum one of the template modeling scores.

[00233] In some implementations, the system can further comprise a tensor generator that voxel-wise concatenates amino acid-wise distance channels for the alpha-carbon atoms with the one-hot encoding of the alternative allele to generate the tensor. The tensor generator can voxel-wise concatenate amino acid-wise distance channels for the beta-carbon atoms with the one-hot encoding of the alternative allele to generate the tensor. The tensor generator can voxel-wise concatenate the amino acid-wise distance channels for the alpha-carbon atoms, the amino acid-wise distance channels for the beta-carbon atoms, and the one-hot encoding of the alternative allele to generate the tensor. The tensor generator can voxel-wise concatenate the amino acid-wise distance channels for the alpha-carbon atoms, the amino acid-wise distance channels for the beta-carbon atoms, the one-hot encoding of the alternative allele, and pan-amino acid conservation frequencies to generate the tensor. The tensor generator can voxel-wise concatenate the amino acid-wise distance channels for the alpha-carbon atoms, the amino acid-wise distance channels for the beta-carbon atoms, the one-hot encoding of the alternative allele, the pan-amino acid conservation frequencies, and the annotation channels to generate the tensor. The tensor generator can voxel-wise concatenate the amino acid-wise distance channels for the alpha-carbon atoms, the amino acid-wise distance channels for the beta-carbon atoms, the one-hot encoding of the alternative allele, the pan-amino acid conservation frequencies, the annotation channels, and the structure confidence channels to generate the tensor. The tensor generator can voxel-wise concatenate the amino acid-wise distance channels for the alpha-carbon atoms, the amino acid-wise distance channels for the beta-carbon atoms, the one-hot encoding of the alternative allele, and per-amino acid conservation frequencies for each of the amino acids to generate the tensor. The tensor generator can voxel-wise concatenate the amino acid-wise distance channels for the alpha-carbon atoms, the amino acid-wise distance channels for the beta-carbon atoms, the one-hot encoding of the alternative allele, per-amino acid conservation frequencies for each of the amino acids, and the annotation channels to generate the tensor. The tensor generator can voxel-wise concatenate the amino acid-wise distance channels for the alpha-carbon atoms, the amino acid-wise distance channels for the beta-carbon atoms, the one-hot encoding of the alternative allele, per-amino acid conservation frequencies for each of the amino acids, the annotation

channels, and the structure confidence channels to generate the tensor. The tensor generator can voxel-wise concatenate the amino acid-wise distance channels for the alpha-carbon atoms, the amino acid-wise distance channels for the beta-carbon atoms, the one-hot encoding of the alternative allele, and the one-hot encoding of the reference allele to generate the tensor. The tensor generator can voxel-wise concatenate the amino acid-wise distance channels for the alpha-carbon atoms, the amino acid-wise distance channels for the beta-carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, and the pan-amino acid conservation frequencies to generate the tensor. The tensor generator can voxel-wise concatenate the amino acid-wise distance channels for the alpha-carbon atoms, the amino acid-wise distance channels for the beta-carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, the pan-amino acid conservation frequencies, and the annotation channels to generate the tensor. The tensor generator can voxel-wise concatenate the amino acid-wise distance channels for the alpha-carbon atoms, the amino acid-wise distance channels for the beta-carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, the pan-amino acid conservation frequencies, the annotation channels, and the structure confidence channels to generate the tensor. The tensor generator can voxel-wise concatenate the amino acid-wise distance channels for the alpha-carbon atoms, the amino acid-wise distance channels for the beta-carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, and the per-amino acid conservation frequencies for each of the amino acids to generate the tensor. The tensor generator can voxel-wise concatenate the amino acid-wise distance channels for the alpha-carbon atoms, the amino acid-wise distance channels for the beta-carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, the per-amino acid conservation frequencies for each of the amino acids, and the annotation channels to generate the tensor. The tensor generator can voxel-wise concatenate the amino acid-wise distance channels for the alpha-carbon atoms, the amino acid-wise distance channels for the beta-carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, the per-amino acid conservation frequencies for each of the amino acids, the annotation channels, and the structure confidence channels to generate the tensor.

[00234] In some implementations, the system can further comprise an atoms rotation engine that rotates atoms of the amino acids before the amino acid-wise distance channels are generated. The pathogenicity determination engine can be a neural network. In particular implementations, the pathogenicity determination engine can be a convolutional neural network. The convolutional neural network can use $1 \times 1 \times 1$ convolutions, $3 \times 3 \times 3$ convolutions, rectified linear unit activation layers, batch normalization layers, a fully-connected layer, a dropout regularization

layer, and a softmax classification layer. The $1 \times 1 \times 1$ convolutions and the $3 \times 3 \times 3$ convolutions can be three-dimensional convolutions.

[00235] In some implementations, a layer of the $1 \times 1 \times 1$ convolutions can process the tensor and produce an intermediate output that is a convolved representation of the tensor. A sequence of layers of the $3 \times 3 \times 3$ convolutions can process the intermediate output and produce a flattened output. The fully-connected layer can process the flattened output and produce unnormalized outputs. The softmax classification layer can process the unnormalized outputs and produce exponentially normalized outputs that identify likelihoods of the variant being pathogenic and benign. A sigmoid layer can process the unnormalized outputs and produce a normalized output that identifies a likelihood of the variant being pathogenic. The voxels, the atoms, and the distances can have three-dimensional coordinates. The tensor can have at least three dimensions, the intermediate output can have at least three dimensions, and the flattened output can have one dimension.

[00236] In some implementations, the pathogenicity determination engine is a recurrent neural network. In other implementations, the pathogenicity determination engine is an attention-based neural network. In still other implementations, the pathogenicity determination engine is a gradient-boosted tree. In still other implementations, the pathogenicity determination engine is a state vector machine.

[00237] In other implementations, a system can comprise memory storing atom category-wise distance channels for amino acids in a protein. The amino acids can have atoms for a plurality of atom categories, and atom categories in the plurality of atom categories can specify atomic elements of the amino acids. The atom category-wise distance channels can have voxel-wise distance values for voxels in a plurality of voxels. The voxel-wise distance values can specify distances from corresponding voxels in the plurality of voxels to atoms in corresponding atom categories in the plurality of atom categories. The system can further comprise a pathogenicity determination engine configured to process a tensor that includes the atom category-wise distance channels and an alternative allele of the protein expressed by a variant, and to determine a pathogenicity of the variant based at least in part on the tensor.

[00238] The system can further comprise a distance channels generator that centers a voxel grid of the voxels on respective atoms of respective atom categories in the plurality of atom categories. The distance channels generator can center the voxel grid on an alpha-carbon atom of a residue of at least one variant amino acid in the protein. The distances can be nearest-atom distances from corresponding voxel centers in the voxel grid to nearest atoms in the corresponding atom categories. The nearest-atom distances can be Euclidean distances. The nearest-atom distances can be normalized by dividing the Euclidean distances with a maximum

nearest-atom distances. The distances can be nearest-atom distances from the corresponding voxel centers in the voxel grid to nearest atoms irrespective of the amino acids and the atom categories of the amino acids. The nearest-atom distances can be Euclidean distances. The nearest-atom distances can be normalized by dividing the Euclidean distances with a maximum nearest-atom distances.

[00239] Other implementations of the method described in this section can include a non-transitory computer readable storage medium storing instructions executable by a processor to perform any of the methods described above. Yet another implementation of the method described in this section can include a system including memory and one or more processors operable to execute instructions, stored in the memory, to perform any of the methods described above.

Clauses 1

1. A computer-implemented method, comprising:
storing amino acid-wise distance channels for a plurality of amino acids in a protein,
wherein each of the amino acid-wise distance channels has voxel-wise distance values for voxels
in a plurality of voxels, and
wherein the voxel-wise distance values specify distances from corresponding voxels in the
plurality of voxels to atoms of corresponding amino acids in the plurality of amino acids;
processing a tensor that includes the amino acid-wise distance channels and an alternative allele
of the protein expressed by a variant; and
determining a pathogenicity of the variant based at least in part on the tensor.
2. The computer-implemented method of claim 1, further comprising centering a voxel grid
of the voxels on an alpha carbon atom of respective residues of the amino acids.
3. The computer-implemented method of claim 2, further comprising centering the voxel
grid on an alpha carbon atom of a residue of a particular amino acid that corresponds to at least
one variant amino acid in the protein.
4. The computer-implemented method of claim 3, further comprising encoding, in the
tensor, a directionality of the amino acids and a position of the particular amino acid by
multiplying, with a directionality parameter, voxel-wise distance values for those amino acids
that precede the particular amino acid.

5. The computer-implemented method of claim 3, wherein the distances are nearest-atom distances from corresponding voxel centers in the voxel grid to nearest atoms of the corresponding amino acids.
6. The computer-implemented method of claim 5, wherein the nearest-atom distances are Euclidean distances.
7. The computer-implemented method of claim 6, wherein the nearest-atom distances are normalized by dividing the Euclidean distances with a maximum nearest-atom distance.
8. The computer-implemented method of claim 5, wherein the amino acids have alpha carbon atoms, and wherein the distances are nearest-alpha carbon atom distances from the corresponding voxel centers to nearest alpha carbon atoms of the corresponding amino acids.
9. The computer-implemented method of claim 5, wherein the amino acids have beta carbon atoms and wherein the distances are nearest-beta carbon atom distances from the corresponding voxel centers to nearest beta carbon atoms of the corresponding amino acids.
10. The computer-implemented method of claim 5, wherein the amino acids have backbone atoms and wherein the distances are nearest-backbone atom distances from the corresponding voxel centers to nearest backbone atoms of the corresponding amino acids.
11. The computer-implemented method of claim 5, wherein the amino acids have sidechain atom and wherein the distances are nearest-sidechain atom distances from the corresponding voxel centers to nearest sidechain atoms of the corresponding amino acids.
12. The computer-implemented method of claim 3, further comprising encoding, in the tensor, a nearest atom channel that specifies a distance from each voxel to a nearest atom, wherein the nearest atom is selected irrespective of the amino acids and atomic elements of the amino acids.
13. The computer-implemented method of claim 12, wherein the distance is a Euclidean distance.
14. The computer-implemented method of claim 13, wherein the distance is normalized by dividing the Euclidean distance with a maximum distance.
15. The computer-implemented method of claim 12, wherein the amino acids include non-standard amino acids.

