



Deep graph learning in molecular docking: Advances and opportunities

Norberto Sánchez-Cruz^{a,b}

^a Instituto de Química, Unidad Mérida, Universidad Nacional Autónoma de México, Carretera Mérida-Tetiz Km. 4.5, Ucu, Yucatán 97357, Mexico

^b Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas Unidad Mérida, Universidad Nacional Autónoma de México, Sierra Papacal, Mérida, Yucatán 97302, Mexico

ARTICLE INFO

Keywords:

Structure-based drug discovery
Molecular docking
Machine learning
Deep graph learning

ABSTRACT

One of the main computational tools for structure-based drug discovery is molecular docking. Due to the natural representation of molecules as graphs (a set of nodes/atoms connected through edges/bonds), Deep Graph Learning has been successfully applied for multiple tasks on this area. This work presents an overview of Deep Graph Learning methods developed within this research field, as well as opportunities for future development.

Introduction

Structure-Based Drug Design (SBDD) is an essential component in medicinal chemistry and drug discovery [1,2]. One of the most relevant computational tools for SBDD is molecular docking, whose main goals are the identification of compounds able to bind to a macromolecular target (screening), the prediction of the binding mode of such compounds (docking), and the accurate estimation of its binding affinity (scoring) [3]. Over the past decade, the application of multiple Machine Learning (ML) methods has been explored in this field [4–6]. The most general approach involves the development of scoring functions to predict the binding affinity of a given protein-ligand complex, such predictions are then employed to sort different compounds and multiple binding poses for the selection of true binders and their binding mode. Due to the natural representation of molecules as graphs (a set of nodes/atoms connected through edges/bonds), Deep Graph Learning (DGL) [7,8], a deep neural network based approach able to learn from graph structured data, has been increasingly applied in this research field. Herein, an overview of DGL methods applied to molecular docking is provided (Table 1), highlighting the graph representations employed, their processing, the tasks learned within the molecular docking field, as well as opportunities for future development.

Brief introduction to deep graph learning in the context of molecular docking

In this section, a brief introduction to deep graph learning in molecular docking is provided. Detailed descriptions of the field and its applications in multiple areas can be found in [7,8]. In simple terms, a graph is a data structure consisting of a set of nodes connected by edges.

The nodes representing objects, and the edges representing some relationship between the nodes. These two elements can be annotated with one or multiple features/descriptors. DGL can be understood as the use of neural networks (typically called Graph Neural Networks) to process graph-structured data and obtain embedded representations of the nodes, edges, or graphs into an n-dimensional space that can be used to train ML algorithms (typically additional fully connected layers) for specific tasks at a node, edge, or graph level (e.g., predicting the classification of a node, predicting the presence of a link between two nodes or predicting a graph property, etc.). In molecular docking, most DGL applications fall in the last case, representing protein-ligand complexes as graphs and trying to learn graph level properties, such as its classification (binder/non binder) or its binding affinity. This approach is illustrated in Fig. 1.

Representation of protein-ligand complexes

DGL methods rely on the representation of proteins, ligands and/or protein-ligand complexes as graphs with node and edge features. Chemical structures of small molecules are usually represented as molecular graphs, with atoms as nodes and bonds as edges. However, although this representation can also be employed for macromolecules such as proteins, its application has been limited to methods centered on modeling the protein binding pocket only [11,12,14–16,18,19]. For methods that intend to model the docking of small molecules when no binding pocket is known (blind docking), a coarse representation of the protein is often preferred, modeling protein residues as nodes and establishing the edges on a distance basis [9,10]. On the other hand, methods such as HOLOPROT [17] and DeepDock [13] chose to obtain the graph representation of the protein from the vertices and edges of a polygonal

Abbreviations: DGL, deep graph learning; DGM, diffusion generative model; GAT, graph attention networks; GCN, graph convolutional networks; IEGMN, E(3)-equivariant graph matching network; ML, machine learning; SBDD, structure-based drug design.

E-mail address: norberto.sanchez@iquimica.unam.mx

<https://doi.org/10.1016/j.ailsci.2023.100062>

Received 31 October 2022; Received in revised form 1 February 2023; Accepted 1 February 2023

Available online 3 February 2023

2667-3185/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Table 1
Deep graph learning methods applied to molecular docking.

