weights to a small number (few dozen) of hand-engineered features, such as the number of rotatable bonds, hydrogen bond donor-acceptor pairs, aromatic stacks, electrostatics, steric complementarity or strain, or solvent-accessible hydrophobic area. The development of these features requires expert knowledge and extensive manual tuning, yet any such feature will necessarily be a limited approximation since, as discussed above, even expert chemists cannot consistently disentangle the forces governing molecular interactions. Force-field based scoring functions are designed to be computationally efficient, which requires approximations to theoretical results from gas phase predictions. For example, such systems ignore or crudely approximate the important mediation of field strength by solvent molecules.

[0006] For the reasons stated, prior art binding affinity prediction solutions may not be as accurate as is desirable.

[0007] There is a need for solutions that provide more accurate and/or more efficient binding affinity prediction, or at least provide an alternate way to provide binding affinity prediction.

SUMMARY

[0008] According to one aspect, there is provided

a system for predicting binding affinity. The system includes: at least one electronic datastore configured to store records reflecting input data; a data encoder module configured to perform a method including: constructing a data structure from the stored input data in the least one electronic datastore, the data structure providing a geometric representation of binding input features, wherein the constructing includes: augmenting the stored input data by generating additional instances of the stored input data through application of one or more pre-defined transformations of the stored input data wherein the one or more pre-defined transformations of the stored input data include a combination of rotation, translation, and mirror operators in any combination of the three X, Y, and Z planes of the stored input data thereby resulting in geometric data, and truncating each instance of the stored input data in the geometric data to fit within a bounding box; and populating the data structure by encoding data relating to the one molecule and the one target protein in the geometric data for each instance of the stored input data fitted to the bounding box in the geometric data, wherein the populating includes: partitioning the instance of the stored input data fitted to the bounding box in the geometric data into a corresponding voxel grid by transforming the instance of the input data fitted to the bounding box into a discrete set of regularly-spaced cubic voxels using a spacing that yields an integer number of voxels, wherein each respective voxel in the respective voxel grid includes one or more input channels that encode a type of atom represented by the respective voxel in the respective voxel grid; and a predictive model module configured to: apply a predictive model to the data structure to generate an indicator of a binding affinity for the one molecule to the one target protein represented by the data structure, wherein the predictive model is a deep neural network or convolutional neural network including a convolutional filter layer comprising a set of function. computation elements each of which takes as input a region of spatially contiguous

voxels from a respective voxel grid in the data structure or convolutional functions thereof, and the indicator of the binding affinity is an output of the predictive model.

[0009] There is also provided a computer-implemented method for predicting binding affinity of one molecule to one target protein. The method includes: storing, in at least one electronic datastore, records reflecting input data; constructing, at a data encoder module, a data structure from the stored input data in the least one electronic datastore, the data structure providing a geometric representation of binding input features, wherein the constructing includes augmenting the stored input data by generating additional instances of the stored input data through application of one or more pre-defined transformations of the stored input data wherein the one or more pre-defined transformations of the stored input data include a combination of rotation, translation, and mirror operators in any combination of the three X, Y, and Z planes of the stored input data thereby resulting in geometric data, and truncating each instance of the stored input data in the geometric data to fit within a bounding box; populating, at the at least one processor, the data structure by encoding data relating to the one molecule and the one target protein in the geometric data by, for each instance of the stored input data fitted to the bounding box in the geometric data, wherein the populating includes partitioning the instance of the stored input data fitted to the bounding box in the geometric data into a corresponding voxel grid by transforming the instance of the input data fitted to the bounding box into a discrete set of regularly-spaced cubic voxels using a spacing that yields an integer number of voxels, wherein each respective voxel in the respective voxel grid includes one or more input channels that encode a type of atom represented by the respective voxel in the respective voxel grid; and applying, at the at least one processor, a predictive model to the data structure to generate an indicator of a binding affinity for the one molecule to the one target protein represented by the data structure, wherein the predictive model is a deep neural network or convolutional neural network comprising a convolutional filter layer comprising a set of function computation elements each of which takes as input a region of spatially contiguous voxels from a respective voxel grid in the data structure or convolutional functions thereof, and the indicator of the binding affinity is an output of the predictive model.