16. The computer-implemented method of claim 1, wherein the tensor further includes an absentee atom channel that specifies atoms not found within a predefined radius of a voxel center, and wherein the absentee atom channel is one-hot encoded.
17. The computer-implemented method of claim 1, wherein the tensor further includes a one-hot encoding of the alternative allele that is voxel-wise encoded to each of the amino acid-wise distance channels.
18. The computer-implemented method of claim 1, wherein the tensor further includes a reference allele of the protein.
19. The computer-implemented method of claim 18, wherein the tensor further includes a one-hot encoding of the reference allele that is voxel-wise encoded to each of the amino acid-wise distance channels.
20. The computer-implemented method of claim 1, wherein the tensor further includes evolutionary profiles that specify conservation levels of the amino acids across a plurality of species.
21. The computer-implemented method of claim 20, further comprising, for each of the voxels,
selecting a nearest atom across the amino acids and the atom categories,
selecting a pan-amino acid conservation frequencies sequence for a residue of an amino acid that includes the nearest atom, and
making the pan-amino acid conservation frequencies sequence available as one of the evolutionary profiles.
22. The computer-implemented method of claim 21, wherein the pan-amino acid conservation frequencies sequence is configured for a particular position of the residue as observed in the plurality of species.
23. The computer-implemented method of claim 21, wherein the pan-amino acid conservation frequencies sequence specifies whether there is a missing conservation frequency for a particular amino acid.
24. The computer-implemented method of claim 21, further comprising, for each of the voxels,
selecting respective nearest atoms in respective ones of the amino acids,

selecting respective per-amino acid conservation frequencies for respective residues of the amino acids that include the nearest atoms, and making the per-amino acid conservation frequencies available as one of the evolutionary profiles.

25. The computer-implemented method of claim 24, wherein the per-amino acid conservation frequencies are configured for a particular position of the residues as observed in the plurality of species.

26. The computer-implemented method of claim 24, wherein the per-amino acid conservation frequencies specify whether there is a missing conservation frequency for a particular amino acid.

27. The computer-implemented method of claim 1, wherein the tensor further includes annotation channels for the amino acids, wherein the annotation channels are one-hot encoded in the tensor.

28. The computer-implemented method of claim 27, wherein the annotation channels are molecular processing annotations that include initiator methionine, signal, transit peptide, propeptide, chain, and peptide.

29. The computer-implemented method of claim 27, wherein the annotation channels are regions annotations that include topological domain, transmembrane, intramembrane, domain, repeat, calcium binding, zinc finger, deoxyribonucleic acid (DNA) binding, nucleotide binding, region, coiled coil, motif, and compositional bias.

30. The computer-implemented method of claim 27, wherein the annotation channels are sites annotations that include active site, metal binding, binding site, and site.

31. The computer-implemented method of claim 27, wherein the annotation channels are amino acid modifications annotations that include non-standard residue, modified residue, lipidation, glycosylation, disulfide bond, and cross-link.

32. The computer-implemented method of claim 27, wherein the annotation channels are secondary structure annotations that include helix, turn, and beta strand.

33. The computer-implemented method of claim 27, wherein the annotation channels are experimental information annotations that include mutagenesis, sequence uncertainty, sequence conflict, non-adjacent residues, and non-terminal residue.

34. The computer-implemented method of claim 1, wherein the tensor further includes structure confidence channels for the amino acids that specify quality of respective structures of the amino acids.
35. The computer-implemented method of claim 34, wherein the structure confidence channels are global model quality estimations (GMQEs).
36. The computer-implemented method of claim 34, wherein the structure confidence channels include qualitative model energy analysis (QMEAN) scores.
37. The computer-implemented method of claim 34, wherein the structure confidence channels are temperature factors that specify a degree to which the residues satisfy physical constraints of respective protein structures.
38. The computer-implemented method of claim 34, wherein the structure confidence channels are template structures alignments that specify a degree to which residues of atoms nearest to the voxels have aligned template structures.
39. The computer-implemented method of claim 38, wherein the structure confidence channels are template modeling scores of the aligned template structures.
40. The computer-implemented method of claim 39, wherein the structure confidence channels are a minimum one of the template modeling scores, a mean of the template modeling scores, and a maximum one of the template modeling scores.
41. The computer-implemented method of claim 1, further comprising voxel-wise concatenating amino acid-wise distance channels for the alpha carbon atoms with the one-hot encoding of the alternative allele to generate the tensor.
42. The computer-implemented method of claim 41, further comprising voxel-wise concatenating amino acid-wise distance channels for the beta carbon atoms with the one-hot encoding of the alternative allele to generate the tensor.
43. The computer-implemented method of claim 42, further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, and the one-hot encoding of the alternative allele to generate the tensor.

44. The computer-implemented method of claim 43, further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, and pan-amino acid conservation frequencies sequences to generate the tensor.

45. The computer-implemented method of claim 44, further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, the pan-amino acid conservation frequencies sequences, and the annotation channels to generate the tensor.

46. The computer-implemented method of claim 45, further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, the pan-amino acid conservation frequencies sequences, the annotation channels, and the structure confidence channels to generate the tensor.

47. The computer-implemented method of claim 46, further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, and per-amino acid conservation frequencies for each of the amino acids to generate the tensor.

48. The computer-implemented method of claim 47, further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, per-amino acid conservation frequencies for each of the amino acids, and the annotation channels to generate the tensor.

49. The computer-implemented method of claim 48, further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, per-amino acid conservation frequencies for each of the amino acids, the annotation channels, and the structure confidence channels to generate the tensor.

50. The computer-implemented method of claim 49, further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, and the one-hot encoding of the reference allele to generate the tensor.

51. The computer-implemented method of claim 50, further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, and the pan-amino acid conservation frequencies sequences to generate the tensor.
52. The computer-implemented method of claim 51, further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, the pan-amino acid conservation frequencies sequences, and the annotation channels to generate the tensor.
53. The computer-implemented method of claim 52, further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, the pan-amino acid conservation frequencies sequences, the annotation channels, and the structure confidence channels to generate the tensor.
54. The computer-implemented method of claim 53, further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, and the per-amino acid conservation frequencies for each of the amino acids to generate the tensor.
55. The computer-implemented method of claim 54, further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, the per-amino acid conservation frequencies for each of the amino acids, and the annotation channels to generate the tensor.
56. The computer-implemented method of claim 55, further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, the per-amino acid conservation frequencies for each of the amino acids, the annotation channels, and the structure confidence channels to generate the tensor.

57. The computer-implemented method of claim 1, further comprising rotating atoms of the amino acids before the amino acid-wise distance channels are generated.
58. The computer-implemented method of claim 1, further comprising using $1 \times 1 \times 1$ convolutions, $3 \times 3 \times 3$ convolutions, rectified linear unit activation layers, batch normalization layers, a fully-connected layer, a dropout regularization layer, and a softmax classification layer in a convolutional neural network.
59. The computer-implemented method of claim 58, wherein the $1 \times 1 \times 1$ convolutions and the $3 \times 3 \times 3$ convolutions are three-dimensional convolutions.
60. The computer-implemented method of claim 58, wherein a layer of the $1 \times 1 \times 1$ convolutions processes the tensor and produces an intermediate output that is a convolved representation of the tensor, wherein a sequence of layers of the $3 \times 3 \times 3$ convolutions processes the intermediate output and produces a flattened output, wherein the fully-connected layer processes the flattened output and produces unnormalized outputs, and wherein the softmax classification layer processes the unnormalized outputs and produces exponentially normalized outputs that identify likelihoods of the variant being pathogenic and benign.
61. The computer-implemented method of claim 60, wherein a sigmoid layer processes the unnormalized outputs and produces a normalized output that identifies a likelihood of the variant being pathogenic.
62. The computer-implemented method of claim 60, wherein the voxels, the atoms, and the distances have three-dimensional coordinates, wherein the tensor has at least three dimensions, wherein the intermediate output has at least three dimensions, and wherein the flattened output has one dimension.
63. A computer-implemented method, comprising:
storing atom category-wise distance channels for amino acids in a protein,
wherein the amino acids have atoms for a plurality of atom categories,
wherein atom categories in the plurality of atom categories specify atomic elements of the amino acids,
wherein each of the atom category-wise distance channels has voxel-wise distance values for voxels in a plurality of voxels, and

wherein the voxel-wise distance values specify distances from corresponding voxels in the plurality of voxels to atoms in corresponding atom categories in the plurality of atom categories;

processing a tensor that includes the atom category-wise distance channels and an alternative allele of the protein expressed by a variant; and
determining a pathogenicity of the variant based at least in part on the tensor.

64. The computer-implemented method of claim 63, further comprising centering a voxel grid of the voxels on respective atoms of respective atom categories in the plurality of atom categories.

65. The computer-implemented method of claim 64, further comprising centering the voxel grid on an alpha carbon atom of a residue of at least one variant amino acid in the protein.

66. The computer-implemented method of claim 65, wherein the distances are nearest-atom distances from corresponding voxel centers in the voxel grid to nearest atoms in the corresponding atom categories.

67. The computer-implemented method of claim 66, wherein the nearest-atom distances are Euclidean distances.

68. The computer-implemented method of claim 67, wherein the nearest-atom distances are normalized by dividing the Euclidean distances with a maximum nearest-atom distances.

69. The computer-implemented method of claim 68, wherein the distances are nearest-atom distances from the corresponding voxel centers in the voxel grid to nearest atoms irrespective of the amino acids and the atom categories of the amino acids.

70. The computer-implemented method of claim 69, wherein the nearest-atom distances are Euclidean distances.

71. The computer-implemented method of claim 70, wherein the nearest-atom distances are normalized by dividing the Euclidean distances with a maximum nearest-atom distances.

1. One or more computer-readable media storing computer-executable instructions that, when executed on one or more processors, configure a computer to perform operations comprising:

storing amino acid-wise distance channels for a plurality of amino acids in a protein,

wherein each of the amino acid-wise distance channels has voxel-wise distance values for voxels in a plurality of voxels, and
wherein the voxel-wise distance values specify distances from corresponding voxels in the plurality of voxels to atoms of corresponding amino acids in the plurality of amino acids;
processing a tensor that includes the amino acid-wise distance channels and an alternative allele of the protein expressed by a variant; and
determining a pathogenicity of the variant based at least in part on the tensor.

