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 computer system for characterization of a test object using spatial data, the computer system comprising: a graphical processing unit having a graphical processing memory; a general processor; and general memory addressable by the general processing unit, the general memory storing at least one program for execution by the at least one general processor, the at least one program comprising instructions for: (A) obtaining spatial coordinates for a target object; (B) modeling the test object with the target object in each pose of a plurality of different poses, thereby creating a plurality of voxel maps, wherein each respective voxel map in the plurality of voxel maps comprises the test object in a respective pose in the plurality of different poses; (C) unfolding each voxel map in the plurality of voxel maps into a corresponding vector, thereby creating a plurality of vectors, wherein each vector in the plurality of vectors is the same size; (D) providing the plurality of vectors to the graphical processing unit memory, wherein the graphical processing memory comprises a network architecture that includes an input layer for sequentially receiving the plurality of vectors, a plurality of convolutional layers, and a scorer, wherein the plurality of convolutional layers includes an initial convolutional layer and a final convolutional layer, each layer in the plurality of convolutional layers is associated with a different set of weights, responsive to input of a respective vector in the plurality of vectors, the input layer feeds a first plurality of values into the initial convolutional layer as a first function of values in the respective vector,

wherein the first function is computed using the graphical processing unit, each respective convolutional layer, other than the final convolutional layer, feeds intermediate values, as a respective second function of (i) the different set of weights associated with the respective convolutional layer and (ii) input values received by the respective convolutional layer, into another convolutional layer in the plurality of convolutional layers, wherein the second function is computed using the graphical processing unit, and the final convolutional layer feeds final values, as a third function of (i) the different set of weights associated with the final convolutional layer and (ii) input values received by the final convolutional layer that is computed using the graphical processing unit, into the scorer; (E) obtaining a plurality of scores from the scorer, wherein each score in the plurality of scores corresponds to the input of a vector in the plurality of vectors into the input layer; and (F) using the plurality of scores to characterize the test object. There is also provided a system for predicting binding affinity. The system includes: at least one electronic datastore configured to store records reflecting input data; an encoder module configured to: construct a data structure providing a geometric representation of binding input features; populate the data structure by encoding data relating to at least one molecule and at least one target protein, the data for encoding selected from the stored input data; and a prediction module configured to: apply a predictive model to the data structure to generate an indicator of a binding affinity for at least one molecule to at least one target protein.

[0009] There is also provided a computer-implemented method for predicting binding affinity. The method includes: storing, in at least one electronic datastore, records reflecting input data; constructing, at at least one processor, a data structure providing a geometric representation of binding input features; populating, at the at least one processor, the data structure by encoding data relating to at least one molecule and at least one target protein, said data for encoding selected from the stored input data; and applying, at the at least one processor, a predictive model to the data structure to generate an indicator of a binding affinity for at least one molecule to at least one target protein.

Claims

1. A computer system for characterization of a test object using spatial data, the computer system comprising:

a graphical processing unit having a graphical processing memory;

a general processor; and

general memory addressable by the general processing unit, the general memory storing at least one program for execution by the at least one general processor, the at least one program comprising instructions for:

(A) obtaining spatial coordinates for a target object;

(B) modeling the test object with the target object in each pose of a plurality of different poses, thereby creating a plurality of voxel maps, wherein each respective voxel map in the plurality of voxel maps comprises the test object in a respective pose in the plurality of different poses;

(C) unfolding each voxel map in the plurality of voxel maps into a corresponding vector, thereby creating a plurality of vectors, wherein each vector in the plurality of vectors is the same size;

(D) providing the plurality of vectors to the graphical processing unit memory, wherein the graphical processing memory comprises a network architecture that includes an input layer for sequentially receiving the plurality of vectors, a plurality of convolutional layers, and a scorer, wherein

the plurality of convolutional layers includes an initial convolutional layer and a final convolutional layer,

each layer in the plurality of convolutional layers is associated with a different set of weights,

responsive to input of a respective vector in the plurality of vectors, the input layer feeds a first plurality of values into the initial convolutional layer as a first function of values in the respective vector, wherein the first function is computed using the graphical processing unit,

each respective convolutional layer, other than the final convolutional layer, feeds intermediate values, as a respective second function of (i) the different set of weights associated with the respective convolutional layer and (ii) input values received by the respective convolutional layer, into another convolutional layer in the plurality of

convolutional layers, wherein the second function is computed using the graphical processing unit, and

the final convolutional layer feeds final values, as a third function of (i) the different set of weights associated with the final convolutional layer and (ii) input values received by the final convolutional layer that is computed using the graphical processing unit, into the scorer;

(E) obtaining a plurality of scores from the scorer, wherein each score in the plurality of scores corresponds to the input of a vector in the plurality of vectors into the input layer; and

(F) using the plurality of scores to characterize the test object.

2. The computer system of claim 1, wherein the scorer comprises a nearest neighbor analysis or a support vector machine,

3. The computer system of claim 1, wherein the plurality of different poses is obtained using a docking scoring function in Monte Carlo sampling

4. The computer system of claim 1, wherein the target object is a polymer.

5. The computer system of claim 1, wherein the target object is a polymer and the spatial coordinates are a set of three-dimensional coordinates $\{x_1, \ldots, x_N\}$ for a crystal structure of the polymer

6. The computer system of claim 1, wherein the test object is a chemical compound and using the plurality of scores to characterize the test object comprises taking a measure of central tendency of the plurality of scores, wherein

when the measure of central tendency satisfies a predetermined threshold value or predetermined threshold value range, the test object is deemed to have a first classification, and

when the measure of central tendency fails to satisfy the predetermined threshold value or predetermined threshold value range, the test object is deemed to have a second classification.

7. The computer system of claim 1, wherein the using the plurality of scores to characterize the test object comprises taking a weighted average of the plurality of scores, wherein

when the weighted average satisfies a predetermined threshold value or predetermined threshold value range, the test object is deemed to have a first classification, and

when the weighted average fails to satisfy the predetermined threshold value or predetermined threshold value range, the test object is deemed to have a second classification.

8. The computer system of claim 33 wherein the weighted average is a Boltzman average of the plurality of scores.

9. The computer system of claim 1, wherein the using the plurality of scores to characterize the test object comprises taking a weighted average of the plurality of scores, wherein when the weighted average satisfies a respective threshold value range in a plurality of threshold value ranges, the test object is deemed to have a respective classification in a plurality of a respective classifications that uniquely corresponds to the respective threshold value range.

10. The computer system of claim 1, wherein the target object is a polymer with an active site, the test object is a chemical composition, and the modeling comprises docking the test object into the active site of the polymer.