Method	Protein Graph	Ligand Graph	Complex Graph	Screening	Docking	Scoring	Refs.
DiffDock			X		X		[9]
EQUIBIND	X	X			X		[10]
MedusaGraph			X		X		[11]
PIGNet			X	X	X	X	[12]
DeepDock	X	X		X	X	X	[13]
InteractionGraphNet	X	X	X	X	X	X	[14]
SG-CNN			X			X	[15]
SIGN			X			X	[16]
HOLOPROT	X	X				X	[17]
GraphBAR			X			X	[18]
Lim et al.			X	X			[19]

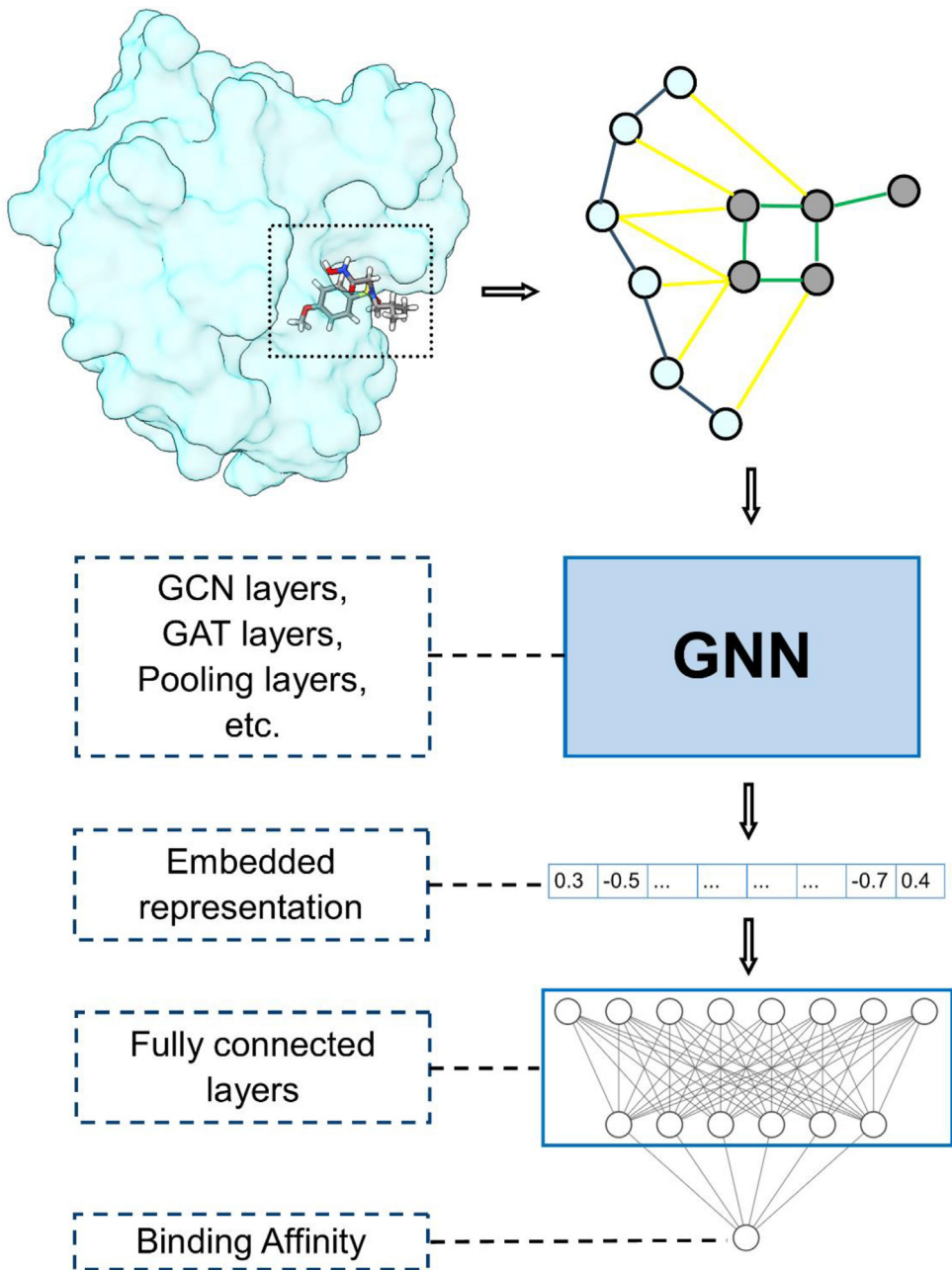


Fig. 1. Most common deep graph learning approach employed in molecular docking.

mesh representation of the protein surface. Regardless of the graph representation employed, nodes and edges encode chemical properties at a surface, atom, or residue level (shape index, atom type, residue type, etc. for nodes; and distance, bond type, etc. for edges).