2. The computer-readable media of claim 1, the operations further comprising centering a voxel grid of the voxels on an alpha carbon atom of respective residues of the amino acids.
3. The computer-readable media of claim 2, the operations further comprising centering the voxel grid on an alpha carbon atom of a residue of a particular amino acid that corresponds to at least one variant amino acid in the protein.
4. The computer-readable media of claim 3, the operations further comprising encoding, in the tensor, a directionality of the amino acids and a position of the particular amino acid by multiplying, with a directionality parameter, voxel-wise distance values for those amino acids that precede the particular amino acid.
5. The computer-readable media of claim 3, wherein the distances are nearest-atom distances from corresponding voxel centers in the voxel grid to nearest atoms of the corresponding amino acids.
6. The computer-readable media of claim 5, wherein the nearest-atom distances are Euclidean distances.
7. The computer-readable media of claim 6, wherein the nearest-atom distances are normalized by dividing the Euclidean distances with a maximum nearest-atom distance.
8. The computer-readable media of claim 5, wherein the amino acids have alpha carbon atoms, and wherein the distances are nearest-alpha carbon atom distances from the corresponding voxel centers to nearest alpha carbon atoms of the corresponding amino acids.
9. The computer-readable media of claim 5, wherein the amino acids have beta carbon atoms and wherein the distances are nearest-beta carbon atom distances from the corresponding voxel centers to nearest beta carbon atoms of the corresponding amino acids.

10. The computer-readable media of claim 5, wherein the amino acids have backbone atoms and wherein the distances are nearest-backbone atom distances from the corresponding voxel centers to nearest backbone atoms of the corresponding amino acids.
11. The computer-readable media of claim 5, wherein the amino acids have sidechain atom and wherein the distances are nearest-sidechain atom distances from the corresponding voxel centers to nearest sidechain atoms of the corresponding amino acids.
12. The computer-readable media of claim 3, the operations further comprising encoding, in the tensor, a nearest atom channel that specifies a distance from each voxel to a nearest atom, wherein the nearest atom is selected irrespective of the amino acids and atomic elements of the amino acids.
13. The computer-readable media of claim 12, wherein the distance is a Euclidean distance.
14. The computer-readable media of claim 13, wherein the distance is normalized by dividing the Euclidean distance with a maximum distance.
15. The computer-readable media of claim 12, wherein the amino acids include non-standard amino acids.
16. The computer-readable media of claim 1, wherein the tensor further includes an absentee atom channel that specifies atoms not found within a predefined radius of a voxel center, and wherein the absentee atom channel is one-hot encoded.
17. The computer-readable media of claim 1, wherein the tensor further includes a one-hot encoding of the alternative allele that is voxel-wise encoded to each of the amino acid-wise distance channels.
18. The computer-readable media of claim 1, wherein the tensor further includes a reference allele of the protein.
19. The computer-readable media of claim 18, wherein the tensor further includes a one-hot encoding of the reference allele that is voxel-wise encoded to each of the amino acid-wise distance channels.
20. The computer-readable media of claim 1, wherein the tensor further includes evolutionary profiles that specify conservation levels of the amino acids across a plurality of species.

21. The computer-readable media of claim 20, the operations further comprising, for each of the voxels,
- selecting a nearest atom across the amino acids and the atom categories,
- selecting a pan-amino acid conservation frequencies sequence for a residue of an amino acid that includes the nearest atom, and
- making the pan-amino acid conservation frequencies sequence available as one of the evolutionary profiles.
22. The computer-readable media of claim 21, wherein the pan-amino acid conservation frequencies sequence is configured for a particular position of the residue as observed in the plurality of species.
23. The computer-readable media of claim 21, wherein the pan-amino acid conservation frequencies sequence specifies whether there is a missing conservation frequency for a particular amino acid.
24. The computer-readable media of claim 21, the operations further comprising, for each of the voxels,
- selecting respective nearest atoms in respective ones of the amino acids,
- selecting respective per-amino acid conservation frequencies for respective residues of the amino acids that include the nearest atoms, and
- making the per-amino acid conservation frequencies available as one of the evolutionary profiles.
25. The computer-readable media of claim 24, wherein the per-amino acid conservation frequencies are configured for a particular position of the residues as observed in the plurality of species.
26. The computer-readable media of claim 24, wherein the per-amino acid conservation frequencies specify whether there is a missing conservation frequency for a particular amino acid.
27. The computer-readable media of claim 1, wherein the tensor further includes annotation channels for the amino acids, wherein the annotation channels are one-hot encoded in the tensor.

28. The computer-readable media of claim 27, wherein the annotation channels are molecular processing annotations that include initiator methionine, signal, transit peptide, propeptide, chain, and peptide.
29. The computer-readable media of claim 27, wherein the annotation channels are regions annotations that include topological domain, transmembrane, intramembrane, domain, repeat, calcium binding, zinc finger, deoxyribonucleic acid (DNA) binding, nucleotide binding, region, coiled coil, motif, and compositional bias.
30. The computer-readable media of claim 27, wherein the annotation channels are sites annotations that include active site, metal binding, binding site, and site.
31. The computer-readable media of claim 27, wherein the annotation channels are amino acid modifications annotations that include non-standard residue, modified residue, lipidation, glycosylation, disulfide bond, and cross-link.
32. The computer-readable media of claim 27, wherein the annotation channels are secondary structure annotations that include helix, turn, and beta strand.
33. The computer-readable media of claim 27, wherein the annotation channels are experimental information annotations that include mutagenesis, sequence uncertainty, sequence conflict, non-adjacent residues, and non-terminal residue.
34. The computer-readable media of claim 1, wherein the tensor further includes structure confidence channels for the amino acids that specify quality of respective structures of the amino acids.
35. The computer-readable media of claim 34, wherein the structure confidence channels are global model quality estimations (GMQEs).
36. The computer-readable media of claim 34, wherein the structure confidence channels include qualitative model energy analysis (QMEAN) scores.
37. The computer-readable media of claim 34, wherein the structure confidence channels are temperature factors that specify a degree to which the residues satisfy physical constraints of respective protein structures.

38. The computer-readable media of claim 34, wherein the structure confidence channels are template structures alignments that specify a degree to which residues of atoms nearest to the voxels have aligned template structures.
39. The computer-readable media of claim 38, wherein the structure confidence channels are template modeling scores of the aligned template structures.
40. The computer-readable media of claim 39, wherein the structure confidence channels are a minimum one of the template modeling scores, a mean of the template modeling scores, and a maximum one of the template modeling scores.
41. The computer-readable media of claim 1, the operations further comprising voxel-wise concatenating amino acid-wise distance channels for the alpha carbon atoms with the one-hot encoding of the alternative allele to generate the tensor.
42. The computer-readable media of claim 41, the operations further comprising voxel-wise concatenating amino acid-wise distance channels for the beta carbon atoms with the one-hot encoding of the alternative allele to generate the tensor.
43. The computer-readable media of claim 42, the operations further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, and the one-hot encoding of the alternative allele to generate the tensor.
44. The computer-readable media of claim 43, the operations further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, and pan-amino acid conservation frequencies sequences to generate the tensor.
45. The computer-readable media of claim 44, the operations further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, the pan-amino acid conservation frequencies sequences, and the annotation channels to generate the tensor.
46. The computer-readable media of claim 45, the operations further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele,

the pan-amino acid conservation frequencies sequences, the annotation channels, and the structure confidence channels to generate the tensor.

47. The computer-readable media of claim 46, the operations further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, and per-amino acid conservation frequencies for each of the amino acids to generate the tensor.

48. The computer-readable media of claim 47, the operations further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, per-amino acid conservation frequencies for each of the amino acids, and the annotation channels to generate the tensor.

49. The computer-readable media of claim 48, the operations further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, per-amino acid conservation frequencies for each of the amino acids, the annotation channels, and the structure confidence channels to generate the tensor.

50. The computer-readable media of claim 49, the operations further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, and the one-hot encoding of the reference allele to generate the tensor.

51. The computer-readable media of claim 50, the operations further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, and the pan-amino acid conservation frequencies sequences to generate the tensor.

52. The computer-readable media of claim 51, the operations further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, the pan-amino acid conservation frequencies sequences, and the annotation channels to generate the tensor.

53. The computer-readable media of claim 52, the operations further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, the pan-amino acid conservation frequencies sequences, the annotation channels, and the structure confidence channels to generate the tensor.

54. The computer-readable media of claim 53, the operations further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, and the per-amino acid conservation frequencies for each of the amino acids to generate the tensor.

55. The computer-readable media of claim 54, the operations further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, the per-amino acid conservation frequencies for each of the amino acids, and the annotation channels to generate the tensor.

56. The computer-readable media of claim 55, the operations further comprising voxel-wise concatenating the amino acid-wise distance channels for the alpha carbon atoms, the amino acid-wise distance channels for the beta carbon atoms, the one-hot encoding of the alternative allele, the one-hot encoding of the reference allele, the per-amino acid conservation frequencies for each of the amino acids, the annotation channels, and the structure confidence channels to generate the tensor.

57. The computer-readable media of claim 1, the operations further comprising rotating atoms of the amino acids before the amino acid-wise distance channels are generated.

58. The computer-readable media of claim 1, the operations further comprising using 1 x 1 x 1 convolutions, 3 x 3 x 3 convolutions, rectified linear unit activation layers, batch normalization layers, a fully-connected layer, a dropout regularization layer, and a softmax classification layer in a convolutional neural network.

59. The computer-readable media of claim 58, wherein the 1 x 1 x 1 convolutions and the 3 x 3 x 3 convolutions are three-dimensional convolutions.

60. The computer-readable media of claim 58, wherein a layer of the 1 x 1 x 1 convolutions processes the tensor and produces an intermediate output that is a convolved representation of

the tensor, wherein a sequence of layers of the $3 \times 3 \times 3$ convolutions processes the intermediate output and produces a flattened output, wherein the fully-connected layer processes the flattened output and produces unnormalized outputs, and wherein the softmax classification layer processes the unnormalized outputs and produces exponentially normalized outputs that identify likelihoods of the variant being pathogenic and benign.

61. The computer-readable media of claim 60, wherein a sigmoid layer processes the unnormalized outputs and produces a normalized output that identifies a likelihood of the variant being pathogenic.

62. The computer-readable media of claim 60, wherein the voxels, the atoms, and the distances have three-dimensional coordinates, wherein the tensor has at least three dimensions, wherein the intermediate output has at least three dimensions, and wherein the flattened output has one dimension.

63. One or more computer-readable media storing computer-executable instructions that, when executed on one or more processors, configure a computer to perform operations comprising:

storing atom category-wise distance channels for amino acids in a protein,
wherein the amino acids have atoms for a plurality of atom categories,
wherein atom categories in the plurality of atom categories specify atomic elements of the amino acids,

wherein each of the atom category-wise distance channels has voxel-wise distance values for voxels in a plurality of voxels, and

wherein the voxel-wise distance values specify distances from corresponding voxels in the plurality of voxels to atoms in corresponding atom categories in the plurality of atom categories;

processing a tensor that includes the atom category-wise distance channels and an alternative allele of the protein expressed by a variant; and
determining a pathogenicity of the variant based at least in part on the tensor.

