A SYSTEM AND METHOD FOR DETERMINING A BIOACTIVITY PARAMETER FOR A LIGAND

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This patent application claims priority to and the benefit of Singapore Patent Application No. 10202400124Q, filed on 16 January 2024 in the Intellectual Property Office of Singapore, the disclosure of which is incorporated by reference herein in its entirety.

TECHNICAL FIELD

[0002] Various aspects of this disclosure relate to a system and method for determining a bioactivity parameter for a ligand.

BACKGROUND

[0003] The following discussion of the background art is intended to facilitate an understanding of the present disclosure only. It should be appreciated that the discussion is not an acknowledgement or admission that any of the material referred to was published, known or is part of the common general knowledge of the person skilled in the art in any jurisdiction as of the priority date of the disclosure.

[0004] The drug discovery process includes stages such as target identification, hit molecule identification, lead molecule optimization, preclinical research, clinical trials, and more. A crucial phase is hit molecule identification, which screens potential therapeutic candidates with high bioactivities. Bioactivity prediction using deep learning has emerged as a tool for early-stage molecule screening, aiding in identifying active molecules efficiently. This prediction also guides lead molecule refinement, enhancing bioactivity and other properties. However, the accuracy of current deep learning-based bioactivity prediction methods is hindered by insufficient high-quality training data.

[0005] To improve bioactivity prediction accuracy, emerging deep learning techniques address data scarcity challenges. Early methods tackled imbalanced samples by using high-quality negative data to distinguish active and inactive molecules. Current approaches include utilizing untested molecule-target pairs as negative examples, leveraging protein family information as priors for transferable inactive ligands, and employing transfer learning and pre-trained neural network layers. Other techniques involve generating adversarial representations and combining cell morphology and chemical structure models. Despite these efforts, many deep learning methods overlook fundamental elements for producing bioactivity, focusing

mainly on Quantitative Structure-Activity Relationship (QSAR) modeling between ligand structures and bioactivity.

[0006] For example, current ligand bioactivity prediction models are based on two prevalent approaches. The first approach includes Ligand Based Virtual Screening (LBVS) models which predict bioactivity solely based on ligand structures and corresponding bioactivities. However, ligand bioactivities are influenced by more than just ligand structures. The other approach includes Structure-Based Virtual Screening (SBVS) models which docks candidates to proteins and estimates their binding affinities. However, it is important to note that achieving a certain threshold of binding affinity is a prerequisite for producing bioactivity, but having a higher binding affinity beyond this threshold does not necessarily guarantee higher bioactivity.

[0007] Bioactivity is determined not only by ligand structure but also by drug-target interactions, for example, binding poses and binding affinities, reaction environments, pharmacokinetic properties, and species differences. However, high-quality data concerning drug-target interactions are scarcer compared to molecular structures. This is also a reason why current deep learning methods predominantly model molecular structures of the drug.

[0008] Therefore, there exists a need for an improved system and/or method for determining the bioactivity of a drug, e.g. ligand, that seeks to address at least one of the aforementioned issues.

SUMMARY

[0009] The disclosure was conceptualized to provide an improved system and method for determining a bioactivity of a ligand, e.g. drug, that circumvents the above difficulties by generating ligand-protein complexes through molecular docking and using model self-attention to precisely identify docking pockets and docking poses. To this end, a Drug-Target, e.g. ligand-protein Interaction Graph Neural Network (DTIGN) is constructed for determining ligand-protein non-covalent parameters which considers the interatomic distances in Coulomb and London forces to capture force field information that reflects binding stability, which is a prerequisite for producing bioactivity. In addition, a machine learning algorithm, specifically a multi-head self-attention mechanism, is employed to further identify binding pockets and poses within molecular docking complexes. To verify the correctness of model attention, crystal structure data were not used for supervised training, demonstrating accurate identification of binding pockets and poses. Lastly, although crystal structure data are sparse,

the present disclosure utilizes these limited high-quality data to conduct semi-supervised learning in the DTIGN to refine model attention and bioactivity prediction.