While the representation of proteins and ligands as independent graphs is relatively simple, the representation of the whole protein-ligand complex is less trivial. The most common representation involves the construction of a graph containing all the protein and ligand atoms as nodes, whose edges are defined through constraints on the maximum distance allowed and/or number of neighbors [9,11,16], or through multiple adjacency matrices to differentiate between distance thresholds [18] or connection types (covalent/non covalent) [12,15,19], which are subsequently processed directly by Graph Convolutional Networks (GCN) or Graph Attention Networks (GAT). Other representations involve the use of independent graphs for the protein and ligand, which are processed through GCN to obtain embeddings that are concatenated and used for the subsequent tasks [13,17], or processed through additional blocks allowing the message passing between the embedded representations of the protein and ligand nodes [10,14].

Learned tasks within molecular docking

The most general ML approach to deal with the tasks of screening, docking and scoring at once is to train a regression model on distinct representations of experimentally determined protein-ligand complexes with known affinity to obtain a scoring function to be employed for these tasks. However, regarding the DGL models discussed herein, only PIGNet [12] and DeepDock [13] employed a similar strategy by building a single model for the three tasks. The rest of the methods were focused on the development of task specific models, with InteractionGraphNet [14] being the only approach that developed a different model for each task.

DGL models developed for screening addressed this task as a binary classification problem, where the models estimate the probability of binding for a given protein-ligand complex; these models include InteractionGraphNet [14], and the model developed by Lim et al. [19], which were trained on datasets containing binder and non-binder compounds coming from the DUD-E dataset [20]. On the other hand, most DGL models focused on scoring address this task as a regression problem to directly predict the binding affinity of a given protein-ligand complex [14–18], for which they are trained on 3D structures of protein-ligand complexes with known binding affinity, showing state-of-the-art performances when tested on well-established benchmarks from PDBBind [21,22].

DGL models developed for docking are a particular case. While most ML models employed for docking are limited to the development of a scoring function capable of distinguishing the experimental binding pose within a set of decoys (often generated by several runs from standard docking software). Recently developed DGL models attempt to predict the movement of the ligand into the binding pocket of the protein. MedusaGraph [11] addressed this problem by training a regression model based on Transformer Convolutional layers [23], to learn the moving vector of each node needed to move the ligand atoms from a starting docking pose to the experimental binding mode, claiming improved accuracy and speed in comparison to standard docking software. Two other examples of models tackling the docking task under a less typical approach are EQUIBIND [10] and DiffDock [9]. Although these approaches are available only as preliminary publications, and thus lack of rigorous peer review, they suggest interesting approaches to tackle the docking problem. On the one hand, EQUIBIND implements a regression model based on Independent E(3)-Equivariant Graph Matching Network (IEGMN) [24] to predict in a single shot a set of keypoints on the protein and ligand, whose alignment is in principle the transformation needed to go from a random ligand conformation to the experimental binding mode, although this approach claims a speed-up of at least 10 times in comparison to standard docking software, its performance seems to be

far from competitive when its output conformations are not further refined by well-established docking software. On the other hand, DiffDock addressed the docking task as a generative modeling problem, implementing a Diffusion Generative Model (DGM) over the degrees of freedom involved in the docking process (translation, rotation and ligand torsional angles), learning a distribution over ligand poses conditioned on the protein structure. In addition, a confidence model is also trained for DiffDock to score the generated poses; preliminary results available for these two models combined suggest an improved performance in terms of speed and accuracy in comparison to standard docking software.

