A geometric deep learning approach to predict binding conformations of bioactive molecules

Oscar Méndez-Lucio, *,[a] Mazen Ahmad,[a] Ehecatl Antonio del Rio-Chanona,[b] Jörg Kurt Wegner *,[a]

[a] Generative AI Team, High Dimensional Biology and Discovery Data Sciences, Janssen Research & Development, Janssen Pharmaceutica N.V., Turnhoutseweg 30, Beerse B-2340, Belgium

[b] Centre for Process Systems Engineering (CPSE), Department of Chemical Engineering, Imperial College London, UK

* E-mail: omendezl@its.jnj.com, jwegner@its.jnj.com

Abstract

Understanding the interactions formed between a ligand and its molecular target is key to guide the optimization of molecules. Different experimental and computational methods have been key to understand better these intermolecular interactions. Herein, we report a method based on geometric deep learning that is capable of predicting the binding conformations of ligands to protein targets. Concretely, the model learns a statistical potential based on distance likelihood which is tailor-made for each ligand-target pair. This potential can be coupled with global optimization algorithms to reproduce experimental binding conformations of ligands. We show that the potential based on distance likelihood described in this paper performs similar or better than well-established scoring functions for docking and screening tasks. Overall, this method represents an example of how artificial intelligence can be used to improve structure-based drug design.

Introduction

There is no doubt that drug design is a challenging task. One of the difficulties arises from the fact that only a small portion of the large chemical space (circa 10^{60} drug-like molecules[1,2]) will bind to a specific biological target resulting in a therapeutical effect. In this context, knowing up-front the biological target and its three-dimensional structure seems to be associated with higher success rates[3]. To a very large extent, this success results from the use of experimental and computational methods that can help understand the key interactions between a ligand and its molecular target to guide the optimization of molecules. In fact, it is known that these intermolecular interactions are a key factor driving drug potency and selectivity[4]. Experimental methods such as X-ray diffraction, NMR crystallography and more recently Cryo-EM have been of paramount importance for drug discovery projects to explore and understand these intermolecular interactions[3,5]. In a similar way, computational methods have also played an important role since they allow to virtually study compounds that have not been synthesized yet. In particular, molecular docking has been recently used to virtually screen ultra-large compound libraries[6,7], although other methods such as molecular dynamics are also commonly used for drug discovery.

In the recent years, the explosion of experimental structural data has also allowed the application of machine learning and artificial intelligence to study ligand-target interactions. For example, machine learning has been successfully applied to identify regions of a protein where a ligand can directly bind[8–10]. Additionally, a wide a variety of methods have been developed to predict binding affinity from the three-dimensional structure of a ligand-target complex[11,12]. Many of these methods make use of engineered descriptors that capture the main ligand-target interactions which can be fed into a predictive algorithm[13–16], while others directly use convolutional neural networks (CNNs)[17–20] or graph convolutional neural networks (GNNs)[21,22] for the prediction task.

Despite the need for more computationally efficient methods for structure-based design, there are few efforts to accelerate or improve the structure prediction of a bound ligand by using artificial intelligence or machine learning. Most of current artificial intelligence methods applied to structure-based drug discovery rely on the 3D structure of a ligand-target complex previously obtained either by experimental or computational approaches. Herein, we report DeepDock, a method based on geometric deep learning that is capable of predicting the binding conformation of ligands to protein targets. For this, the method learns a statistical potential based on distance likelihood which is tailor-made for each ligand-target. Statistical potentials have been used to sample small molecule conformations in an efficient manner[23–26]. In particular, the work of Klebe and Mietzner[26] pave the road for using statistical potentials on torsion angles to generate molecular conformations. Nonetheless, learning these potentials using deep learning confers some advantages such as taking larger portions of the molecule into account or inferring the potential for a combination of atoms not included in the training set. Similar advantages have been observed in deep learning potentials recently used for protein structure prediction[27]. In this work we show that the proposed potential based on distance likelihood performs similar or better than well-established scoring function for docking and screening tasks. In addition, it can be coupled with global optimization algorithms to reproduce experimental binding conformations of ligands.

Results

Learning a customized potential based on distance likelihood

Contrary to most computational methods that predict the binding conformation of a ligand (e.g., docking), our geometric deep learning approach learns a potential that is specific for each ligand-target complex and which global minimum corresponds to the optimal binding conformation. To learn this potential we trained DeepDock using experimental three-dimensional data of ligands bound to protein targets (e.g., X-ray crystallography), extracted from the PDBbind database[28]. DeepDock is a neural network responsible for two main tasks: feature extraction from the input data and identify key ligand-target interactions, as shown in Fig. 1.

In a first step, the neural network extracts relevant representations of the input data, namely ligand and target structures. Our approach directly uses the molecular surface of the binding site in the form of a polygon mesh. In this mesh, a collection of nodes, edges and faces defines the shape of the molecular surface as a polygon (Fig. 1a). Moreover, the nodes also contain features encoding chemical and topological information at that specific point of the molecular surface, whereas edge features encode the connectivity between nodes. In a similar way, ligands are represented as a two-dimensional undirected graph, where atoms are designated by nodes and bonds are represented by edges (Fig. 1b). In this case, node and edge features encode the atom and bond types, respectively. Both, the target mesh and the ligand graph, are processed by independent residual graph convolutional neural networks (GNNs). Through this procedure, the processed node features not only contain information of an individual atom or point in the molecular surface, but also have information about the other nodes around them. In other words, the processed atom features encode all the atomic environment around a specific atom, whereas the target features encode a patch of the molecular surface around a specific point. A more detailed description of the feature extraction can be found in the Methods section.