[0010] In addition, the disclosure provides a new benchmark dataset for evaluating bioactivity prediction models in the context of ligand-protein complexes. This new benchmark dataset is unique as it incorporates a substantial number of ligand-protein complexes generated via molecular docking, and sets it apart from other bioactivity datasets, which mainly focus on ligand structures and from binding affinity datasets that independently utilize ligand structures, target sequences, target structures or binding sites.

[0011] According to a first aspect of the disclosure, there is provided a system for determining, a bioactivity parameter for a ligand, when the ligand is bound to a protein to form a ligand-protein complex, the ligand being bound to the protein in a plurality of docking poses, the system comprising a processor configured to: obtain, a ligand-protein complex structural data indicative of a structure of the ligand-protein complex and a reference bioactivity data of the ligand, the protein and/or the ligand-protein complex; the ligand-protein complex structural data comprising, a plurality of ligand nodes of the ligand, each ligand node having a ligand attribute; a plurality of protein nodes of the protein, each protein node having a protein attribute; a plurality of non-covalent edges connecting a ligand node to a respective protein node, each non-covalent edge having a non-covalent bond attribute; determine, for each docking pose of the plurality of docking poses, a ligand-protein complex interaction parameter indicative of intermolecular interaction data of the ligand-protein complex, in a respective docking pose, wherein determining the ligand-protein complex interaction parameter comprises, determining, for each non-covalent edge, a non-covalent parameter indicative of intermolecular interaction data of a first ligand node, and a first protein node positioned adjacent thereto and bound to the first ligand node by a first non-covalent edge, based on, a first ligand attribute of the first ligand node, a first protein attribute of the first protein node, and a first ligand-protein non-covalent parameter indicative of a binding stability of a type of dispersion force of the first non-covalent edge; and determining, the ligand-protein complex interaction parameter based on an aggregation of the non-covalent parameters, for the plurality of non-covalent edges, of the ligand-protein complex; determine, an aggregated ligand-protein complex interaction parameter indicative of an aggregated intermolecular interaction data of the ligand-protein complex, in the plurality of docking poses; and determine, the bioactivity parameter based on the aggregated ligand-protein complex interaction parameter.

[0012] In various embodiments, the first ligand-protein non-covalent parameter may be based on a predetermined ligand-protein vector representative of a magnitude of a force field corresponding to the type of dispersion force of the first non-covalent edge, as a function of the non-covalent bond attribute of the first non-covalent edge.

[0013] In various embodiments, the predetermined ligand-protein vector may comprise a Coulomb ligand-protein vector comprising a predetermined radial basis function (RBF) parametrized by a Coulomb power value indicative of the magnitude of the force field of a Coulomb dispersion force of the first non-covalent edge. In various other embodiments, the predetermined ligand-protein vector may comprise a London ligand-protein vector comprising a predetermined RBF parametrized by a predetermined London power value indicative of the magnitude of the force field a London dispersion force of the first non-covalent edge.

[0014] In various embodiments, the non-covalent bond attribute of the first non-covalent edge may comprise, a ligand-protein distance attribute representative of a physical distance value between the first ligand node and the first protein node, of the ligand-protein complex. The physical distance value may be indicative of a force field interaction between the first ligand node and the first protein node. In some embodiments, the processor may be further configured to, determine, the ligand-protein distance attribute by, measuring, on the ligand-protein complex structural data, the physical distance value between the first ligand node and the first protein node; comparing, the physical distance value with a predetermined distance threshold value indicative of a presence of the first non-covalent edge; and determining, the presence of the first non-covalent edge based on the comparison of the physical distance value and the predetermined distance threshold value.

[0015] In various embodiments, the ligand-protein complex structural data may further comprise, a plurality of ligand edges binding the plurality of ligand nodes to each other, in the ligand, each ligand edge having a ligand covalent bond chemical attribute, and a ligand covalent bond physical attribute, wherein to determine the ligand-protein complex interaction parameter, the processor is further configured to, determine, for each ligand edge of the plurality of ligand edges, a ligand edge parameter indicative of intramolecular interaction data of the first ligand node and a second ligand node positioned adjacent thereto, and bound to the first ligand node by a first ligand edge, the ligand edge parameter determined based on, the first ligand attribute of the first ligand node, a second ligand attribute of the second ligand node, the ligand covalent bond chemical attribute of the first ligand edge, and a first ligand covalent parameter indicative of a binding stability of the first ligand edge, and determine, the ligand-

protein complex interaction parameter further based on, an aggregation of the ligand edge parameters, for the plurality of ligand edges.