64. The computer-readable media of claim 63, the operations further comprising centering a voxel grid of the voxels on respective atoms of respective atom categories in the plurality of atom categories.

65. The computer-readable media of claim 64, the operations further comprising centering the voxel grid on an alpha carbon atom of a residue of at least one variant amino acid in the protein.
66. The computer-readable media of claim 65, wherein the distances are nearest-atom distances from corresponding voxel centers in the voxel grid to nearest atoms in the corresponding atom categories.
67. The computer-readable media of claim 66, wherein the nearest-atom distances are Euclidean distances.
68. The computer-readable media of claim 67, wherein the nearest-atom distances are normalized by dividing the Euclidean distances with a maximum nearest-atom distances.
69. The computer-readable media of claim 68, wherein the distances are nearest-atom distances from the corresponding voxel centers in the voxel grid to nearest atoms irrespective of the amino acids and the atom categories of the amino acids.
70. The computer-readable media of claim 69, wherein the nearest-atom distances are Euclidean distances.
- [00240] 71. The computer-readable media of claim 70, wherein the nearest-atom distances are normalized by dividing the Euclidean distances with a maximum nearest-atom distances.

Particular Implementations 2

[00241] In some implementations, a system comprises a voxelizer that accesses a three-dimensional structure of a reference amino acid sequence of a protein and fits a three-dimensional grid of voxels on atoms in the three-dimensional structure on an amino acid-basis to generate amino acid-wise distance channels. Each of the amino acid-wise distance channels has a three-dimensional distance value for each voxel in the three-dimensional grid of voxels. The three-dimensional distance value specifies a distance from a corresponding voxel in the three-dimensional grid of voxels to atoms of a corresponding reference amino acid in the reference amino acid sequence. The system further comprises an alternative allele encoder that encodes an alternative allele amino acid to each voxel in the three-dimensional grid of voxels. The alternative allele amino acid is a three-dimensional representation of a one-hot encoding of a variant amino acid expressed by a variant nucleotide. The system further comprises an evolutionary conservation encoder that encodes an evolutionary conservation sequence to each voxel in the three-dimensional grid of voxels. The evolutionary conservation sequence can be a

three-dimensional representation of amino acid-specific conservation frequencies across a plurality of species. The amino acid-specific conservation frequencies can be selected in dependence upon amino acid proximity to the corresponding voxel. The system further comprises a convolutional neural network configured to apply three-dimensional convolutions to a tensor that includes the amino acid-wise distance channels encoded with the alternative allele amino acid and respective evolutionary conservation sequences. The convolutional neural network can be also configured to determine a pathogenicity of the variant nucleotide based at least in part on the tensor.

[00242] The voxelizer can center the three-dimensional grid of voxels on an alpha-carbon atom of respective residues of reference amino acids in the reference amino acid sequence. The voxelizer can center the three-dimensional grid of voxels on an alpha-carbon atom of a residue of a particular reference amino acid positioned at the variant amino acid.

[00243] In some implementations, the system can be further configured to encode, in the tensor, a directionality of the reference amino acids in the reference amino acid sequence and a position of the particular reference amino acid by multiplying, with a directionality parameter, three-dimensional distance values for those reference amino acids that precede the particular reference amino acid. The distances can be nearest-atom distances from corresponding voxel centers in the three-dimensional grid of voxels to nearest atoms of the corresponding reference amino acids. The nearest-atom distances can be Euclidean distances and can be normalized by dividing the Euclidean distances with a maximum nearest-atom distance.

[00244] In some implementations, the reference amino acids can have alpha-carbon atoms and the distances can be nearest-alpha-carbon atom distances from the corresponding voxel centers to nearest alpha-carbon atoms of the corresponding reference amino acids. In some implementations, the reference amino acids can have beta-carbon atoms and the distances can be nearest-beta-carbon atom distances from the corresponding voxel centers to nearest beta-carbon atoms of the corresponding reference amino acids. In some implementations, the reference amino acids can have backbone atoms and the distances can be nearest-backbone atom distances from the corresponding voxel centers to nearest backbone atoms of the corresponding reference amino acids. In some implementations, the amino acids can have sidechain atoms and the distances can be nearest-sidechain atom distances from the corresponding voxel centers to nearest sidechain atoms of the corresponding reference amino acids.

[00245] In some implementations, the system can be further configured to encode, in the tensor, a nearest atom channel that specifies a distance from each voxel to a nearest atom. The nearest atom can be selected irrespective of the amino acids and atomic elements of the amino acids. The distance can be a Euclidean distance and can be normalized by dividing the Euclidean

distance with a maximum distance. The amino acids can include non-standard amino acids. The tensor can further include an absentee atom channel that specifies atoms not found within a predefined radius of a voxel center. The absentee atom channel can be one-hot encoded.

[00246] In some implementations, the system can further comprise a reference allele encoder that voxel-wise encodes a reference allele amino acid to each three-dimensional distance value on the amino acid position-basis. The reference allele amino acid can be a three-dimensional representation of a one-hot encoding of the reference amino acid sequence. The amino acid-specific conservation frequencies can specify conservation levels of respective amino acids across the plurality of species.

[00247] In some implementations, the evolutionary conservation encoder can select a nearest atom to the corresponding voxel across the reference amino acids and the atom categories, can select pan-amino acid conservation frequencies for a residue of a reference amino acid that includes the nearest atom, and can use a three-dimensional representation of the pan-amino acid conservation frequencies as the evolutionary conservation sequence. The pan-amino acid conservation frequencies can be configured for a particular position of the residue as observed in the plurality of species. The pan-amino acid conservation frequencies can specify whether there is a missing conservation frequency for a particular reference amino acid.

[00248] In some implementations, the evolutionary conservation encoder can select respective nearest atoms to the corresponding voxel in respective ones of the reference amino acids, can select respective per-amino acid conservation frequencies for respective residues of the reference amino acids that include the nearest atoms, and can use a three-dimensional representation of the per-amino acid conservation frequencies as the evolutionary conservation sequence. The per-amino acid conservation frequencies can be configured for a particular position of the residues as observed in the plurality of species. The per-amino acid conservation frequencies can specify whether there is a missing conservation frequency for a particular reference amino acid.

[00249] In some implementations, the system can further comprise an annotations encoder that voxel-wise encodes one or more annotation channels to each three-dimensional distance value. The annotation channels can be three-dimensional representations of a one-hot encoding of residue annotations and can be molecular processing annotations that include initiator methionine, signal, transit peptide, propeptide, chain, and peptide. In some implementations, the annotation channels can be regions annotations that include topological domain, transmembrane, intramembrane, domain, repeat, calcium binding, zinc finger, deoxyribonucleic acid (DNA) binding, nucleotide binding, region, coiled coil, motif, and compositional bias or can be sites annotations that include active site, metal binding, binding site, and site. In some implementations, the annotation channels can be amino acid modifications annotations that

include non-standard residue, modified residue, lipidation, glycosylation, disulfide bond, and cross-link or can be secondary structure annotations that include helix, turn, and beta strand. The annotation channels can be experimental information annotations that include mutagenesis, sequence uncertainty, sequence conflict, non-adjacent residues, and non-terminal residue.

[00250] In some implementations, the system can further comprise a structure confidence encoder that voxel-wise encodes one or more structure confidence channels to each three-dimensional distance value. The structure confidence channels can be three-dimensional representations of confidence scores that specify quality of respective residue structures. The structure confidence channels can be global model quality estimations (GMQEs), can be qualitative model energy analysis (QMEAN) scores, can be temperature factors that specify a degree to which the residues satisfy physical constraints of respective protein structures, can be template structures alignments that specify a degree to which residues of atoms nearest to the voxels have aligned template structures, can be template modeling scores of the aligned template structures, or can be a minimum one of the template modeling scores, a mean of the template modeling scores, and a maximum one of the template modeling scores.

[00251] In some implementations, the system can further comprise an atoms rotation engine that rotates the atoms before the amino acid-wise distance channels are generated.

[00252] The convolutional neural network can use $1 \times 1 \times 1$ convolutions, $3 \times 3 \times 3$ convolutions, rectified linear unit activation layers, batch normalization layers, a fully-connected layer, a dropout regularization layer, and a softmax classification layer. The $1 \times 1 \times 1$ convolutions and the $3 \times 3 \times 3$ convolutions can be the three-dimensional convolutions. In some implementations, a layer of the $1 \times 1 \times 1$ convolutions can process the tensor and produce an intermediate output that is a convolved representation of the tensor. A sequence of layers of the $3 \times 3 \times 3$ convolutions can process the intermediate output and produce a flattened output. The fully-connected layer can process the flattened output and produce unnormalized outputs. The softmax classification layer can process the unnormalized outputs and produce exponentially normalized outputs that identify likelihoods of the variant nucleotide being pathogenic and benign.

[00253] In some implementations, a sigmoid layer can process the unnormalized outputs and produce a normalized output that identifies a likelihood of the variant nucleotide being pathogenic. The convolutional neural network can be an attention-based neural network. The tensor can include the amino acid-wise distance channels further encoded with the reference allele amino acid, can include the amino acid-wise distance channels further encoded with the annotation channels, or can include the amino acid-wise distance channels further encoded with the structure confidence channels.

[00254] In some implementations, a system can comprise a voxelizer that accesses a three-dimensional structure of a reference amino acid sequence of a protein and fits a three-dimensional grid of voxels on atoms in the three-dimensional structure on an amino acid-basis to generate atom category-wise distance channels. The atoms span a plurality of atom categories, which specify atomic elements of the amino acids. Each of the atom category-wise distance channels has a three-dimensional distance value for each voxel in the three-dimensional grid of voxels. The three-dimensional distance value specifies a distance from a corresponding voxel in the three-dimensional grid of voxels to atoms of corresponding atom categories in the plurality of atom categories. The system further comprises an alternative allele encoder that encodes an alternative allele amino acid to each voxel in the three-dimensional grid of voxels. The alternative allele amino acid is a three-dimensional representation of a one-hot encoding of a variant amino acid expressed by a variant nucleotide. The system further comprises an evolutionary conservation encoder that encodes an evolutionary conservation sequence to each voxel in the three-dimensional grid of voxels. The evolutionary conservation sequence can be a three-dimensional representation of amino acid-specific conservation frequencies across a plurality of species. The amino acid-specific conservation frequencies can be selected in dependence upon amino acid proximity to the corresponding voxel. The system further comprises a convolutional neural network configured to apply three-dimensional convolutions to a tensor that includes the atom category-wise distance channels encoded with the alternative allele amino acid and respective evolutionary conservation sequences, and to determine a pathogenicity of the variant nucleotide based at least in part on the tensor.