In a following step, the processed node features from the target and ligand were combined in order to model the interaction of the ligand with the target (Fig. 1c). For this, we concatenate all node features in a pairwise manner meaning each ligand atom will be paired with each node in the molecular surface of the target. In a final step, these concatenated features are processed by mixture density network (MDN)[29]. This network is composed by a feed forward neural network that predict a set of means, standard deviations and mixing coefficients needed to parametrize a mixture density model for each ligand-target node-pair. The mixture model represents the conditional probability density function of distance for any given ligand-target node pair $P\left(d_{ij}\middle|v_i^l, v_j^t\right)$. In other words, using this probability density function we can estimate the likelihood of finding ligand node i separated from a target node j by any distance d_{ij}. Using an MDN is essential since it allows to learn the distribution of distance data i.e., the distribution of all values d_{ij} separating ligand node i from target node j observed in the training set. On the contrary, a simple feed forward neural trained by minimizing the error (e.g., using RMSE or MAE as loss function) only approximates the average distance $\overline{d_{ij}}$ that separates ligand node i from a target node j in the training data. This would be inadequate to model the multi-valued nature of the data used to train the model since d_{ij} can take an infinite number of valid values, but some of them are more likely to be observed than others.

Finally, the probability density functions of all pairwise combinations of ligand atoms and points in the molecular surface are aggregated into a statistical potential. This is simply done by adding up all the independent negative log likelihood values calculated for each ligand-target pair. This results in an energy function that can be minimized, and whose minimum correspond to the conformation of ligand in which all atoms are separated from all points in the target surface by the most likely distance.

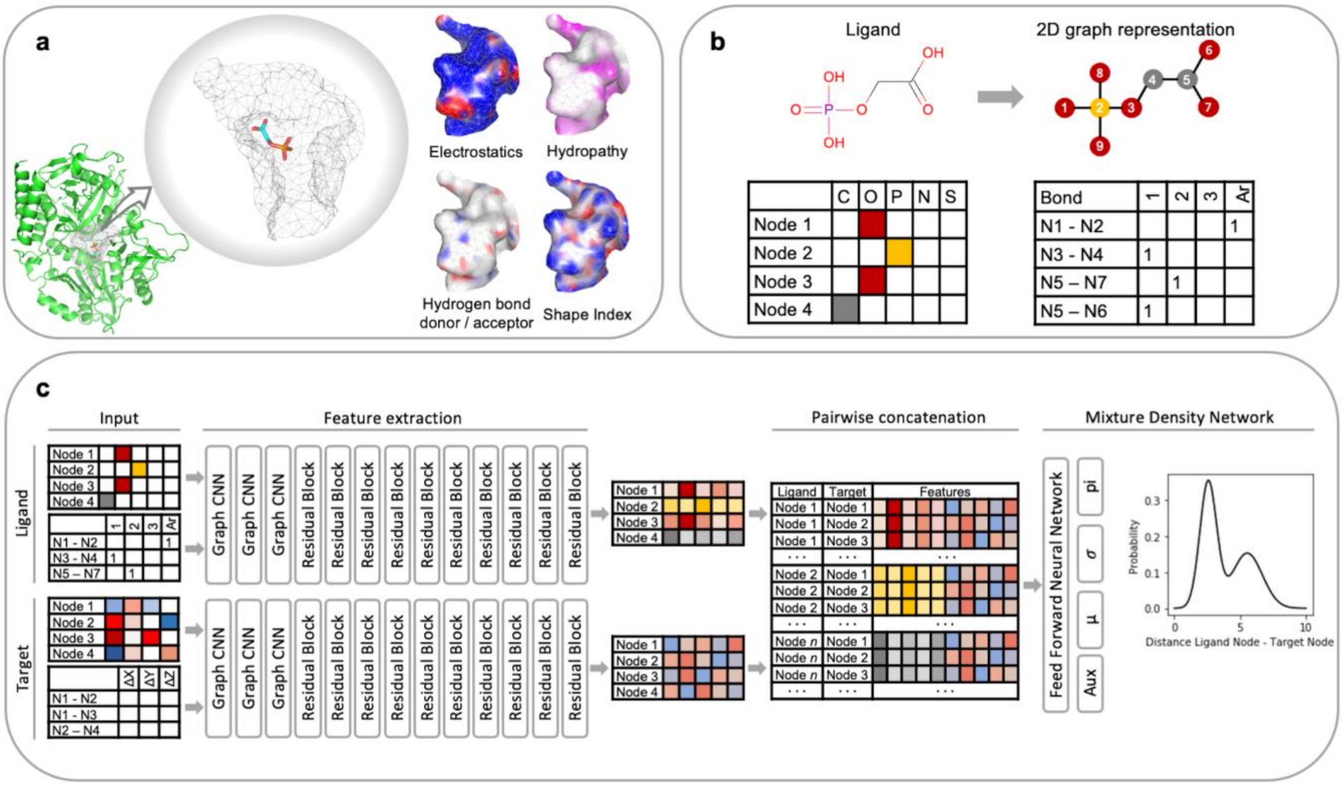

Fig. 1 Deep learning model used to learn a potential to predict binding conformations. **a**, The protein target is represented as a polygon mesh of the molecular surface with four properties encoded in each node, namely electrostatics, hydropathy, hydrogen bond donor / acceptor and shape index. **b**, The ligand is represented as a graph where each node corresponds to one atom and each edge to one bond. **c**, Ligand and target representations are processed by a neural network that extracts features using graph convolutions, which then are pairwise concatenated and used as input of a mixture density network. As a result, the model predicts a set of probability distributions that are assembled into a potential.