As stated above, only PIGNet [12] and DeepDock [13] have built a scoring function with applicability on the three tasks. However, their approaches do not rely on the training of a regression model to directly predict the binding affinity for a given protein-ligand complex. The PIGNet model was trained to predict parameters of physics-informed equations that describe intermolecular interactions (hydrogen bond, hydrophobic, etc.); with them, the energy components of each protein-ligand atom pair are calculated, and the total energy is obtained as the sum of these components divided by an entropy term related to the number of rotatable bonds in the ligand. The training set of PIGNet consisted of an augmented version of the PDBBind that included the binding poses of random ligands and conformations as an attempt to provide wrong examples to the model and improve its generalizability, which resulted in improved docking and screening accuracies. On the other hand, DeepDock consists of a model trained to learn statistics of the distribution of distances between atom types, which are subsequently used to build a statistical potential for the distance between protein-ligand atom pairs, which is employed to calculate the energy of the protein-ligand complexes. Although this model was trained on structures from the PDBBind with no data augmentation strategies, it showed state-of-the-art performance for the three tasks, which is a significant achievement.

Opportunities for future development

Although significant advances have been achieved in the development of DGL models for molecular docking, important challenges need to be addressed. Most of the developed models have been tested in datasets where the protein is treated as a rigid body. The graph representations employed to describe the complexes often include the distance between atoms as an edge feature, allowing an implicit description of the flexibility of both protein and ligand (since different conformations would yield to different graphs). However, the impact of this flexibility in the models has not been studied in detail.

Another difficult question to answer for these models is its generalization potential. This is due to the sparsity of protein-ligand complexes, as well as the known bias of the PDBBind and DUD-E datasets (the gold standard datasets to train and test these models) towards preferred crystallographic targets [25,26]. These problems will diminish as more experimental data on protein-ligand complexes become available in the public domain. In the meantime, validation splits based on time or structural similarity have been suggested to partially address this problem, although there are still no well-established protocols incorporating them, which makes direct comparison between different models difficult.

A couple of other problems arise from the fact that one of the claims of deep learning is its ability to automatically extract features relevant to the task of interest. On the one hand, this triggers assumptions that DGL models are capable of capturing the physicochemical foundations of the intermolecular interactions involved in molecular docking, which has been strongly questioned [27,28] and explored to a limited extent [14]. On the other hand, the increasing availability of open-source packages for deep learning and their ease of use have led to the development of increasingly complex models with different architectures. These aspects can lead to the misconception that the representation of the input data is irrelevant for the construction of the model, which can be reflected by

the fact that the number of edges on graph representations of protein-ligand complexes is often limited by a maximum number of neighbors or distance thresholds to speed up their processing, even when they are processed by models relying on hundreds of thousands of parameters to be adjusted, while the impact of employing complete graphs as representation has not been studied. The interpretability and limitations of these approaches should be investigated in detail for further development of the field.

DGL has been successfully applied to different tasks involved in molecular docking. Although their applications have been extensively validated in well-established benchmarks, their real potential will only be validated when their applications in drug discovery projects are demonstrated, which hopefully will occur in the next few years.

Funding

This work was funded by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) [IA210023](#).

Declaration of Competing Interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

