CLAIMS

1. A system comprising:

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,

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;

an alternative allele encoder that encodes an alternative allele amino acid to each voxel in the three-dimensional grid of voxels, wherein 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;

an evolutionary conservation encoder that encodes an evolutionary conservation sequence to each voxel in the three-dimensional grid of voxels,

wherein the evolutionary conservation sequence 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

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, and

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

2. The system of claim 1, wherein the voxelizer centers 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 system of claim ~~1 or~~ 2, wherein the voxelizer centers 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.

4. The system of ~~any of claims 1-3~~claim 3, 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.

5. The system of any of claims 1-4, wherein the distances from corresponding voxels to atoms 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 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~~6. The system of any of claims ~~1-7~~2-5, wherein the reference amino acids have alpha-carbon atoms, wherein the distances from corresponding voxels to the atoms are nearest-alpha-carbon atom distances from ~~the~~ corresponding voxel centers to nearest alpha-carbon atoms of the corresponding reference amino acids.

~~9~~7. The system of any of claims 1-6~~7~~, wherein the reference amino acids have beta-carbon atoms, wherein the distances from corresponding voxels to atoms are nearest-beta-carbon atom distances from ~~the~~ corresponding voxel centers to nearest beta-carbon atoms of the corresponding reference amino acids.

~~10~~8. The system of any of claims 1-~~7~~6, wherein the reference amino acids have backbone atoms, wherein the distances from corresponding voxels to atoms are nearest-backbone atom distances from ~~the~~ corresponding voxel centers to nearest backbone atoms of the corresponding reference amino acids.

~~11~~9. The system of any of claims 1-~~7~~6, wherein the reference amino acids have sidechain atoms, wherein the distances from corresponding voxels to atoms are nearest-sidechain atom distances from ~~the~~ corresponding voxel centers to nearest sidechain atoms of the corresponding reference amino acids.

~~12~~10. The system of any of claims 1-9~~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~~ an amino acid~~s~~ to which the nearest atom belongs and atomic elements of the amino acid~~s~~.

~~13. The system of any of claims 1-12, wherein the distance is a Euclidean distance.~~

~~14. The system of any of claims 1-13, wherein the distance is normalized by dividing the Euclidean distance with a maximum distance.~~

~~15. The system of any of claims 1-14, 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.~~

~~17. The system of claim 16, wherein the absentee atom channel is one-hot encoded.~~

~~18~~11. The system of any of claims 1-10~~17~~, further comprising a reference allele encoder that voxel-wise encodes a reference allele amino acid to each voxel in the three-dimensional grid of voxels.

~~19~~12. The system of claim ~~18~~11, wherein 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~~13. The system of any of claims 1-12~~19~~, wherein the amino acid-specific conservation frequencies specify conservation levels of respective amino acids across the plurality of species.

~~21. The system of any of claims 1-20, wherein the evolutionary conservation encoder~~

~~selects a nearest atom to the corresponding voxel across the reference amino acids and the atom categories,~~

~~selects pan-amino acid conservation frequencies for a residue of a reference amino acid that includes the nearest atom, and~~

~~uses a three-dimensional representation of the pan-amino acid conservation frequencies as the evolutionary conservation sequence.~~

~~22. The system of claim 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 system of claim 21 or 22, wherein the pan-amino acid conservation frequencies specify whether there is a missing conservation frequency for a particular reference amino acid.~~

~~24. The system of any of claims 1-23, wherein the evolutionary conservation encoder~~

~~selects respective nearest atoms to the corresponding voxel in respective ones of the reference amino acids,~~

~~selects respective per-amino acid conservation frequencies for respective residues of the reference amino acids that include the nearest atoms, and~~

~~uses a three-dimensional representation of the per-amino acid conservation frequencies as the evolutionary conservation sequence.~~

~~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 reference amino acid.~~

14~~27~~. The system of any of claims 1-13~~26~~, further comprising an annotations encoder that voxel-wise encodes one or more annotation channels to each voxel in the three-dimensional grid of voxels, and wherein the one or more annotation channels are three-dimensional representations of a one-hot encoding of residue annotations.