Potential based on distance likelihood can be used as an accurate scoring function

We used the CASF-2016 benchmark[30,31] in order to evaluate if our approach is suitable to be used as an accurate scoring function for an optimization algorithm. The CASF-2016 benchmark is composed of 285 protein-ligand complexes carefully selected to contain diverse proteins in terms of amino acid sequence and unique ligands with a wide binding affinity range. This particular benchmark is designed to assess scoring functions in four demanding tasks, namely scoring power, ranking power, docking power and screening power. Since DeepDock is not specifically trained to predict binding affinities, only the docking and screening power tasks are relevant in this study.

The evaluation of docking power measures the ability of a scoring function to identify native ligand binding poses among a set of decoys. For this, the CASF-2016 benchmark provides a set of ~100 decoy conformations for each of the 285 ligand-protein complexes, with an RMSD ranging from 0 to 10 Å from the native binding pose. The scoring function under evaluation is used to rank all decoys expecting those with similar conformations to the native ligand-binding pose (i.e. RMSD < 2 Å) to be among the top-ranked. Fig. 2a shows the results of our approach compared to results obtained for other 34 frequently used scoring functions evaluated in the same benchmark by Su et al.[31]. In 87% of the cases, the top ranked decoy using our approach was within an RMSD < 2 Å from the native ligand-binding pose and this amount increased to 94.7% if the 3 top-ranked decoys are considered. Based on these results, DeepDock is ranked among the top 5 best performing scoring function in this benchmark, and not far from the best performing scoring function, Autodock Vina, in which the conformation of the best ranked decoy was similar to the native binding pose in 90% of the cases. In addition, it is important to mention that our approach presented a Spearman's rank correlation of 0.83 between the computed score and the decoy RMSD from the native binding pose (Supplementary Fig. 1). In other words, this value indicates that the more similar the decoy conformation is to the native binding pose, the higher the score computed by the scoring function. This correlation has been used as an indicator of the efficiency of a scoring function since it is believed that scoring functions with a high rank correlation can improve conformation sampling to find the native pose[31].

The evaluation of screening power in CASF-2016 is designed to measure the ability of a scoring function to identify true binders of a specific target from a pool of random compounds. The CASF-2016 benchmark is composed of 57 protein targets each with a set of 5 true ligands (i.e., a total of 285 compounds) covering a wide range of binding affinity (at least 100-fold). The benchmark provides 100 precomputed binding conformations for each of the 285 ligands in each of the protein target (i.e., 28,500 conformations per target and 1,624,500 in total). First, the scoring function is used to assess all conformations in each target, then all compounds are ranked based on the score of their best conformation, and it is expected that true binders are among the top ranked compounds. The ability of distinguishing true binders from random molecules is evaluated using an enhancement factor (EF)[30,31]. DeepDock presented a mean EF of 16.41 (90% CI, 12.67 - 19.91) when the top 1% ranked compounds are considered, which is the highest compared to other scoring functions previously evaluated in this benchmark (mean EF < 12) as shown in Supplementary Fig. 2. In addition, Fig. 2b shows the success rate of identifying the most potent true binder among the top 1%, 5% or 10% ranked compounds using different scoring functions previously evaluated[31]. Our approach showed the best performance by finding the most potent ligand among the top 1% ranked compounds for 25 protein targets (43.9%), among the top 5% for 35 targets (61.4%) and among the top 10% for 47 targets (82.5%). Other scoring functions among the best performers are $\Delta_{\text{Vina}}RF_{20}$, GlideScore-SP, ChemPLP@GOLD and Autodock Vina which were able to rank the most potent ligand among the top 10% ranked compounds for just 37 targets or less (< 65%).

The above framework can also be used to evaluate the reverse screening power of a scoring function, that is the ability of identify the real target of a molecule among a set of random targets. Fig. 2c shows the reverse screening power of our approach compared to the performance of other scoring functions previously evaluated[31]. Our approach identified the true target among the 1% top ranked targets for 68 ligands (23.9%), among the 5% for 112 ligands (39.3%) and among the 10% for 145 ligands (50.9%).