- [1] Amy CA. The Process of structure-based drug design. *Chem Biol* 2003;10:787–97. doi:[10.1016/j.chembiol.2003.09.002](#).
- [2] Maria B, Bilal A, Sangdun C. A structure-based drug discovery paradigm. *Int J Mol Sci* 2019;20:2783. doi:[10.3390/ijms20112783](#).
- [3] Lyu J, Wang S, Balias TE, et al. Ultra-large library docking for discovering new chemotypes. *Nature* 2019;566:224–9. doi:[10.1038/s41586-019-0917-9](#).
- [4] Crampon K, Giorkallos A, Deldossi M, et al. Machine-learning methods for ligand–protein molecular docking. *Drug Discov Today* 2022;27:151–64. doi:[10.1016/j.drudis.2021.09.007](#).
- [5] Li H, Sze K, Lu G, Ballester PJ. Machine-learning scoring functions for structure-based drug lead optimization. *WIREs Comput Mol Sci* 2020;10:1–20. doi:[10.1002/wcms.1465](#).
- [6] Li Y, Su M, Liu Z, et al. Assessing protein–ligand interaction scoring functions with the CASF-2013 benchmark. *Nat Protoc* 2018;13:666–80. doi:[10.1038/nprot.2017.114](#).
- [7] Georgousis S, Kenning MP, Xie X. Graph deep learning: state of the art and challenges. *IEEE Access* 2021;9:22106–40. doi:[10.1109/ACCESS.2021.3055280](#).
- [8] Xia F, Sun K, Yu S, et al. Graph learning: a survey. *IEEE Trans Artif Intell* 2021;2:109–27. doi:[10.1109/TAI.2021.3076021](#).
- [9] Corso G., Stärk H., Jing B., et al. (2022) DiffDock: diffusion steps, twists, and turns for molecular docking. *arXiv. Preprint*. 10.48550/arXiv.2210.01776
- [10] Stärk H., Ganea O.-E., Pattanaik L., et al. (2022) EquiBind: geometric deep learning for drug binding structure prediction. *arXiv. Preprint*. 10.48550/arXiv.2202.05146
- [11] Jiang H, Wang J, Cong W, et al. Predicting protein–ligand docking structure with graph neural network. *J Chem Inf Model* 2022;62:2923–32. doi:[10.1021/acs.jcim.2c00127](#).
- [12] Moon S, Zhong W, Yang S, et al. PIGNet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chem Sci* 2022;13:3661–73. doi:[10.1039/D1SC06946B](#).
- [13] Méndez-Lucio O, Ahmad M, del Rio-Chanona EA, Wegner JK. A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nat Mach Intell* 2021;3:1033–9. doi:[10.1038/s42256-021-00409-9](#).
- [14] Jiang D, Hsieh CY, Wu Z, et al. InteractionGraphNet: a novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions. *J Med Chem* 2021;64:18209–32. doi:[10.1021/acs.jmedchem.1c01830](#).
- [15] Jones D, Kim H, Zhang X, et al. Improved protein–ligand binding affinity prediction with structure-based deep fusion inference. *J Chem Inf Model* 2021;61:1583–92. doi:[10.1021/acs.jcim.0c01306](#).
- [16] Li S., Zhou J., Xu T., et al. (2021) Structure-aware interactive graph neural networks for the prediction of protein–ligand binding affinity. *arXiv. Preprint*. 10.48550/arXiv.2107.10670
- [17] Somnath V.R., Bunne C., Krause A. (2021) Multi-scale representation learning on proteins. *arXiv. Preprint*. 10.48550/arXiv.2204.02337
- [18] Son J, Kim D. Development of a graph convolutional neural network model for efficient prediction of protein–ligand binding affinities. *PLoS One* 2021;16:e0249404. doi:[10.1371/journal.pone.0249404](#).
- [19] Lim J, Ryu S, Park K, et al. Predicting drug–target interaction using a novel graph neural network with 3d structure-embedded graph representation. *J Chem Inf Model* 2019;59:3981–8. doi:[10.1021/acs.jcim.9b00387](#).
- [20] Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* 2012;55:6582–94. doi:[10.1021/jm300687e](#).
- [21] Su M, Yang Q, Du Y, et al. Comparative assessment of scoring functions: the CASF-2016 update. *J Chem Inf Model* 2019;59:895–913. doi:[10.1021/acs.jcim.8b00545](#).
- [22] Liu Z, Su M, Han L, et al. Forging the basis for developing protein–ligand interaction scoring functions. *Acc Chem Res* 2017;50:302–9. doi:[10.1021/acs.accounts.6b00491](#).
- [23] Shi Y, Huang Z, Feng S, et al. Masked label prediction: unified message passing model for semi-supervised classification. *IJCAI Int Jt Conf Artif Intell* 2021:1548–54. doi:[10.24963/ijcai.2021/214](#).
- [24] Ganea O.-E., Huang X., Bunne C., et al. (2021) Independent SE(3)-equivariant models for end-to-end rigid protein docking. *arXiv. Preprint*. 10.48550/arXiv.2111.07786
- [25] Yang J, Shen C, Huang N. Predicting or pretending: artificial intelligence for protein–ligand interactions lack of sufficiently large and unbiased datasets. *Front Pharmacol* 2020;11. doi:[10.3389/fphar.2020.00069](#).
- [26] Wang J, Dokholyan NV. Yuel: improving the generalizability of structure-free compound–protein interaction prediction. *J Chem Inf Model* 2022;62:463–71. doi:[10.1021/acs.jcim.1c01531](#).
- [27] Bajorath J. Deep learning of protein–ligand interactions—remembering the actors. *Artif Intell Life Sci* 2022;2:100037. doi:[10.1016/j.ailsci.2022.100037](#).
- [28] Volkov M, Turk JA, Drizard N, et al. On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. *J Med Chem* 2022;65:7946–58. doi:[10.1021/acs.jmedchem.2c00487](#).