CLAIMS

1. A system, comprising:

memory storing amino acid-wise distance channels for a plurality of amino acids in an amino acid sequence of 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 neural network-based variant pathogenicity classifier, running on at least one processor coupled to the memory, wherein the neural network-based variant pathogenicity classifier is trained to:

process as input a tensor that includes the amino acid-wise distance channels and an alternative allele amino acid of the protein expressed by a variant, and

classify the variant as benign or pathogenic based at least in part on the tensor.

2. The system of claim 1, further comprising a distance channels generator, running on at least one processor coupled to the memory, that centers a voxel grid of the voxels on an alpha-carbon atom of respective residues of the corresponding amino acids and calculates the voxel-wise distance values by specifying a distance between centers of the voxels in the voxel grid and the atoms of the corresponding 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 that corresponds to at least one variant amino acid in the protein.

4. The system of claim 3, further configured to encode, in the tensor, a directionality of the corresponding amino acids and a position of the particular amino acid by multiplying, with a directionality parameter, voxel-wise distance values for preceding amino acids that precede the particular amino acid.

5. The system 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 system of claim 5, wherein the corresponding 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.

7. The system of claim 5, wherein the corresponding 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.

8. The system of claim 5, wherein the corresponding 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.

9. The system of claim 3, 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 an amino acid to which the nearest atom belongs and atomic elements of the amino acid.

10. The system of claim 1, wherein the tensor further includes evolutionary profiles that specify conservation levels of the corresponding amino acids across a plurality of species with sequences that are homologous to the amino acid sequence of the protein.

11. The system of claim 10, further comprising an evolutionary profiles generator, running on at least one processor coupled to the memory, that, for each of the voxels,

uses a multi-sequence alignment to determine pan-amino acid conservation frequencies,

selects a nearest atom across the plurality of amino acids and atom categories,

selects a pan-amino acid conservation frequencies sequence for a residue of an amino acid that includes the nearest atom,

voxelizes the pan-amino acid conservation frequencies for the residue of the amino acid, and

makes the pan-amino acid conservation frequencies sequence available as one of the evolutionary profiles.

12. The system of claim 11, wherein the pan-amino acid conservation frequencies sequence is configured for a particular position of the residue as observed in the plurality of species.

13. The system of claim 11, wherein the evolutionary profiles generator, for each of the voxels,

uses the multi-sequence alignment to determine per-amino acid conservation frequencies,

selects respective nearest atoms in respective ones of the plurality of amino acids,

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

voxelizes the per-amino acid conservation frequencies for the respective residues of the plurality of amino acids, and

makes the per-amino acid conservation frequencies available as one of the evolutionary profiles.

14. A computer-implemented method, comprising:

storing amino acid-wise distance channels for a plurality of amino acids in an amino acid sequence of 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 as input a tensor that includes the amino acid-wise distance channels and an alternative allele amino acid of the protein expressed by a variant; and

classifying the variant as benign or pathogenic based at least in part on the tensor.

15. The computer-implemented method of claim 14, 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.

16. The computer-implemented method of claim 14, 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.

17. The computer-implemented method of claim 14, wherein the tensor further includes a reference allele amino acid in the amino acid sequence of the protein.

18. The computer-implemented method of claim 17, 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.

19. The computer-implemented method of claim 14, wherein the tensor further includes one or more of:

annotation channels for the corresponding amino acids that annotate characteristics of the corresponding amino acids, wherein the annotation channels are one-hot encoded in the tensor; or

structure confidence channels for the corresponding amino acids that specify quality of respective structures of the corresponding amino acids.

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

storing amino acid-wise distance channels for a plurality of amino acids in an amino acid sequence of 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 as input a tensor that includes the amino acid-wise distance channels and an alternative allele amino acid of the protein expressed by a variant; and

classifying the variant as benign or pathogenic based at least in part on the tensor.