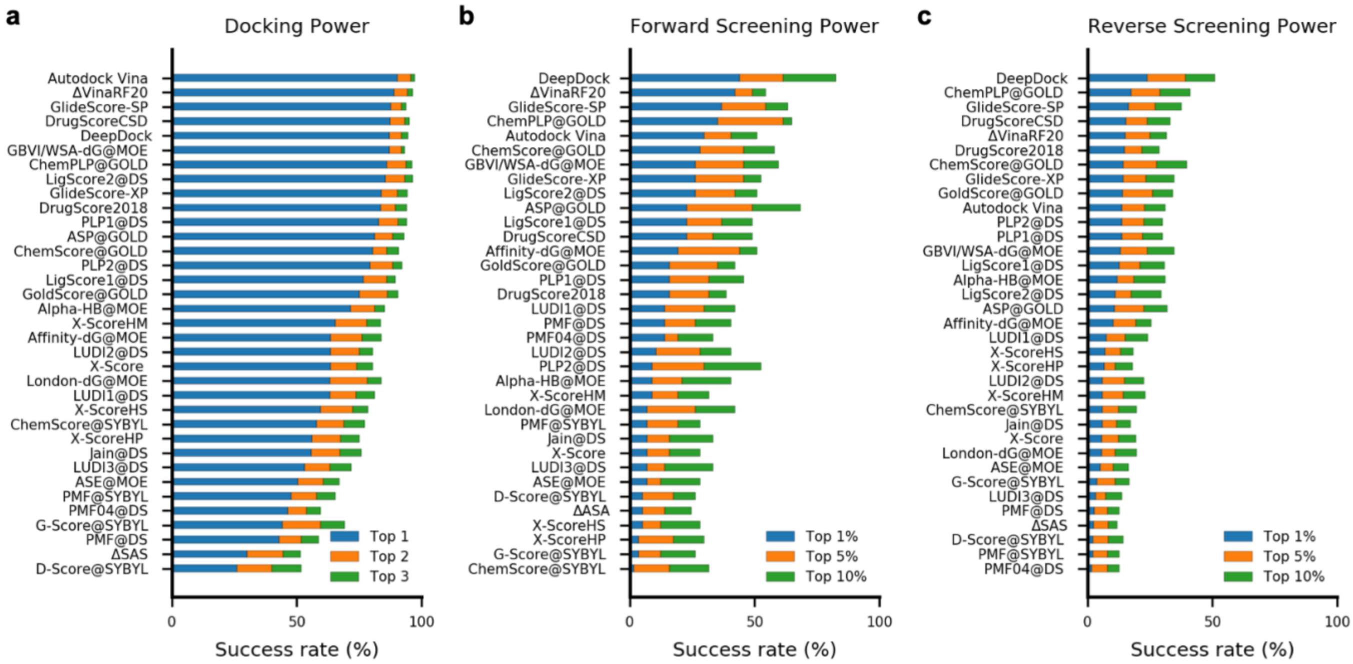

Fig. 2 Results of the distance likelihood potential in the CASF-2016 benchmark compared to other scoring functions reported by Su et al.[31]. **a**, Success rate of detecting real binding pose of a ligand (with an RMSD < 2 Å) among the top 1, 2, and 3 ranked poses during the docking power evaluation task. **b**, Success rate of detecting the highest affinity ligand for a given target (among the top 1%, 5%, and 10% candidates) during the forward screening task. **c**, Success rate of detecting the best target protein for a given ligand (among the top 1%, 5%, and 10% possible targets) during the reverse screening task.

Potential based on distance likelihood can reproduce experimental binding conformations

An advantage of this deep learning approach is that it can be easily combined with optimization algorithms in order to find the ligand conformation associated with the global minimum of the potential. In other words, it can find the ligand conformation with highest likelihood of binding. For this, the optimization algorithm carefully rotates each rotatable bond in the molecule, and at the same time it translates and rotates the whole ligand using a transformation matrix until it finds the conformation that best fits the binding pocket (Fig. 3a). In this case we used differential evolution[32] as the optimization algorithm, but others such as particle swarm optimization (PSO), simulated annealing (SA), or even gradient-based algorithms can be adapted to DeepDock. For example, gradient descent has recently been used to minimize a potential learnt by neural networks in order to predict protein structures[27].

To assess if an optimization algorithm can minimize a potential based on distance likelihood, we tried to reproduce real binding poses of different ligand-target when starting from a random conformation and position. For a more realistic test, this was done using the 285 ligand-target pairs in the CASF-2016 coreset plus 1,367 ligand-target pairs used as the validation set. It is worth mentioning that none of these complexes were included in the training set. DeepDock was able to find conformations corresponding to a minimum for 225 (87%) of the compounds in the CASF-2016 coreset and for 917 (67%) of the molecules in the validation set. Interestingly, the optimization failed for most of the compounds with more than 10 rotatable bonds (Fig. *3*g-h). The effect of the number of rotatable bonds on optimization has been noticed before and is linked to the inefficiency of the optimization algorithms when dealing with a large number of degrees of freedom[33]. In general, all compounds for which the optimization finished correctly presented a conformation very similar to the real binding pose, that is, a median (IQR) RMSD of 1.33 (0.81 to 1.99) Å for the CASF-2016 molecules and a median (IQR) RMSD of 1.47 (1.00 to 2.11) Å for molecules in the validation set (Fig. *3*i-j). These similarities are also evident from the high correlation between the scores produced by the predicted and the real binding pose (R^2 = 0.81 for CASF-2016 coreset and R^2 = 0.82 for the validation set). It is important to mention that no correlation was found between the compound binding affinity and the score of the predicted or real binding pose (Supplementary Fig. 3-4).