[0016] In various embodiments, the first ligand covalent parameter may be based on a predetermined ligand vector representative of a distance property of the intramolecular covalent bond of the first ligand edge, as a function of the ligand covalent bond physical attribute of the first ligand edge

[0017] In various embodiments, the predetermined ligand vector may comprise a predetermined RBF parametrized by a first power value indicative of the distance property of the intramolecular covalent bond, of the first ligand edge.

[0018] In various embodiments, the ligand covalent bond physical attribute of the first ligand edge may comprise, a ligand distance attribute representative of a physical distance value between the first ligand node and the second ligand node, of the ligand.

[0019] In various embodiments, the ligand-protein complex structural data may further comprise, a plurality of protein edges binding the plurality of protein nodes to each other, in the protein, each protein edge having a protein covalent bond chemical attribute, and a protein covalent bond physical attribute, wherein to determine the ligand-protein complex interaction parameter, the processor is further configured to, determine, for each protein edge of the plurality of protein edges, a protein edge parameter indicative of intramolecular interaction data of the first protein node and a second protein node positioned adjacent thereto, and bound to the first protein node by a first protein edge of the plurality of protein edges, the protein edge parameter determined based on, the first protein attribute of the first protein node, a second protein attribute of the second protein node, the protein covalent bond chemical attribute of the first protein edge, and a first protein covalent parameter indicative of a binding stability of the first protein edge, and determine, the ligand-protein complex interaction parameter further based on, an aggregation of the protein edge parameters, for the plurality of protein edges.

[0020] In various embodiments, the first protein covalent parameter may be based on a predetermined protein vector representative of a distance property of the intramolecular covalent bond of the first protein edge, as a function of the protein covalent bond physical attribute of the first protein edge.

[0021] In various embodiments, the predetermined protein vector may comprise a predetermined RBF parametrized by a second power value indicative of the distance property of the intramolecular covalent bond, of the first protein edge.

[0022] In various embodiments, the protein covalent bond physical attribute of the first protein edge may comprise, a protein distance attribute representative of a physical distance value between the first protein node and the second protein node, of the protein.

[0023] In various embodiments, the processor may further comprise a first machine learning algorithm, the first machine learning algorithm optionally including a deep learning algorithm, such as, but not limited to a neural network-based deep learning algorithm.

[0024] In various embodiments, the processor may further comprise, a second machine learning algorithm, for determining the ligand-protein complex interaction parameter, wherein the second machine learning algorithm optionally comprises, a multi-layer perceptron (MLP).

[0025] In various embodiments, the processor may further comprise a third machine learning algorithm, for determining the aggregated ligand-protein complex interaction parameter, wherein the third machine learning algorithm optionally comprises, a multi-head self-attention algorithm.

[0026] In various embodiments, the bioactivity parameter may comprise at least one of, an EC50 parameter indicative of a half maximal effective concentration of the ligand; and/or an IC50 parameter indicative of a half maximal inhibitory concentration of the ligand.

[0027] According to a second aspect of the disclosure, there is provided a drug discovery platform comprising the system of the first aspect of the disclosure.

[0028] According to a third aspect of the disclosure, there is provided a method for determining, a bioactivity parameter for a ligand, when the ligand is bound to a protein to form a ligand-protein complex, the ligand being bound to the protein in a plurality of docking poses, the method comprising providing a processor for: obtaining, a ligand-protein complex structural data indicative of a structure of the ligand-protein complex and a reference bioactivity data of the ligand, the protein and the ligand-protein complex; the ligand-protein complex structural data comprising, a plurality of ligand nodes of the ligand, each ligand node having a ligand attribute; a plurality of protein nodes, each protein node having a protein attribute; a plurality of non-covalent edges connecting a ligand node to a respective protein node, each non-covalent edge having a non-covalent bond attribute; determining, for each docking pose of the plurality of docking poses, a ligand-protein complex interaction parameter indicative of intermolecular interaction data of the ligand-protein complex, in a respective docking pose, wherein determining the ligand-protein complex interaction parameter comprises, determining, for each non-covalent edge, a non-covalent parameter indicative of intermolecular interaction data of a first ligand node, and a first protein node positioned