[00255] In some implementations, a system comprises a voxelizer that accesses a three-dimensional structure of a reference amino acid sequence of a protein and fits a three-dimensional grid of voxels on atoms in the three-dimensional structure on an amino acid-basis to generate amino acid-wise distance channels. Each of the amino acid-wise distance channels can have a three-dimensional distance value for each voxel in the three-dimensional grid of voxels. The three-dimensional distance value can specify a distance from a corresponding voxel in the three-dimensional grid of voxels to atoms of a corresponding reference amino acid in the reference amino acid sequence. The system further comprises an alternative allele encoder that encodes an alternative allele amino acid to each voxel in the three-dimensional grid of voxels. The alternative allele amino acid is a three-dimensional representation of a one-hot encoding of a variant amino acid expressed by a variant nucleotide. The system further comprises an evolutionary conservation encoder that encodes an evolutionary conservation sequence to each voxel in the three-dimensional grid of voxels. The evolutionary conservation sequence can be a three-dimensional representation of amino acid-specific conservation frequencies across a

plurality of species. The amino acid-specific conservation frequencies can be selected in dependence upon amino acid proximity to the corresponding voxel. The system further comprises a tensor generator configured to generate a tensor that includes the amino acid-wise distance channels encoded with the alternative allele amino acid and respective evolutionary conservation sequences.

[00256] In some implementations, a system comprises a voxelizer that accesses a three-dimensional structure of a reference amino acid sequence of a protein and fits a three-dimensional grid of voxels on atoms in the three-dimensional structure on an amino acid-basis to generate atom category-wise distance channels. The atoms can span a plurality of atom categories, which specify atomic elements of the amino acids. Each of the atom category-wise distance channels can have a three-dimensional distance value for each voxel in the three-dimensional grid of voxels. The three-dimensional distance value can specify a distance from a corresponding voxel in the three-dimensional grid of voxels to atoms of corresponding atom categories in the plurality of atom categories. The system further comprises an alternative allele encoder that encodes an alternative allele amino acid to each voxel in the three-dimensional grid of voxels. The alternative allele amino acid is a three-dimensional representation of a one-hot encoding of a variant amino acid expressed by a variant nucleotide. The system further comprises an evolutionary conservation encoder that encodes an evolutionary conservation sequence to each voxel in the three-dimensional grid of voxels. The evolutionary conservation sequence can be a three-dimensional representation of amino acid-specific conservation frequencies across a plurality of species. The amino acid-specific conservation frequencies can be selected in dependence upon amino acid proximity to the corresponding voxel. The system further comprises a tensor generator configured to generate a tensor that includes the atom category-wise distance channels encoded with the alternative allele amino acid and respective evolutionary conservation sequences.

Clauses 2

1. A computer-implemented method, comprising:
accessing a three-dimensional structure of a reference amino acid sequence of a protein, and
fitting a three-dimensional grid of voxels on atoms in the three-dimensional structure on an amino acid-basis to generate amino acid-wise distance channels,
wherein each of the amino acid-wise distance channels has a three-dimensional distance value for each voxel in the three-dimensional grid of voxels, and

wherein the three-dimensional distance value specifies a distance from a corresponding voxel in the three-dimensional grid of voxels to atoms of a corresponding reference amino acid in the reference amino acid sequence;

encoding an alternative allele channel to each voxel in the three-dimensional grid of voxels, wherein the alternative allele channel is a three-dimensional representation of a one-hot encoding of a variant amino acid expressed by a variant nucleotide;

encoding an evolutionary conservation channel to each sequence of three-dimensional distance values across the amino acid-wise distance channels on a voxel position-basis, wherein the evolutionary conservation channel is a three-dimensional representation of amino acid-specific conservation frequencies across a plurality of species, and wherein the amino acid-specific conservation frequencies are selected in dependence upon amino acid proximity to the corresponding voxel;

applying three-dimensional convolutions to a tensor that includes the amino acid-wise distance channels encoded with the alternative allele channel and respective evolutionary conservation channels; and

determining a pathogenicity of the variant nucleotide based at least in part on the tensor.

2. The computer-implemented method of clause 1, further comprising centering the three-dimensional grid of voxels on an alpha carbon atom of respective residues of reference amino acids in the reference amino acid sequence.
3. The computer-implemented method of clause 2, further comprising centering the three-dimensional grid of voxels on an alpha carbon atom of a residue of a particular reference amino acid that corresponds to the variant amino acid.
4. The computer-implemented method of clause 3, further comprising encoding, in the tensor, a directionality of the reference amino acids in the reference amino acid sequence and a position of the particular reference amino acid by multiplying, with a directionality parameter, three-dimensional distance values for those reference amino acids that precede the particular reference amino acid.
5. The computer-implemented method of clause 4, wherein the distances are nearest-atom distances from corresponding voxel centers in the three-dimensional grid of voxels to nearest atoms of the corresponding reference amino acids.
6. The computer-implemented method of clause 5, wherein the nearest-atom distances are Euclidean distances.

7. The computer-implemented method of clause 6, wherein the nearest-atom distances are normalized by dividing the Euclidean distances with a maximum nearest-atom distance.
8. The computer-implemented method of clause 5, wherein the reference amino acids have alpha carbon atoms and wherein the distances are nearest-alpha carbon atom distances from the corresponding voxel centers to nearest alpha carbon atoms of the corresponding reference amino acids.
9. The computer-implemented method of clause 5, wherein the reference amino acids have beta carbon atoms and wherein the distances are nearest-beta carbon atom distances from the corresponding voxel centers to nearest beta carbon atoms of the corresponding reference amino acids.
10. The computer-implemented method of clause 5, wherein the reference amino acids have backbone atoms and wherein the distances are nearest-backbone atom distances from the corresponding voxel centers to nearest backbone atoms of the corresponding reference amino acids.
11. The computer-implemented method of clause 5, wherein the amino acids have sidechain atoms and wherein the distances are nearest-sidechain atom distances from the corresponding voxel centers to nearest sidechain atoms of the corresponding reference amino acids.
12. The computer-implemented method of clause 3, further comprising encoding, in the tensor, a nearest atom channel that specifies a distance from each voxel to a nearest atom, wherein the nearest atom is selected irrespective of the amino acids and atomic elements of the amino acids.
13. The computer-implemented method of clause 12, wherein the distance is a Euclidean distance.
14. The computer-implemented method of clause 13, wherein the distance is normalized by dividing the Euclidean distance with a maximum distance.
15. The computer-implemented method of clause 12, wherein the amino acids include non-standard amino acids.
16. The computer-implemented method of clause 1, wherein the tensor further includes an absentee atom channel that specifies atoms not found within a predefined radius of a voxel center.

17. The computer-implemented method of clause 16, wherein the absentee atom channel is one-hot encoded.
18. The computer-implemented method of clause 1, further comprising voxel-wise encoding a reference allele channel to each voxel in the three-dimensional grid of voxels.
19. The computer-implemented method of clause 18, the reference allele amino acid is a three-dimensional representation of a one-hot encoding of a reference amino acid that experiences the variant amino acid.
20. The computer-implemented method of clause 1, wherein the amino acid-specific conservation frequencies specify conservation levels of respective amino acids across the plurality of species.
21. The computer-implemented method of clause 20, further comprising:
selecting a nearest atom to the corresponding voxel across the reference amino acids and the atom categories,
selecting pan-amino acid conservation frequencies for a residue of a reference amino acid that includes the nearest atom, and
using a three-dimensional representation of the pan-amino acid conservation frequencies as the evolutionary conservation channel.
22. The computer-implemented method of clause 21, wherein the pan-amino acid conservation frequencies are configured for a particular position of the residue as observed in the plurality of species.
23. The computer-implemented method of clause 21, wherein the pan-amino acid conservation frequencies specify whether there is a missing conservation frequency for a particular reference amino acid.
24. The computer-implemented method of clause 21, further comprising:
selecting respective nearest atoms to the corresponding voxel in respective ones of the reference amino acids,
selecting respective per-amino acid conservation frequencies for respective residues of the reference amino acids that include the nearest atoms, and
using a three-dimensional representation of the per-amino acid conservation frequencies as the evolutionary conservation channel.

25. The computer-implemented method of clause 24, wherein the per-amino acid conservation frequencies are configured for a particular position of the residues as observed in the plurality of species.
26. The computer-implemented method of clause 24, wherein the per-amino acid conservation frequencies specify whether there is a missing conservation frequency for a particular reference amino acid.
27. The computer-implemented method of clause 1, further comprising voxel-wise encoding one or more annotation channels to each voxel in the three-dimensional grid of voxels, wherein the annotation channels are three-dimensional representations of a one-hot encoding of residue annotations.
28. The computer-implemented method of clause 27, wherein the annotation channels are molecular processing annotations that include initiator methionine, signal, transit peptide, propeptide, chain, and peptide.
29. The computer-implemented method of clause 27, wherein the annotation channels are regions annotations that include topological domain, transmembrane, intramembrane, domain, repeat, calcium binding, zinc finger, deoxyribonucleic acid (DNA) binding, nucleotide binding, region, coiled coil, motif, and compositional bias.
30. The computer-implemented method of clause 27, wherein the annotation channels are sites annotations that include active site, metal binding, binding site, and site.
31. The computer-implemented method of clause 27, wherein the annotation channels are amino acid modifications annotations that include non-standard residue, modified residue, lipidation, glycosylation, disulfide bond, and cross-link.
32. The computer-implemented method of clause 27, wherein the annotation channels are secondary structure annotations that include helix, turn, and beta strand.
33. The computer-implemented method of clause 27, wherein the annotation channels are experimental information annotations that include mutagenesis, sequence uncertainty, sequence conflict, non-adjacent residues, and non-terminal residue.
34. The computer-implemented method of clause 1, further comprising voxel-wise encoding one or more structure confidence channels to each voxel in the three-dimensional grid of voxels,

wherein the structure confidence channels are three-dimensional representations of confidence scores that specify quality of respective residue structures.

35. The computer-implemented method of clause 34, wherein the structure confidence channels are global model quality estimations (GMQEs).

36. The computer-implemented method of clause 34, wherein the structure confidence channels are qualitative model energy analysis (QMEAN) scores.

37. The computer-implemented method of clause 34, wherein the structure confidence channels are temperature factors that specify a degree to which the residues satisfy physical constraints of respective protein structures.

38. The computer-implemented method of clause 34, wherein the structure confidence channels are template structures alignments that specify a degree to which residues of atoms nearest to the voxels have aligned template structures.

39. The computer-implemented method of clause 38, wherein the structure confidence channels are template modeling scores of the aligned template structures.