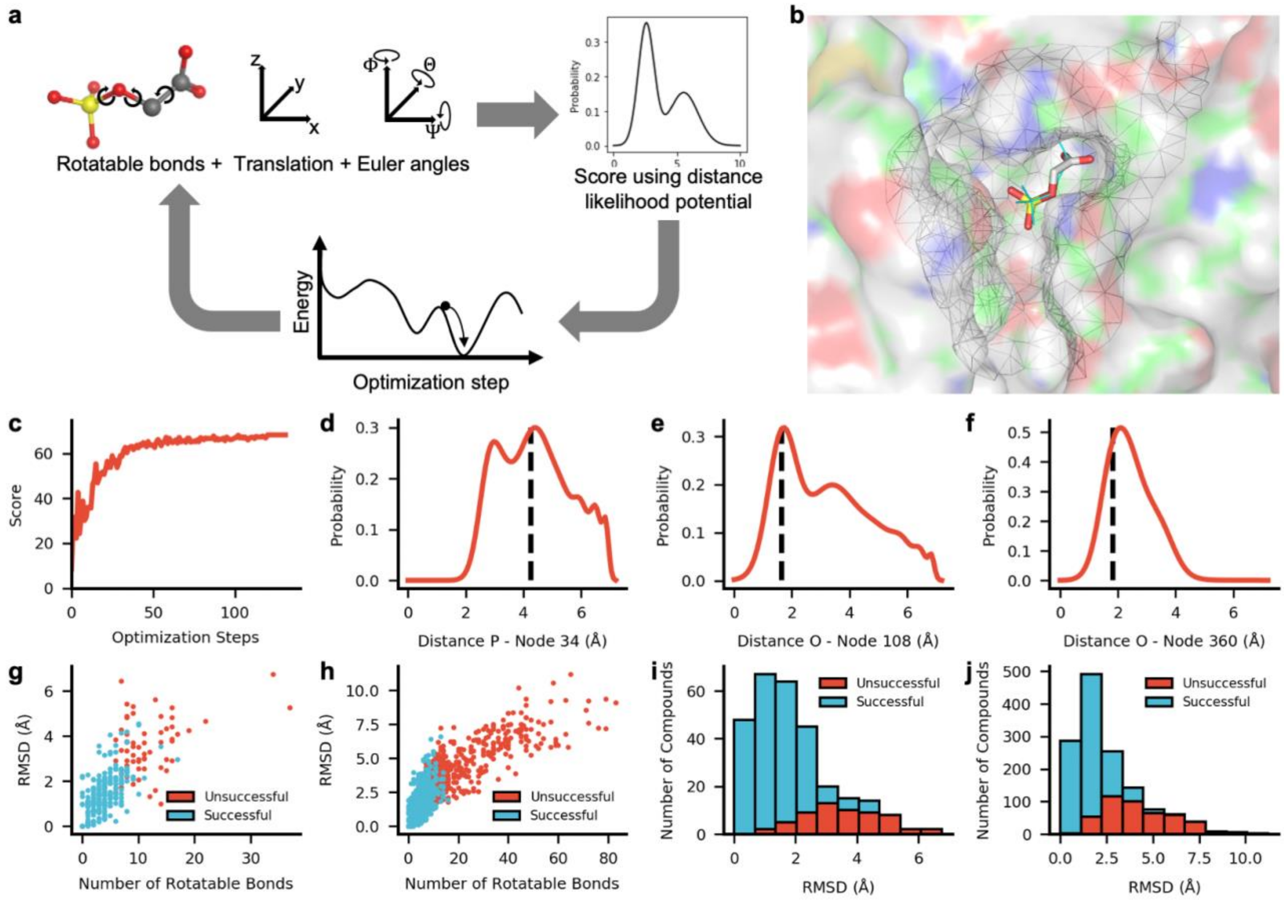

Fig. 3 Use of distance likelihood potential to predict ligand binding conformations. **a**, Representation of the optimization process where a ligand conformation is represented by a vector of the values of all rotatable bonds in the molecule, the displacement across the three dimensions of the Euclidean space and the three Euler angles that represent the rotation of the molecule. This conformation is scored using the distance likelihood potential and then optimized using differential evolution to produce a new conformation, which follows the same procedure until the optimization has successfully finished. **b**, Example of the predicted binding conformation of 2-phosphoglycolic acid to rat PEPCK (PDB ID: 2RKA). Experimental binding conformation is in depicted in cyan lines and the polygon mesh in gray lines. **c**, Optimization process of 2-phosphoglycolic acid to rat PEPCK. **d-f**, Examples of predicted distance probability distributions between ligand atoms and target nodes for 2RKA. The dashed line indicates the distance of between ligand atoms and target node for the predicted binding conformation. **g-h**, Scatter plots for 285 compounds in CASF-2016 (**g**) and 1,367 compounds in the validation set (**h**) showing that RMSD between predicted and experimental binding conformations is lower for compounds with less rotatable bonds. The optimization using differential evolution successfully finished for most compounds bearing less than 10 rotatable bonds. **g-h**, Distributions of RMSD between predicted and experimental binding conformations in CASF-2016 (**g**) and in the validation set (**h**). Color code refers to compounds for which the optimization successfully finished or not.

Conclusions

In this work, we report a method that exploits geometric deep learning to predict ligand binding conformations. Contrary to docking methods where a one-fits-all scoring function is used, here a deep neural network learns a potential that is specific for each protein-ligand complex, which is then used to find the optimal binding conformation. In a first instance, the deep neural network learns the parameters of a mixture model that is employed as a probability density function. This probability density function is used to determine the most likely distance separating a ligand atom from a specific point in the molecular surface of the binding site. The potential is determined as the combination of the negative log likelihood of all pairwise combination of ligand atoms and points in the molecular surface. The optimal conformation is the one that minimizes the potential, that is, the ligand conformation in which every atom is separated from the target surface by the most likely distance. We demonstrate that this potential can be used as an accurate scoring function for molecular docking and virtual screening. In fact, this potential performs equally or better than many of the most widely used scoring functions in the CASF-2016 benchmark. It is important to mention that this benchmark only contains a small number of compounds compared to real screening libraries (usually composed by millions of molecules). Finally, we also show that this potential can be minimized using global optimization methods, such as differential evolution[32], in order to find the most likely binding conformation of a ligand. More in concrete, we reproduce the binding conformation of 868 ligands (177 from CASF-2016 and 691 form the test set) with an RMSD < 2 Å from the experimental structure using the method described in this paper. Overall, we have presented evidence that geometric deep learning can be used to predict the binding conformation of ligands to their biological target. Although the results presented in this work were mainly focused on small molecules, similar approaches can be used to predict binding conformations of larger molecules such as peptides or even protein-protein interactions. We anticipate that further developments in geometric deep learning will help to significantly improve and speed up structure-based virtual screening.

Methods

Data set: The model reported in this study was trained using the general set of the PDBbind database (v.2019)[28], which contains a collection of 17,679 protein-ligand structures with their respective potency (e.g., IC_{50}, K_d, etc). From these, we removed those complexes that are included in the CASF-2016 benchmark and those that failed during the pre-processing step leaving a total 16,367 protein-ligand complexes which were randomly divided in a training set containing 15,000 complexes and a test set with 1,367. Each of these complexes was processed in order to be used as an input for the model.