adjacent thereto and bound to the first ligand node by a first non-covalent edge, based on, a first ligand attribute of the first ligand node, a first protein attribute of the first protein node, and a first ligand-protein non-covalent parameter indicative of a binding stability of a type of dispersion force of the first non-covalent edge; and determining, the ligand-protein complex interaction parameter based on an aggregation of the non-covalent parameters, for the plurality of non-covalent edges, of the ligand-protein complex; determining, an aggregated ligand-protein complex interaction parameter indicative of an aggregated intermolecular interaction data of the ligand-protein complex, in the plurality of docking poses; and determining, the bioactivity parameter based on the aggregated ligand-protein complex interaction parameter.

[0029] In various embodiments, the first ligand-protein non-covalent parameter may be based on a predetermined ligand-protein vector representative of a magnitude of a force field corresponding to the type of dispersion force of the first non-covalent edge, as a function of the non-covalent bond attribute of the first non-covalent edge.

[0030] In various embodiments, the predetermined ligand-protein vector may comprise a Coulomb ligand-protein vector comprising a predetermined RBF parametrized by a Coulomb power value indicative of the magnitude of the force field of a Coulomb dispersion force of the first non-covalent edge. In various other embodiments, the predetermined ligand-protein vector may comprise, a London ligand-protein vector comprising a predetermined RBF parametrized by a predetermined London power value indicative of the magnitude of the force field a London dispersion force of the first non-covalent edge.

[0031] In various embodiments, the non-covalent bond attribute of the first non-covalent edge may comprise, a ligand-protein distance attribute representative of a physical distance value between the first ligand node and the first protein node, of the ligand-protein complex. The physical distance value may be indicative of a force field interaction between the first ligand node and the first protein node. In some embodiments, determining, the ligand-protein distance attribute may comprise: measuring, on the ligand-protein complex structural data, the physical distance value between the first ligand node and the first protein node; comparing, the physical distance value with a predetermined distance threshold value indicative of a presence of the first non-covalent edge; and determining, the presence of the first non-covalent edge based on the comparison of the physical distance value and the predetermined distance threshold value.

[0032] In various embodiments, the ligand-protein complex structural data may further comprise, a plurality of ligand edges binding the plurality of ligand nodes to each other, in the

ligand, each ligand edge having a ligand covalent bond chemical attribute, and a ligand covalent bond physical attribute, wherein the method further comprises: determining, for each ligand edge of the plurality of ligand edges, a ligand edge parameter indicative of intramolecular interaction data of the first ligand node and a second ligand node positioned adjacent thereto, and bound to the first ligand node by a first ligand edge, the ligand edge parameter determined based on, the first ligand attribute of the first ligand node, a second ligand attribute of the second ligand node, the ligand covalent bond chemical attribute of the first ligand edge, and a first ligand covalent parameter indicative of a binding stability of the first ligand edge, and determining, the ligand-protein complex interaction parameter further based on, an aggregation of the ligand edge parameters, for the plurality of ligand edges.

[0033] In various embodiments, the first ligand covalent parameter may be based on a predetermined ligand vector representative of a distance property of the intramolecular covalent bond of the first ligand edge, as a function of the ligand covalent bond physical attribute of the first ligand edge

[0034] In various embodiments, the predetermined ligand vector may comprise a predetermined RBF parametrized by a first power value indicative of the distance property of the intramolecular covalent bond, of the first ligand edge.

[0035] In various embodiments, the ligand covalent bond physical attribute of the first ligand edge may comprise, a ligand distance attribute representative of a physical distance value between the first ligand node and the second ligand node, of the ligand.