Claims

1. A system for predicting binding affinity of one molecule to one target protein, the system comprising:

at least one electronic datastore configured to store records reflecting input data;

a data encoder module configured to perform a method comprising:

constructing a data structure from the stored input data in the least one electronic datastore, the data structure providing a geometric representation of binding input features, wherein the constructing comprises:

augmenting the stored input data by generating additional instances of the stored input data through application of one or more pre-defined transformations of the stored input data wherein the one or more pre-defined transformations of the stored input data include a combination of rotation, translation, and mirror operators in any combination of the three X, Y, and Z planes of the stored input data thereby resulting in geometric data; and

truncating each instance of the stored input data in the geometric data to fit within a bounding box; and

populating the data structure by encoding data relating to the one molecule and the one target protein in the geometric data for each instance of the stored input data fitted to the bounding box in the geometric data, wherein the populating comprises:

partitioning the instance of the stored input data fitted to the bounding box in the geometric data into a corresponding voxel grid by transforming the instance of the input data fitted to the bounding box into a discrete set of regularly-spaced cubic voxels using a spacing that yields an integer number of voxels, wherein each respective voxel in the respective voxel grid comprises one or more input channels that encode a type of atom represented by the respective voxel in the respective voxel grid; and

a predictive model module configured to:

apply a predictive model to the data structure to generate an indicator of a binding affinity for the one molecule to the one target protein represented by the data structure, wherein the predictive model is a deep neural network or convolutional

neural network comprising a convolutional filter layer comprising a set of function computation elements each of which takes as input a region of spatially contiguous voxels from a respective voxel grid in the data structure or convolutional functions thereof, and the indicator of the binding affinity is an output of the predictive model.

2. The system of claim 1, wherein the geometric representation comprises a plurality of coordinates of atom centres.

3. The system of claim 1 or 2, wherein the geometric representation has a resolution between 0.1Å and 10Å.

4. The system of claim 3, wherein the geometric representation has a resolution of 1Å.

5. The system of any one of claims 1 to 4, wherein the predictive model module is configured to train the predictive model over time.

6. The system of any one of claims 1 to 5, wherein the records reflecting input data include records reflecting input data relating to at least one of atoms, ligand molecules, organic cofactors, inorganic cofactors, and proteins.

7. The system of any of claims 1 to 6, wherein the one or more input channels is a plurality of input channels and wherein the plurality of input channels further includes an input channel that specifies partial charge, polarizability, electronegativity, solvent accessible space, or electron density represented by the respective voxel in the respective voxel grid.

8. The system of any of claims 1 to 6, wherein the atom represented by a respective voxel in the respective voxel grid only partially occupies the respective voxel and the input channel in the one or more input channels that encodes the type of atom represented by the respective voxel is partially turned on proportional to the amount the atom represented by the respective voxel occupies the respective voxel.

9. The system of any of claims 1 to 6, wherein more than one atom type shares the same input channel in the one or more input channels.

10. A computer-implemented method for predicting binding affinity of one molecule to one target protein, the method comprising:

storing, in at least one electronic datastore, records reflecting input data;

constructing, at a data encoder module, a data structure from the stored input data in the least one electronic datastore the data structure providing a geometric representation of binding input features, wherein the constructing comprises:

augmenting the stored input data by generating additional instances of the stored input data through application of one or more pre-defined transformations of the stored input data wherein the one or more pre-defined transformations of the stored input data include a combination of rotation, translation, and mirror operators in any combination of the three X, Y, and Z planes of the stored input data thereby resulting in geometric data, and

truncating each instance of the stored input data in the geometric data to fit within a bounding box; and

populating, at the at least one processor, the data structure by encoding data relating to the one molecule and the one target protein in the geometric data by, for each instance of the stored input data fitted to the bounding box in the geometric data, wherein the populating comprises:

partitioning the instance of the stored input data fitted to the bounding box in the geometric data into a corresponding voxel grid by transforming the instance of the input data fitted to the bounding box into a discrete set of regularly-spaced cubic voxels using a spacing that yields an integer number of voxels, wherein each respective voxel in the respective voxel grid comprises one or more input channels that encode a type of atom represented by the respective voxel in the respective voxel grid; and

applying, at the at least one processor, a predictive model to the data structure to generate an indicator of a binding affinity for the one molecule to the one target protein represented by the data structure, wherein

the predictive model is a deep neural network or convolutional neural network comprising a convolutional filter layer comprising a set of function computation

elements each of which takes as input a region of spatially contiguous voxels from a respective voxel grid in the data structure or convolutional functions thereof, and the indicator of the binding affinity is an output of the predictive model.

11. The method of claim 10, wherein the geometric representation comprises a plurality of coordinates of atom centres.

12. The method of claim 10 or 11, wherein the geometric representation has a resolution between 0.1Å and 10Å.

13. The method of claim 12, wherein the geometric representation has a resolution of 1Å.

14. The method of any one of claims 10 to 13, further comprising training (28) the predictive model over time.

15. The method of any one of claims 10 to 14, wherein the records of input data include records reflecting input data relating to at least one of atoms, ligand molecules, organic cofactors, inorganic cofactors, proteins.