The chemical structures of ligands were represented as undirected graphs $\mathcal{G}^l = (\mathcal{V}^l, \mathcal{E}^l)$ where nodes $v_i^l \in \mathcal{V}^l$ represent atoms in the molecule and edges $e_{i,j}^l \in \mathcal{E}^l$ represent bonds. In this case each node v_i^l is represented by a one-hot vector that indicates the atom type among 28 possibilities (Be, B, C, N, O, F, Mg, Si, P, S, Cl, V, Fe, Co, Cu, Zn, As, Se, Br, Ru, Rh, Sb, I, Re, Os, Ir, Pt, Hg). Similarly, each edge $e_{i,j}^l$ is represented by a one-hot vector that indicates the bond type, either single, double, triple or aromatic. It is important to mention that no information regarding the three-dimensional conformation of the ligand was used for training the model.

The protein targets were processed using a pipeline based on the one previously described by Gainza et al.[34]. As in MaSIF, protein surfaces were triangulated using MSMS[35] with a density of 3.0 $^{nodes}/_{\text{Å}^2}$ and a probe radius of 1.5 Å. The resulting meshes were down sampled to a resolution of 1 Å and processed using pymesh. The resulting mesh $\mathcal{G}^t = (\mathcal{V}^t, \mathcal{E}^t)$ is composed of a fixed set of nodes $v_i^t \in \mathcal{V}^t$ and edges $e_{i,j}^t \in \mathcal{E}^t$. Each node v_i^t is represented by vector of four features calculated using MaSIF, namely, Poisson-Boltzmann continuum electrostatics, free electrons and proton donors, hydropathy and shape index. In a similar way, each edge $e_{i,j}^t$ is represented by a vector defining the relative cartesian coordinates of the linked nodes i.e., $r_{i,j} = (p_i - p_j)$ where $p_i \in \mathbb{R}^3$ represents the coordinates of v_i^t in a three-dimensional Euclidean space. It is worth mentioning that only nodes defining the binding site (i.e., within 10 Å or less from any ligand atom) were used to train the model.

Model: The model construction can be divided into three stages: feature extraction, feature concatenation and a mixed density network. In the first stage, features are extracted by two independent residual graph convolutional neural networks (GNNs), one for the ligand and the other for the target. Despite being independent, both residual GNN have the same architecture. First, the node and edge features are projected to a 128-dimensional embedding using a linear layer as in Eqs. (1) and (2). Then we used a sequence of three GNNs to update each node and edge based on their neighbouring nodes and the type of edges connecting them. The GNN first updates each edge in the graph by applying a multi-layer perceptron (MLP) on the concatenation of the edge features and the features of the two connecting nodes as shown in Eq. (3). The updated edge features $e_{i,j}^{\ell}$ are used to update the node features as shown in Eq. (4). The updated edge and node features ($e_{i,j}^{\ell}$ and v_i^{ℓ}, respectively) contain information of the central atom but also of the neighbouring atoms around it and can be used as input of another convolution round (Eqs. (3) and (4)). In this case, we used three convolutions, i.e. up to $\ell = 3$.

$$e_{i,j}^{0} = Linear\left(e_{i,j}\right) \tag{1}$$

$$v_i^{0} = Linear(v_i) \tag{2}$$

$$e_{i,j}^{\ell} = MLP\left(\left[v_i^{\ell-1}, v_j^{\ell-1}, e_{i,j}^{\ell-1}\right]\right) \tag{3}$$

$$v_i^{\ell} = MLP\left(\left[v_i^{\ell-1}, \frac{1}{\|j\|}\sum_{j} MLP\left(\left[v_j^{\ell-1}, e_{i,j}^{\ell}\right]\right)\right]\right) \tag{4}$$

After the initial processing by the GNNs, the node and edge features were processed by 10 residual GNN blocks. Each residual block starts by projecting the node and edge features (v_i^{h-1} and $e_{i,j}^{h-1}$, respectively) to a 32-dimensional vector using an MLP as shown in Eqs. (5) and (6). The resulting vectors are used as inputs to a GNN (Eq. (7)) resulting in aggregated node and edge features v_i' and $e_{i,j}'$, respectively, which are projected back to 128-dimmensional vectors v_i^{up} and $e_{i,j}^{up}$ as shown in Eqs. (8) and (9). Finally, the resulting vectors (v_i^{up} and $e_{i,j}^{up}$) are added to the input vectors (v_i^{h-1} and $e_{i,j}^{h-1}$,) to create a skip connection and later modified by an activation function (Eqs. (10) and (11)).

$$e_{i,j}^{down} = MLP\left(e_{i,j}^{h-1}\right) \tag{5}$$

$$v_i^{down} = MLP\left(v_i^{h-1}\right) \tag{6}$$

$$v_i', e_{i,j}' = GNN\left(v_i^{down}, v_j^{down}, e_{i,j}^{down}\right) \tag{7}$$

$$e_{i,j}^{up} = Dropout\left(BatchNorm\left(Linear\left(e_{i,j}'\right)\right)\right) \tag{8}$$

$$v_i^{up} = Dropout\left(BatchNorm\left(Linear\left(v_i'\right)\right)\right) \tag{9}$$

$$e_{i,j}^{h} = ELU\left(e_{i,j}^{h-1} + e_{i,j}^{up}\right) \tag{10}$$

$$v_i^{h} = ELU\left(v_i^{h-1} + v_i^{up}\right) \tag{11}$$

The extracted node features by the GNNs and residual GNNs for both target $\bar{v}_r^t$ and ligand $\bar{v}_s^l$ are then pairwise concatenated and used as input of a mixture density network (MND)[29]. The MND uses an MLP to create a hidden representation $h_{r,s}$ that combines the concatenated target and ligand node information as shown in Eq. (12). The hidden representation is used to compute the outputs of the MND, which consist of the means ($\mu_{r,s}$), standard deviations ($\sigma_{r,s}$) and mixing coefficients ($\alpha_{r,s}$) that are necessary to parametrize a mixture of gaussians (Eqs. (13)- (15)). In this particular case, the mixture model uses 10 gaussians to simulate the probability density distribution of the distance between the ligand and target nodes ($\bar{v}_s^l$ and $\bar{v}_r^t$, respectively).