40. The computer-implemented method of clause 39, wherein the structure confidence channels are a minimum one of the template modeling scores, a mean of the template modeling scores, and a maximum one of the template modeling scores.

41. The computer-implemented method of clause 1, further comprising rotating the atoms before the amino acid-wise distance channels are generated.

42. The computer-implemented method of clause 1, further comprising using 1 x 1 x 1 convolutions, 3 x 3 x 3 convolutions, rectified linear unit activation layers, batch normalization layers, a fully-connected layer, a dropout regularization layer, and a softmax classification layer in a convolutional neural network.

43. The computer-implemented method of clause 42, wherein the 1 x 1 x 1 convolutions and the 3 x 3 x 3 convolutions are the three-dimensional convolutions.

44. The computer-implemented method of clause 42, wherein a layer of the 1 x 1 x 1 convolutions processes the tensor and produces an intermediate output that is a convolved representation of the tensor, wherein a sequence of layers of the 3 x 3 x 3 convolutions processes the intermediate output and produces a flattened output, wherein the fully-connected layer

processes the flattened output and produces unnormalized outputs, and wherein the softmax classification layer processes the unnormalized outputs and produces exponentially normalized outputs that identify likelihoods of the variant nucleotide being pathogenic and benign.

45. The computer-implemented method of clause 44, wherein a sigmoid layer processes the unnormalized outputs and produces a normalized output that identifies a likelihood of the variant nucleotide being pathogenic.

46. The computer-implemented method of clause 1, wherein the convolutional neural network is an attention-based neural network.

47. The computer-implemented method of clause 1, wherein the tensor includes the amino acid-wise distance channels further encoded with the reference allele channel.

48. The computer-implemented method of clause 1, wherein the tensor includes the amino acid-wise distance channels further encoded with the annotation channels.

49. The computer-implemented method of clause 1, wherein the tensor includes the amino acid-wise distance channels further encoded with the structure confidence channels.

50. A computer-implemented method, comprising:

accessing a three-dimensional structure of a reference amino acid sequence of a protein, and
fitting a three-dimensional grid of voxels on atoms in the three-dimensional structure on an amino acid-basis to generate atom category-wise distance channels,
wherein the atoms span a plurality of atom categories,
wherein atom categories in the plurality of atom categories specify atomic elements of the amino acids,
wherein each of the atom category-wise distance channels has a three-dimensional distance value for each voxel in the three-dimensional grid of voxels, and
wherein the three-dimensional distance value specifies a distance from a corresponding voxel in the three-dimensional grid of voxels to atoms of corresponding atom categories in the plurality of atom categories;
encoding an alternative allele channel to each voxel in the three-dimensional grid of voxels,
wherein the alternative allele channel is a three-dimensional representation of a one-hot encoding of a variant amino acid expressed by a variant nucleotide;
encoding an evolutionary conservation channel to each sequence of three-dimensional distance values across the atom category-wise distance channels on a voxel position-basis,

wherein the evolutionary conservation channel is a three-dimensional representation of amino acid-specific conservation frequencies across a plurality of species, and wherein the amino acid-specific conservation frequencies are selected in dependence upon amino acid proximity to the corresponding voxel;

applying three-dimensional convolutions to a tensor that includes the atom category-wise distance channels encoded with the alternative allele channel and respective evolutionary conservation channels; and

determining a pathogenicity of the variant nucleotide based at least in part on the tensor.

51. A computer-implemented method, comprising:

accessing a three-dimensional structure of a reference amino acid sequence of a protein, and fitting a three-dimensional grid of voxels on atoms in the three-dimensional structure on an amino acid-basis to generate amino acid-wise distance channels,

wherein each of the amino acid-wise distance channels has a three-dimensional distance value for each voxel in the three-dimensional grid of voxels, and

wherein the three-dimensional distance value specifies a distance from a corresponding voxel in the three-dimensional grid of voxels to atoms of a corresponding reference amino acid in the reference amino acid sequence;

encoding an alternative allele channel to each voxel in the three-dimensional grid of voxels, wherein the alternative allele channel is a three-dimensional representation of a one-hot encoding of a variant amino acid expressed by a variant nucleotide;

encoding an evolutionary conservation channel to each sequence of three-dimensional distance values across the amino acid-wise distance channels on a voxel position-basis, wherein the evolutionary conservation channel is a three-dimensional representation of amino acid-specific conservation frequencies across a plurality of species, and wherein the amino acid-specific conservation frequencies are selected in dependence upon amino acid proximity to the corresponding voxel; and

generating a tensor that includes the amino acid-wise distance channels encoded with the alternative allele channel and respective evolutionary conservation channels.

52. A computer-implemented method, comprising:

accessing a three-dimensional structure of a reference amino acid sequence of a protein, and fitting a three-dimensional grid of voxels on atoms in the three-dimensional structure on an amino acid-basis to generate atom category-wise distance channels,

wherein the atoms span a plurality of atom categories,

wherein atom categories in the plurality of atom categories specify atomic elements of the amino acids,

wherein each of the atom category-wise distance channels has a three-dimensional distance value for each voxel in the three-dimensional grid of voxels, and

wherein the three-dimensional distance value specifies a distance from a corresponding voxel in the three-dimensional grid of voxels to atoms of corresponding atom categories in the plurality of atom categories;

encoding an alternative allele channel to each voxel in the three-dimensional grid of voxels,

wherein the alternative allele channel is a three-dimensional representation of a one-hot encoding of a variant amino acid expressed by a variant nucleotide;

encoding an evolutionary conservation channel to each sequence of three-dimensional distance values across the atom category-wise distance channels on a voxel position-basis,

wherein the evolutionary conservation channel is a three-dimensional representation of amino acid-specific conservation frequencies across a plurality of species, and

wherein the amino acid-specific conservation frequencies are selected in dependence upon amino acid proximity to the corresponding voxel; and

generating a tensor that includes the atom category-wise distance channels encoded with the alternative allele channel and respective evolutionary conservation channels.

1. One or more computer-readable media storing computer-executable instructions that, when executed on one or more processors, configure a computer to perform operations comprising:

accessing a three-dimensional structure of a reference amino acid sequence of a protein, and

fitting a three-dimensional grid of voxels on atoms in the three-dimensional structure on an amino acid-basis to generate amino acid-wise distance channels,

wherein each of the amino acid-wise distance channels has a three-dimensional distance value for each voxel in the three-dimensional grid of voxels, and

wherein the three-dimensional distance value specifies a distance from a corresponding voxel in the three-dimensional grid of voxels to atoms of a corresponding reference amino acid in the reference amino acid sequence;

encoding an alternative allele channel to each voxel in the three-dimensional grid of voxels,

wherein the alternative allele channel is a three-dimensional representation of a one-hot encoding of a variant amino acid expressed by a variant nucleotide;

encoding an evolutionary conservation channel to each sequence of three-dimensional distance values across the amino acid-wise distance channels on a voxel position-basis,

wherein the evolutionary conservation channel is a three-dimensional representation of amino acid-specific conservation frequencies across a plurality of species, and wherein the amino acid-specific conservation frequencies are selected in dependence upon amino acid proximity to the corresponding voxel; applying three-dimensional convolutions to a tensor that includes the amino acid-wise distance channels encoded with the alternative allele channel and respective evolutionary conservation channels; and determining a pathogenicity of the variant nucleotide based at least in part on the tensor.

2. The computer-readable media of clause 1, the operations further comprising centering the three-dimensional grid of voxels on an alpha carbon atom of respective residues of reference amino acids in the reference amino acid sequence.

3. The computer-readable media of clause 2, the operations further comprising centering the three-dimensional grid of voxels on an alpha carbon atom of a residue of a particular reference amino acid that corresponds to the variant amino acid.

4. The computer-readable media of clause 3, the operations further comprising encoding, in the tensor, a directionality of the reference amino acids in the reference amino acid sequence and a position of the particular reference amino acid by multiplying, with a directionality parameter, three-dimensional distance values for those reference amino acids that precede the particular reference amino acid.

5. The computer-readable media of clause 4, wherein the distances are nearest-atom distances from corresponding voxel centers in the three-dimensional grid of voxels to nearest atoms of the corresponding reference amino acids.

6. The computer-readable media of clause 5, wherein the nearest-atom distances are Euclidean distances.

7. The computer-readable media of clause 6, wherein the nearest-atom distances are normalized by dividing the Euclidean distances with a maximum nearest-atom distance.

8. The computer-readable media of clause 5, wherein the reference amino acids have alpha carbon atoms and wherein the distances are nearest-alpha carbon atom distances from the corresponding voxel centers to nearest alpha carbon atoms of the corresponding reference amino acids.

9. The computer-readable media of clause 5, wherein the reference amino acids have beta carbon atoms and wherein the distances are nearest-beta carbon atom distances from the corresponding voxel centers to nearest beta carbon atoms of the corresponding reference amino acids.
10. The computer-readable media of clause 5, wherein the reference amino acids have backbone atoms and wherein the distances are nearest-backbone atom distances from the corresponding voxel centers to nearest backbone atoms of the corresponding reference amino acids.
11. The computer-readable media of clause 5, wherein the amino acids have sidechain atoms and wherein the distances are nearest-sidechain atom distances from the corresponding voxel centers to nearest sidechain atoms of the corresponding reference amino acids.
12. The computer-readable media of clause 3, the operations further comprising encoding, in the tensor, a nearest atom channel that specifies a distance from each voxel to a nearest atom, wherein the nearest atom is selected irrespective of the amino acids and atomic elements of the amino acids.
13. The computer-readable media of clause 12, wherein the distance is a Euclidean distance.
14. The computer-readable media of clause 13, wherein the distance is normalized by dividing the Euclidean distance with a maximum distance.
15. The computer-readable media of clause 12, wherein the amino acids include non-standard amino acids.
16. The computer-readable media of clause 1, wherein the tensor further includes an absentee atom channel that specifies atoms not found within a predefined radius of a voxel center.
17. The computer-readable media of clause 16, wherein the absentee atom channel is one-hot encoded.
18. The computer-readable media of clause 1, the operations further comprising voxel-wise encoding a reference allele channel to each voxel in the three-dimensional grid of voxels.
19. The computer-readable media of clause 18, the reference allele amino acid is a three-dimensional representation of a one-hot encoding of a reference amino acid that experiences the variant amino acid.

20. The computer-readable media of clause 1, wherein the amino acid-specific conservation frequencies specify conservation levels of respective amino acids across the plurality of species.

