CLAIMS

Claim 1. A system, comprising:

one or more processors; and

a non-transitory storage medium comprising computer instructions that, when executed by the one or more processors, cause the system to:

input, into a pathogenicity prediction neural network:

an encoded reference amino acid sequence for a protein;

an encoded alternative amino acid sequence for the protein comprising a variant amino acid;

at least a conservation profile generated from amino acids of related species aligned with a reference amino acid sequence corresponding to the encoded reference amino acid sequence; and

based on the encoded alternative amino acid sequence, the encoded reference amino acid sequence, and at least the conservation profile, generate, as an output of the pathogenicity prediction neural network, a pathogenicity prediction indicating a likelihood that the variant amino acid is benign or pathogenic.

Claim 2. The system of claim 1, wherein the conservation profile comprises a position-specific frequency matrix generated from a multiple sequence alignment of the related species.

Claim 3. The system of claim 1, wherein the conservation profile comprises a position-specific score matrix generated from a multiple sequence alignment of the related species.

Claim 4. The system of claim 1, wherein at least the conservation profile comprises one or more of:

a primate conservation profile generated from amino acids of primate species aligned with the reference amino acid sequence;

a mammal conservation profile generated from amino acids from mammal species aligned with the reference amino acid sequence; or

a vertebrate conservation profile generated from amino acids from vertebrate species aligned with the reference amino acid sequence.

Claim 5. The system of claim 1, further comprising computer instructions that, when executed by the one or more processors, cause the system to input, into the pathogenicity prediction neural

network, a three-state secondary structure prediction for each amino acid in the encoded reference amino acid sequence.

Claim 6. The system of claim 5, further comprising computer instructions that, when executed by the one or more processors, cause the system to generate, utilizing a secondary structure neural network, the three-state secondary structure prediction based on the conservation profile.

Claim 7. The system of claim 1, further comprising computer instructions that, when executed by the one or more processors, cause the system to input, into the pathogenicity prediction neural network, a three-state solvent accessibility prediction for each amino acid in the encoded reference amino acid sequence.

Claim 8. The system of claim 7, further comprising computer instructions that, when executed by the one or more processors, cause the system to generate, utilizing a solvent accessibility neural network, the three-state solvent accessibility prediction based on the conservation profile.

Claim 9. The system of claim 1, wherein the variant amino acid of the encoded alternative amino acid sequence is caused by a nucleotide substitution of a nucleotide within the reference amino acid sequence.

Claim 10. The system of claim 1, further comprising computer instructions that, when executed by the one or more processors, cause the system to:

input the encoded reference amino acid sequence into a first convolutional layer of the pathogenicity prediction neural network;

input the encoded alternative amino acid sequence into a second convolutional layer of the pathogenicity prediction neural network; and

input the conservation profile into a third convolutional layer of the pathogenicity prediction neural network.

Claim 11. A computer-implemented method comprising:

inputting, into a pathogenicity prediction neural network:

an encoded reference amino acid sequence for a protein;

an encoded alternative amino acid sequence for the protein comprising a variant amino acid;

at least a conservation profile generated from amino acids of related species aligned with a reference amino acid sequence corresponding to the encoded reference amino acid sequence; and

based on the encoded alternative amino acid sequence, the encoded reference amino acid sequence, and at least the conservation profile, generate, as an output of the pathogenicity prediction neural network, a pathogenicity prediction indicating a likelihood that the variant amino acid is benign or pathogenic.

Claim 12. The computer-implemented method of claim 11, wherein the conservation profile comprises a position-specific frequency matrix generated from a multiple sequence alignment of the related species.

Claim 13. The computer-implemented method of claim 11, wherein the conservation profile comprises a position-specific score matrix generated from a multiple sequence alignment of the related species.

Claim 14. The computer-implemented method of claim 11, wherein at least the conservation profile comprises one or more of:

a primate conservation profile generated from amino acids of primate species aligned with the reference amino acid sequence;

a mammal conservation profile generated from amino acids from mammal species aligned with the reference amino acid sequence; or

a vertebrate conservation profile generated from amino acids from vertebrate species aligned with the reference amino acid sequence.

Claim 15. The computer-implemented method of claim 11, further comprising inputting, into the pathogenicity prediction neural network, a three-state secondary structure prediction for each amino acid in the encoded reference amino acid sequence.

Claim 16. The computer-implemented method of claim 15, further comprising generating, utilizing a secondary structure neural network, the three-state secondary structure prediction based on the conservation profile.

Claim 17. The computer-implemented method of claim 11, further comprising inputting, into the pathogenicity prediction neural network, a three-state solvent accessibility prediction for each amino acid in the encoded reference amino acid sequence.

Claim 18. The computer-implemented method of claim 17, further comprising generating, utilizing a solvent accessibility neural network, the three-state solvent accessibility prediction based on the conservation profile.

Claim 19. The computer-implemented method of claim 17, wherein the variant amino acid of the encoded alternative amino acid sequence is caused by a nucleotide substitution of a nucleotide within the reference amino acid sequence.