$$h_{r,s} = Dropout\left(MLP\left(\left[\bar{v}_r^t, \bar{v}_s^l\right]\right)\right) \tag{12}$$

$$\mu_{r,s} = ELU\left(Linear\left(h_{r,s}\right)\right) + 1 \tag{13}$$

$$\sigma_{r,s} = ELU\left(Linear\left(h_{r,s}\right)\right) + 1 \tag{14}$$

$$\alpha_{r,s} = Softmax\left(Linear\left(h_{r,s}\right)\right) \tag{15}$$

In addition, the extracted ligand node features $\bar{v}_s^l$ were used to predict auxiliary tasks, namely atom type and bond type with connecting neighbouring nodes. These auxiliary tasks help to learn molecular structures which accelerates training. All MLPs used are composed by a linear layer followed by batch normalization and an ELU activation function. The dropout rate used was of 0.1 in all experiments.

Training: We employed the Adam optimizer with a learning rate of 0.002 to update model weights. The model was trained to minimize the loss function shown in Eq. (16) where $\mathcal{L}_{MDN}$ represents the loss of the mixture density network whereas $\mathcal{L}_{atoms}$ and $\mathcal{L}_{bonds}$ are the cross-entropy cost functions of predicting atom and bond types, respectively, that were used as auxiliary tasks. In particular, $\mathcal{L}_{MDN}$ minimizes the negative log-likelihood of $d_{r,s}$, which represents the distance separating the target node v_r^t from the ligand node v_s^l, computed using the mixture model formed by $k = 10$ gaussians and parametrised by $\alpha_{r,s}$, $\mu_{r,s}$ and $\sigma_{r,s}$ that were predicted by the model (Eq. (17)). The model was trained for 150 epochs using a batch size of 16 protein-ligand complexes. Contributions of ligand-target node pairs separated by a $d_{r,s}$ > 7 Å were masked since we considered that those atoms cannot form relevant interactions.

$$\mathcal{L}_{total} = \mathcal{L}_{MDN} + \mathcal{L}_{atoms} + \mathcal{L}_{bonds} \qquad (16)$$

$$\mathcal{L}_{MDN} = -\log P\left(d_{r,s} \middle| v_r^t, v_s^l\right) = -\log \sum_{k=1}^{K} \alpha_{r,s,k} \mathcal{N}\left(d_{r,s} \middle| \mu_{r,s,k}, \sigma_{r,s,k}\right) \qquad (17)$$

$$U_{(x)} = -\sum_{r=1}^{R} \sum_{s=1}^{S} \log P\left(d_{r,s} \middle| v_r^t, v_s^l\right) \qquad (18)$$

The loss function shown in Eq. (17) can be easily used to define a potential $U_{(x)}$ which is tailored for a particular target-ligand complex (Eq. (18)). It is possible to use this potential to score the 3D structure of a target-ligand complex by computing the distances $d_{r,s}$ separating each target node v_r^t from each ligand node v_s^l in that specific conformation, calculating the negative log likelihood $-\log P\left(d_{r,s} \middle| v_r^t, v_s^l\right)$ for each target-ligand node pair, and summing across all possible pairs. The lower the value of $U_{(x)}$, the more likely is to find the target-ligand complex in that specific conformation.

Benchmark: This approach was evaluated using the CASF-2016 benchmark[30,31], which contains 285 protein-ligand complexes carefully curated. Structures from this benchmark were preprocessed in the same way as the training set. There are four different tasks in this benchmark in order to evaluate the scoring power, ranking power, docking power and screening power of a scoring function. Only the docking and screening power are relevant for this evaluation. The former evaluates the ability of the scoring function to identify the real binding conformation among generated decoy conformations of the same ligand. The latter evaluates if the scoring function can identify true binders for a particular target using the enhancement factor (Eq. (19)) as metric. Results can be directly compared with other scoring functions previously evaluated by Su et al.[31]. The complete protocol and scripts are fully described in the original publication[30].

$$EF_{1\%} = \frac{Number\ of\ true\ binders\ detected\ among\ 1\%}{(Total\ true\ binders) \times 1\%} \qquad (19)$$

Prediction of binding conformations: We represent the ligand conformation as vector of the Euler angels, the relative position of the ligand in the Euclidean space and the dihedral angles of all rotatable bonds in the molecule. We employed differential evolution[32] to find the ligand conformation that minimize the potential $U_{(x)}$ learnt by the model for that specific complex, that is, the resulting ligand conformation will be the most likely to interact with target binding site according to the model. We run the global optimization for a maximum of 500 iterations using a population size of 150, the mutation constant was randomly changed each generation from a (0.5, 1) interval and with a recombination constant of 0.8. The values of dihedrals of rotatable bonds and Euler angles were restricted to be between $-\pi$ and π. In interest of reproducibility, the calculation was seeded but this is not a requirement.

Author contributions

O.M.L. conceived the idea, wrote the code, performed the experiments and wrote the paper. M.A., A.E.D.C. and J.K.W. helped with the preparation of the manuscript and with insightful discussions. A.E.D.C. helped to improve code.