21. The computer-readable media of clause 20, the operations further comprising:
selecting a nearest atom to the corresponding voxel across the reference amino acids and the atom categories,
selecting pan-amino acid conservation frequencies for a residue of a reference amino acid that includes the nearest atom, and
using a three-dimensional representation of the pan-amino acid conservation frequencies as the evolutionary conservation channel.

22. The computer-readable media of clause 21, wherein the pan-amino acid conservation frequencies are configured for a particular position of the residue as observed in the plurality of species.

23. The computer-readable media of clause 21, wherein the pan-amino acid conservation frequencies specify whether there is a missing conservation frequency for a particular reference amino acid.

24. The computer-readable media of clause 21, the operations further comprising:
selecting respective nearest atoms to the corresponding voxel in respective ones of the reference amino acids,
selecting respective per-amino acid conservation frequencies for respective residues of the reference amino acids that include the nearest atoms, and
using a three-dimensional representation of the per-amino acid conservation frequencies as the evolutionary conservation channel.

25. The computer-readable media of clause 24, wherein the per-amino acid conservation frequencies are configured for a particular position of the residues as observed in the plurality of species.

26. The computer-readable media of clause 24, wherein the per-amino acid conservation frequencies specify whether there is a missing conservation frequency for a particular reference amino acid.

27. The computer-readable media of clause 1, the operations further comprising voxel-wise encoding one or more annotation channels to each voxel in the three-dimensional grid of voxels,

wherein the annotation channels are three-dimensional representations of a one-hot encoding of residue annotations.

28. The computer-readable media of clause 27, wherein the annotation channels are molecular processing annotations that include initiator methionine, signal, transit peptide, propeptide, chain, and peptide.
29. The computer-readable media of clause 27, wherein the annotation channels are regions annotations that include topological domain, transmembrane, intramembrane, domain, repeat, calcium binding, zinc finger, deoxyribonucleic acid (DNA) binding, nucleotide binding, region, coiled coil, motif, and compositional bias.
30. The computer-readable media of clause 27, wherein the annotation channels are sites annotations that include active site, metal binding, binding site, and site.
31. The computer-readable media of clause 27, wherein the annotation channels are amino acid modifications annotations that include non-standard residue, modified residue, lipidation, glycosylation, disulfide bond, and cross-link.
32. The computer-readable media of clause 27, wherein the annotation channels are secondary structure annotations that include helix, turn, and beta strand.
33. The computer-readable media of clause 27, wherein the annotation channels are experimental information annotations that include mutagenesis, sequence uncertainty, sequence conflict, non-adjacent residues, and non-terminal residue.
34. The computer-readable media of clause 1, the operations further comprising voxel-wise encoding one or more structure confidence channels to each voxel in the three-dimensional grid of voxels, wherein the structure confidence channels are three-dimensional representations of confidence scores that specify quality of respective residue structures.
35. The computer-readable media of clause 34, wherein the structure confidence channels are global model quality estimations (GMQEs).
36. The computer-readable media of clause 34, wherein the structure confidence channels are qualitative model energy analysis (QMEAN) scores.

37. The computer-readable media of clause 34, wherein the structure confidence channels are temperature factors that specify a degree to which the residues satisfy physical constraints of respective protein structures.
38. The computer-readable media of clause 34, wherein the structure confidence channels are template structures alignments that specify a degree to which residues of atoms nearest to the voxels have aligned template structures.
39. The computer-readable media of clause 38, wherein the structure confidence channels are template modeling scores of the aligned template structures.
40. The computer-readable media of clause 39, wherein the structure confidence channels are a minimum one of the template modeling scores, a mean of the template modeling scores, and a maximum one of the template modeling scores.
41. The computer-readable media of clause 1, the operations further comprising rotating the atoms before the amino acid-wise distance channels are generated.
42. The computer-readable media of clause 1, the operations further comprising using 1 x 1 x 1 convolutions, 3 x 3 x 3 convolutions, rectified linear unit activation layers, batch normalization layers, a fully-connected layer, a dropout regularization layer, and a softmax classification layer in a convolutional neural network.
43. The computer-readable media of clause 42, wherein the 1 x 1 x 1 convolutions and the 3 x 3 x 3 convolutions are the three-dimensional convolutions.
44. The computer-readable media of clause 42, wherein a layer of the 1 x 1 x 1 convolutions processes the tensor and produces an intermediate output that is a convolved representation of the tensor, wherein a sequence of layers of the 3 x 3 x 3 convolutions processes the intermediate output and produces a flattened output, wherein the fully-connected layer processes the flattened output and produces unnormalized outputs, and wherein the softmax classification layer processes the unnormalized outputs and produces exponentially normalized outputs that identify likelihoods of the variant nucleotide being pathogenic and benign.
45. The computer-readable media of clause 44, wherein a sigmoid layer processes the unnormalized outputs and produces a normalized output that identifies a likelihood of the variant nucleotide being pathogenic.

46. The computer-readable media of clause 1, wherein the convolutional neural network is an attention-based neural network.
47. The computer-readable media of clause 1, wherein the tensor includes the amino acid-wise distance channels further encoded with the reference allele channel.
48. The computer-readable media of clause 1, wherein the tensor includes the amino acid-wise distance channels further encoded with the annotation channels.
49. The computer-readable media of clause 1, wherein the tensor includes the amino acid-wise distance channels further encoded with the structure confidence channels.
50. One or more computer-readable media storing computer-executable instructions that, when executed on one or more processors, configure a computer to perform operations comprising:
- accessing a three-dimensional structure of a reference amino acid sequence of a protein, and
fitting a three-dimensional grid of voxels on atoms in the three-dimensional structure on an amino acid-basis to generate atom category-wise distance channels,
wherein the atoms span a plurality of atom categories,
wherein atom categories in the plurality of atom categories specify atomic elements of the amino acids,
wherein each of the atom category-wise distance channels has a three-dimensional distance value for each voxel in the three-dimensional grid of voxels, and
wherein the three-dimensional distance value specifies a distance from a corresponding voxel in the three-dimensional grid of voxels to atoms of corresponding atom categories in the plurality of atom categories;
- encoding an alternative allele channel to each voxel in the three-dimensional grid of voxels,
wherein the alternative allele channel is a three-dimensional representation of a one-hot encoding of a variant amino acid expressed by a variant nucleotide;
- encoding an evolutionary conservation channel to each sequence of three-dimensional distance values across the atom category-wise distance channels on a voxel position-basis,
wherein the evolutionary conservation channel is a three-dimensional representation of amino acid-specific conservation frequencies across a plurality of species, and
wherein the amino acid-specific conservation frequencies are selected in dependence upon amino acid proximity to the corresponding voxel;

applying three-dimensional convolutions to a tensor that includes the atom category-wise distance channels encoded with the alternative allele channel and respective evolutionary conservation channels; and
determining a pathogenicity of the variant nucleotide based at least in part on the tensor.

51. One or more computer-readable media storing computer-executable instructions that, when executed on one or more processors, configure a computer to perform operations comprising:

accessing a three-dimensional structure of a reference amino acid sequence of a protein, and fitting a three-dimensional grid of voxels on atoms in the three-dimensional structure on an amino acid-basis to generate amino acid-wise distance channels,
wherein each of the amino acid-wise distance channels has a three-dimensional distance value for each voxel in the three-dimensional grid of voxels, and
wherein the three-dimensional distance value specifies a distance from a corresponding voxel in the three-dimensional grid of voxels to atoms of a corresponding reference amino acid in the reference amino acid sequence;
encoding an alternative allele channel to each three-dimensional distance value in each of the amino acid-wise distance channels on an amino acid position-basis,
wherein the alternative allele channel is a three-dimensional representation of a one-hot encoding of a variant amino acid expressed by a variant nucleotide;
encoding an evolutionary conservation channel to each sequence of three-dimensional distance values across the amino acid-wise distance channels on a voxel position-basis,
wherein the evolutionary conservation channel is a three-dimensional representation of amino acid-specific conservation frequencies across a plurality of species, and
wherein the amino acid-specific conservation frequencies are selected in dependence upon amino acid proximity to the corresponding voxel; and
generating a tensor that includes the amino acid-wise distance channels encoded with the alternative allele channel and respective evolutionary conservation channels.

52. One or more computer-readable media storing computer-executable instructions that, when executed on one or more processors, configure a computer to perform operations comprising:

accessing a three-dimensional structure of a reference amino acid sequence of a protein, and fitting a three-dimensional grid of voxels on atoms in the three-dimensional structure on an amino acid-basis to generate atom category-wise distance channels,

wherein the atoms span a plurality of atom categories,
wherein atom categories in the plurality of atom categories specify atomic elements of the amino acids,
wherein each of the atom category-wise distance channels has a three-dimensional distance value for each voxel in the three-dimensional grid of voxels, and
wherein the three-dimensional distance value specifies a distance from a corresponding voxel in the three-dimensional grid of voxels to atoms of corresponding atom categories in the plurality of atom categories;
encoding an alternative allele channel to each voxel in the three-dimensional grid of voxels, wherein the alternative allele channel is a three-dimensional representation of a one-hot encoding of a variant amino acid expressed by a variant nucleotide;
encoding an evolutionary conservation channel to each sequence of three-dimensional distance values across the atom category-wise distance channels on a voxel position-basis, wherein the evolutionary conservation channel is a three-dimensional representation of amino acid-specific conservation frequencies across a plurality of species, and
wherein the amino acid-specific conservation frequencies are selected in dependence upon amino acid proximity to the corresponding voxel; and
generating a tensor that includes the atom category-wise distance channels encoded with the alternative allele channel and respective evolutionary conservation channels.

[00257] Other implementations of the method described in this section can include a non-transitory computer readable storage medium storing instructions executable by a processor to perform any of the methods described above. Yet another implementation of the method described in this section can include a system including memory and one or more processors operable to execute instructions, stored in the memory, to perform any of the methods described above.

Particular Implementations 3

Clauses 3

1. A computer-implemented method of efficiently determining which elements of a sequence are nearest to uniformly spaced cells in a grid, wherein the elements have element coordinates, and the cells have dimension-wise cell indices and cell coordinates, including:
generating an element-to-cells mapping that maps, to each of the elements, a subset of the cells, wherein the subset of the cells mapped to a particular element in the sequence includes a nearest cell in the grid and one or more neighborhood cells in the grid,

wherein the nearest cell is selected based on matching element coordinates of the particular element to the cell coordinates, and
wherein the neighborhood cells are contiguously adjacent to the nearest cell and selected based on being within a distance proximity range from the particular element;
generating a cell-to-elements mapping that maps, to each of the cells, a subset of the elements, wherein the subset of the elements mapped to a particular cell in the grid includes those elements in the sequence that are mapped to the particular cell by the element-to-cells mapping; and
using the cell-to-elements mapping to determine, for each of the cells, a nearest element in the sequence,
wherein the nearest element to the particular cell is determined based on distances between the particular cell and the elements in the subset of the elements.