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

~~29. The system of claim 27 or 28, 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 system of any of claims 27-29, wherein the annotation channels are sites annotations that include active site, metal binding, binding site, and site.~~

~~31. The system of any of claims 27-30, 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 system of any of claims 27-31, wherein the annotation channels are secondary structure annotations that include helix, turn, and beta strand.~~

~~33. The system of any of claims 27-32, wherein the annotation channels are experimental information annotations that include mutagenesis, sequence uncertainty, sequence conflict, non-adjacent residues, and non-terminal residue.~~

~~34~~15. The system of any of claims 1-~~33~~14, further comprising a structure confidence encoder that voxel-wise encodes one or more structure confidence channels to each voxel in the three-dimensional grid of voxels, and wherein the one or more structure confidence channels are three-dimensional representations of confidence scores that specify quality of respective residue structures.

16. 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.

17. The computer-implemented method of claim 16, further comprising:

selecting a nearest atom to the corresponding voxel across the reference amino acids and 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.

18. The computer-implemented method of claim 17, wherein the pan-amino acid conservation frequencies are configured for a particular position of the residue as observed in the plurality of species.

19. The computer-implemented method of any of claims 16, further comprising:

selecting respective nearest atoms to the corresponding voxel in respective reference amino acids,

selecting respective per-amino acid conservation frequencies for respective residues of the respective reference amino acids that include the respective nearest atoms, and

using a three-dimensional representation of the respective per-amino acid conservation frequencies as the evolutionary conservation channel.

20. A non-transitory computer readable medium storing instructions that, when executed by at least a processor, cause a system to performing actions 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.

~~35. The system of claim 34, wherein the structure confidence channels are global model quality estimations (GMQEs).~~

~~36. The system of claim 34 or 35, wherein the structure confidence channels are qualitative model energy analysis (QMEAN) scores.~~

~~37. The system of any of claims 34-36, 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 system of any of claims 34-37, 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 system of any of claims 34-38, wherein the structure confidence channels are template modeling scores of the aligned template structures.~~

~~40. The system 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 system of any of claims 1-40, further comprising an atoms rotation engine that rotates the atoms before the amino acid-wise distance channels are generated.~~

~~42. The system of any of claims 1-41, wherein the convolutional neural network uses 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.~~

~~43. The system of claim 42, wherein the 1 x 1 x 1 convolutions and the 3 x 3 x 3 convolutions are the three-dimensional convolutions.~~

~~44. The system of claim 42 or 43, 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 system of claim 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 system of any of claims 1-45, wherein the convolutional neural network is an attention-based neural network.~~

~~47. The system of any of claims 1-46, wherein the tensor includes the amino acid-wise distance channels further encoded with the reference allele amino acid.~~

~~48. The system of any of claims 27-47, wherein the tensor includes the amino acid-wise distance channels further encoded with the annotation channels.~~

49. The system of any of claims 34-48, wherein the tensor includes the amino acid-wise distance channels further encoded with the structure confidence channels.

50. A system, comprising:

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,

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;

an alternative allele encoder that encodes an alternative allele amino acid to each voxel in the three-dimensional grid of voxels,

wherein 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;

an evolutionary conservation encoder that encodes an evolutionary conservation sequence to each voxel in the three-dimensional grid of voxels,

wherein the evolutionary conservation sequence 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

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

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

51. A system, comprising:

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,

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;

an alternative allele encoder that encodes an alternative allele amino acid to each voxel in the three-dimensional grid of voxels,

wherein 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;

an evolutionary conservation encoder that encodes an evolutionary conservation sequence to each voxel in the three-dimensional grid of voxels,

~~wherein the evolutionary conservation sequence 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~~

~~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.~~

~~52. A system, comprising:~~

~~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,~~

~~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;~~

~~an alternative allele encoder that encodes an alternative allele amino acid to each voxel in the three-dimensional grid of voxels,~~

~~wherein 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;~~

~~an evolutionary conservation encoder that encodes an evolutionary conservation sequence to each voxel in the three-dimensional grid of voxels,~~

~~wherein the evolutionary conservation sequence 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~~

~~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.~~