Acknowledgments

Authors thank Dries Van Rompaey, Jonas Verhoeven and Natalia Dyubankova for supporting this project.

Competing interests

O.M.L., M.A. and J.K.W. are employees of Janssen Pharmaceutica N.V.

Data availability

The data that support the findings of this study will be available after the final publication of this manuscript.

Code availability

The code used to generate results shown in this study will be available after the final publication of this manuscript.

References

1. Hert, J., Irwin, J. J., Laggner, C., Keiser, M. J. & Shoichet, B. K. Quantifying biogenic bias in screening libraries. *Nat. Chem. Biol.* **5**, 479–483 (2009).

2. Dobson, C. M. Chemical space and biology. *Nature* **432**, 824–828 (2004).

3. Congreve, M., Murray, C. W. & Blundell, T. L. Keynote review: Structural biology and drug discovery. *Drug Discov. Today* **10**, 895–907 (2005).

4. Klebe, G. Protein--Ligand Interactions as the Basis for Drug Action. in *Drug Design: Methodology, Concepts, and Mode-of-Action* (ed. Klebe, G.) 61–88 (Springer Berlin Heidelberg, 2013). doi:10.1007/978-3-642-17907-5_4.

5. Renaud, J. P. *et al.* Cryo-EM in drug discovery: Achievements, limitations and prospects. *Nat. Rev. Drug Discov.* **17**, 471–492 (2018).

6. Gorgulla, C. *et al.* An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **580**, 663–668 (2020).

7. Lyu, J. *et al.* Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).

8. Krivák, R. & Hoksza, D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J. Cheminform.* **10**, 1–12 (2018).

9. Pu, L., Govindaraj, R. G., Lemoine, J. M., Wu, H. C. & Brylinski, M. Deepdrug3D: Classification of ligand-binding pockets in proteins with a convolutional neural network. *PLoS Comput. Biol.* **15**, 1–23 (2019).

10. Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S. & De Fabritiis, G. DeepSite: Protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **33**, 3036–3042 (2017).

11. Ain, Q. U., Aleksandrova, A., Roessler, F. D. & Ballester, P. J. Machine-learning scoring functions to improve structure-based binding affinity

prediction and virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **5**, 405–424 (2015).

12. Li, H., Sze, K. H., Lu, G. & Ballester, P. J. Machine-learning scoring functions for structure-based virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **11**, (2021).

13. Sánchez-Cruz, N., Medina-Franco, J. L., Mestres, J. & Barril, X. Extended connectivity interaction features: improving binding affinity prediction through chemical description. *Bioinformatics* **btaa982**, (2020).

14. Wójcikowski, M., Ballester, P. J. & Siedlecki, P. Performance of machine-learning scoring functions in structure-based virtual screening. *Sci. Rep.* **7**, 1–10 (2017).

15. Wójcikowski, M., Kukiełka, M., Stepniewska-Dziubinska, M. M. & Siedlecki, P. Development of a protein-ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics* **35**, 1334–1341 (2019).

16. Ballester, P. J. & Mitchell, J. B. O. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics* **26**, 1169–1175 (2010).

17. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. & Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **57**, 942–957 (2017).

18. Stepniewska-Dziubinska, M. M., Zielenkiewicz, P. & Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* **34**, 3666–3674 (2018).

19. Hassan-Harrirou, H., Zhang, C. & Lemmin, T. RosENet: Improving Binding Affinity Prediction by Leveraging Molecular Mechanics Energies with an Ensemble of 3D Convolutional Neural Networks. *J. Chem. Inf. Model.* **60**, 2791–2802 (2020).

20. Jiménez, J., Škalič, M., Martínez-Rosell, G. & De Fabritiis, G. KDEEP:

Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **58**, 287–296 (2018).

21. Feinberg, E. N. *et al.* PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **4**, 1520–1530 (2018).

22. Lim, J. *et al.* Predicting Drug-Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation. *J. Chem. Inf. Model.* **59**, 3981–3988 (2019).

23. Gasteiger, J., Rudolph, C. & Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **3**, 537–547 (1990).

24. Velec, H. F. G., Gohlke, H. & Klebe, G. DrugScoreCSDKnowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of Near-Native Ligand Poses and Better Affinity Prediction. *J. Med. Chem.* **48**, 6296–6303 (2005).

25. Fan, H. *et al.* Statistical Potential for Modeling and Ranking of Protein–Ligand Interactions. *J. Chem. Inf. Model.* **51**, 3078–3092 (2011).

26. Klebe, G. & Mietzner, T. A fast and efficient method to generate biologically relevant conformations. *J. Comput. Aided. Mol. Des.* **8**, 583–606 (1994).

27. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).

28. Liu, Z. *et al.* PDB-wide collection of binding data: Current status of the PDBbind database. *Bioinformatics* **31**, 405–412 (2015).

29. Bishop, C. M. Mixture Density Networks. *Tech. Report. Ast. Univ. Birmingham.* (1994).

30. Li, Y. *et al.* Assessing protein-ligand interaction scoring functions with the CASF-2013 benchmark. *Nat. Protoc.* **13**, 666–680 (2018).

31. Su, M. *et al.* Comparative Assessment of Scoring Functions: The CASF-

2016 Update. *J. Chem. Inf. Model.* **59**, 895–913 (2019).

32. Storn, R. & Price, K. Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *J. Glob. Optim.* **11**, 341–359 (1997).

33. Li, H., Leung, K. S., Ballester, P. J. & Wong, M. H. Istar: A web platform for large-scale protein-ligand docking. *PLoS One* **9**, (2014).

34. Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).

35. Sanner, M. F., Olson, A. J. & Spehner, J. C. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers* **38**, 305–320 (1996).