2. The computer-implemented method of clause 1, wherein the matching the element coordinates of the particular element to the cell coordinates further includes truncating a decimal portion of the element coordinates to generate truncated element coordinates.

3. The computer-implemented method of clause 2, wherein the matching the element coordinates of the particular element to the cell coordinates further includes:

for a first dimension, matching a first truncated element coordinate in the truncated element coordinates to a first cell coordinate of a first cell in the grid, and selecting a first dimension index of the first cell;
for a second dimension, matching a second truncated element coordinate in the truncated element coordinates to a second cell coordinate of a second cell in the grid, and selecting a second dimension index of the second cell;
for a third dimension, matching a third truncated element coordinate in the truncated element coordinates to a third cell coordinate of a third cell in the grid, and selecting a third dimension index of the third cell;
using the selected first, second, and third dimension indices to generate an accumulated sum based on position-wise weighting the selected first, second, and third dimension indices by powers of a radix; and
using the accumulated sum as a cell index for selection of the nearest cell.

4. The computer-implemented method of clause 1, wherein the distances are calculated between cell coordinates of the particular cell and element coordinates of the elements in the subset of the elements.

5. The computer-implemented method of clause 1, wherein the sequence is a protein sequence of amino acids.
6. The computer-implemented method of clause 5, wherein the elements are atoms of the amino acids.
7. The computer-implemented method of clause 6, wherein the steps of generating the element-to-cells mapping, generating the cell-to-elements mapping, and using the cell-to-elements mapping to determine, for each of the cells, the nearest element have a runtime complexity of $O(a * f + v)$, wherein
a is a number of the atoms,
f is a number of the amino acids,
v is a number of the cells, and
* is a multiplication operation.
8. The computer-implemented method of clause 7, wherein the atoms include alpha carbon atoms.
9. The computer-implemented method of clause 7, wherein the atoms include beta carbon atoms.
10. The computer-implemented method of clause 7, wherein the atoms include non-carbon atoms.
11. The computer-implemented method of clause 1, wherein the cells are three-dimensional voxels.
12. The computer-implemented method of clause 11, wherein the cell coordinates are three-dimensional coordinates.
13. The computer-implemented method of clause 12, wherein the element coordinates are three-dimensional coordinates.
14. The computer-implemented method of clause 1, wherein the neighborhood cells are selected based on being within an index adjacency range from the nearest cell.
15. The computer-implemented method of clause 1, wherein the neighborhood cells are selected based on being within a cell neighborhood in the grid that includes the nearest cell.

16. The computer-implemented method of clause 1, wherein the sequence includes M elements, wherein the subset of the elements includes N elements, and wherein $M \gg N$.
17. A computer-implemented method of efficiently determining which atoms in a protein are nearest to voxels in a grid, wherein the atoms have three-dimensional (3D) atom coordinates, and the voxels have 3D voxel coordinates, including:
- generating an atom-to-voxels mapping that maps, to each of the atoms, a containing voxel selected based on matching 3D atom coordinates of a particular atom of the protein to the 3D voxel coordinates in the grid;
- generating a voxel-to-atoms mapping that maps, to each of the voxels, a subset of the atoms, wherein the subset of the atoms mapped to a particular voxel in the grid includes those atoms in the protein that are mapped to the particular voxel by the atom-to-voxels mapping; and
- using the voxel-to-atoms mapping to determine, for each of the voxels, a nearest atom in the protein.
18. The computer-implemented method of clause 17, wherein the steps of clause 17 have a runtime complexity of $O(\text{number of atoms})$.
- [00258]** Other implementations of the method described in this section can include a non-transitory computer readable storage medium storing instructions executable by a processor to perform any of the methods described above. Yet another implementation of the method described in this section can include a system including memory and one or more processors operable to execute instructions, stored in the memory, to perform any of the methods described above.
- [00259]** While the present invention is disclosed by reference to the preferred implementations and examples detailed above, it is to be understood that these examples are intended in an illustrative rather than in a limiting sense. It is contemplated that modifications and combinations will readily occur to those skilled in the art, which modifications and combinations will be within the spirit of the invention and the scope of the following claims.
- [00260]** What is claimed is:

CLAIMS

1. A system, comprising:
memory storing amino acid-wise distance channels for a plurality of amino acids in a protein,
wherein each of the amino acid-wise distance channels has voxel-wise distance values for voxels
in a plurality of voxels, and
wherein the voxel-wise distance values specify distances from corresponding voxels in the
plurality of voxels to atoms of corresponding amino acids in the plurality of amino acids; and
a pathogenicity determination engine configured to
process a tensor that includes the amino acid-wise distance channels and an alternative allele
amino acid of the protein expressed by a variant, and
determine a pathogenicity of the variant based at least in part on the tensor.
2. The system of claim 1, further comprising a distance channels generator that centers a
voxel grid of the voxels on an alpha-carbon atom of respective residues of the amino acids.
3. The system of claim 2, wherein the distance channels generator centers the voxel grid on
an alpha-carbon atom of a residue of a particular amino acid positioned at a variant amino acid in
the protein.
4. The system of any of claims 1-3, further configured to encode, in the tensor, a
directionality of the amino acids and a position of the particular amino acid by multiplying, with
a directionality parameter, voxel-wise distance values for those amino acids that precede the
particular amino acid.
5. The system of any of claims 1-4, wherein the distances are nearest-atom distances from
corresponding voxel centers in the voxel grid to nearest atoms of the corresponding amino acids.
6. The system of claim 5, wherein the nearest-atom distances are Euclidean distances.
7. The system of claim 5 or 6, wherein the nearest-atom distances are normalized by
dividing the Euclidean distances with a maximum nearest-atom distance.
8. The system of any of claims 1-7, wherein the amino acids have alpha-carbon atoms,
wherein the distances are nearest-alpha-carbon atom distances from the corresponding voxel
centers to nearest alpha-carbon atoms of the corresponding amino acids.

9. The system of any of claims 1-8, wherein the amino acids have beta-carbon atoms, wherein the distances are nearest-beta-carbon atom distances from the corresponding voxel centers to nearest beta-carbon atoms of the corresponding amino acids.
10. The system of any of claims 1-9, wherein the amino acids have backbone atoms, wherein the distances are nearest-backbone atom distances from the corresponding voxel centers to nearest backbone atoms of the corresponding amino acids.
11. The system of any of claims 1-10, wherein the amino acids have sidechain atoms, wherein the distances are nearest-sidechain atom distances from the corresponding voxel centers to nearest sidechain atoms of the corresponding amino acids.
12. The system of any of claims 1-11, further configured to encode, in the tensor, a nearest atom channel that specifies a distance from each voxel to a nearest atom, wherein the nearest atom is selected irrespective of the amino acids and atomic elements of the amino acids.
13. The system of claim 12, wherein the distance is a Euclidean distance.
14. The system of claim 12 or 13, wherein the distance is normalized by dividing the Euclidean distance with a maximum distance.
15. The system of any of claims 12-14, wherein the amino acids include non-standard amino acids.
16. The system of any of claims 1-15, wherein the tensor further includes an absentee atom channel that specifies atoms not found within a predefined radius of a voxel center, wherein the absentee atom channel is one-hot encoded.
17. The system of any of claims 1-16, wherein the tensor further includes a one-hot encoding of the alternative allele amino acid that is voxel-wise encoded to each of the amino acid-wise distance channels.
18. The system of any of claims 1-17, wherein the tensor further includes a reference allele amino acid of the protein.
19. The system of claim 18, wherein the tensor further includes a one-hot encoding of the reference allele amino acid that is voxel-wise encoded to each of the amino acid-wise distance channels.

20. The system of any of claims 1-19, wherein the tensor further includes evolutionary profiles that specify conservation levels of the amino acids across a plurality of species.
21. The system of any of claims 1-20, further comprising an evolutionary profiles generator that, for each of the voxels,
selects a nearest atom across the amino acids and the atom categories,
selects a pan-amino acid conservation frequencies sequence for a residue of an amino acid that includes the nearest atom, and
makes the pan-amino acid conservation frequencies sequence available as one of the evolutionary profiles.
22. The system of claim 21, wherein the pan-amino acid conservation frequencies sequence is configured for a particular position of the residue as observed in the plurality of species.
23. The system of claim 21 or 22, wherein the pan-amino acid conservation frequencies sequence specifies whether there is a missing conservation frequency for a particular amino acid.
24. The system of any of claims 21-23, wherein the evolutionary profiles generator, for each of the voxels,
selects respective nearest atoms in respective ones of the amino acids,
selects respective per-amino acid conservation frequencies for respective residues of the amino acids that include the nearest atoms, and
makes the per-amino acid conservation frequencies available as one of the evolutionary profiles.
25. The system of claim 24, wherein the per-amino acid conservation frequencies are configured for a particular position of the residues as observed in the plurality of species.
26. The system of claim 24 or 25, wherein the per-amino acid conservation frequencies specify whether there is a missing conservation frequency for a particular amino acid.
27. The system of any of claims 1-26, wherein the tensor further includes annotation channels for the amino acids, wherein the annotation channels are one-hot encoded in the tensor.
28. The system of any of claims 1-27, wherein the tensor further includes structure confidence channels for the amino acids that specify quality of respective structures of the amino acids.

29. A system, comprising:

memory storing atom category-wise distance channels for amino acids in a protein,
wherein the amino acids have atoms for a plurality of atom categories,
wherein atom categories in the plurality of atom categories specify atomic elements of the amino acids,
wherein each of the atom category-wise distance channels has voxel-wise distance values for voxels in a plurality of voxels, and
wherein the voxel-wise distance values specify distances from corresponding voxels in the plurality of voxels to atoms in corresponding atom categories in the plurality of atom categories; and
a pathogenicity determination engine configured to
process a tensor that includes the atom category-wise distance channels and an alternative allele amino acid of the protein expressed by a variant, and
determine a pathogenicity of the variant based at least in part on the tensor.

30. A computer-implemented method, comprising:

storing amino acid-wise distance channels for a plurality of amino acids in a protein,
wherein each of the amino acid-wise distance channels has voxel-wise distance values for voxels in a plurality of voxels, and
wherein the voxel-wise distance values specify distances from corresponding voxels in the plurality of voxels to atoms of corresponding amino acids in the plurality of amino acids;
processing a tensor that includes the amino acid-wise distance channels and an alternative allele of the protein expressed by a variant; and
determining a pathogenicity of the variant based at least in part on the tensor.