[0036] In various embodiments, the ligand-protein complex structural data may further comprise, a plurality of protein edges binding the plurality of protein nodes to each other, in the protein, each protein edge having a protein covalent bond chemical attribute, and a protein covalent bond physical attribute, wherein the method further comprises: determining, for each protein edge of the plurality of protein edges, a protein edge parameter indicative of intramolecular interaction data of the first protein node and a second protein node positioned adjacent thereto, and bound to the first protein node by a first protein edge of the plurality of protein edges, the protein edge parameter determined based on, the first protein attribute of the first protein node, a second protein attribute of the second protein node, the protein covalent bond chemical attribute of the first protein edge, and a first protein covalent parameter indicative of a binding stability of the first protein edge, and determining, the ligand-protein complex interaction parameter further based on, an aggregation of the protein edge parameters, for the plurality of protein edges.

[0037] In various embodiments, the first protein covalent parameter may be based on a predetermined protein vector representative of a magnitude of a distance property of the intramolecular covalent bond of the first protein edge, as a function of the protein covalent bond physical attribute of the first protein edge.

[0038] In various embodiments, the predetermined protein vector may comprise a predetermined RBF parametrized by a second power value indicative of the distance property of the intramolecular covalent bond, of the first protein edge.

[0039] In various embodiments, the protein covalent bond physical attribute of the first protein edge may comprise, a protein distance attribute representative of a physical distance value between the first protein node and the second protein node, of the protein.

[0040] In various embodiments, the method may further comprise, a first machine learning algorithm, the first machine learning algorithm optionally including a deep learning algorithm, such as, but not limited to a neural network-based deep learning algorithm.

[0041] In various embodiments, the method may further comprise, a second machine learning algorithm, for determining the ligand-protein complex interaction parameter, wherein the second machine learning algorithm optionally comprises, an MLP.

[0042] In various embodiments, the method may further comprise, a third machine learning algorithm, for determining the aggregated ligand-protein complex interaction parameter, wherein the third machine learning algorithm optionally comprises, a multi-head self-attention algorithm.

[0043] In various embodiments, the bioactivity parameter comprises at least one of, an EC50 parameter indicative of a half maximal effective concentration of the ligand; and/or an IC50 parameter indicative of a half maximal inhibitory concentration of the ligand.

[0044] According to a fourth aspect of the disclosure, there is provided a computer readable medium comprising instructions, which when executed by the processor, causes the processor to perform the method of the third aspect of the disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

[0045] The disclosure will be better understood with reference to the detailed description when considered in conjunction with the non-limiting examples and the accompanying drawings, in which:

- FIG. 1 shows an exemplary schematic illustration of a system 100 for determining a bioactivity parameter 180 for a ligand 210, when the ligand 210 is bound to a protein

220 to form a ligand-protein complex 250, the ligand 210 being bound to the protein 220 in a plurality of docking poses;

- FIG. 2 shows an exemplary schematic illustration of a ligand-protein complex structural data 200 indicative of a structure of the ligand-protein complex 250 as a 3D graph;
- FIG. 3 shows a table 300 of exemplary ligand attributes 212a, 214a, 216a, protein attributes 222a, 224a, 226a, 228a, 232a, 234a, ligand covalent bond chemical attributes 215a, 217a, protein covalent bond chemical attributes 221a, 223a, 225a, 227a, 229a, a radial basis function (RBF) $\lambda^{(1)}(d)$ for covalent intramolecular interaction forces between the ligand nodes 212, 214, 216, a radial basis function (RBF) $\lambda^{(1)}(d)$ for covalent intramolecular interaction forces between the protein nodes 222, 224, 226, 228, 232, 234, and two radial basis functions $\lambda^{(-2)}(d)$ and $\lambda^{(-6)}(d)$ for non-covalent intermolecular interaction forces between the ligand nodes 212, 214, 216 and the respective protein nodes 222, 224, 226, 228, 232, 234 of the ligand-protein complex 250;
- FIG. 4 shows an exemplary architecture 400 of a ligand-protein, e.g. drug-target, complex 250 graph neural network (DTIGN), for determining the ligand-protein complex parameter 130;
- FIG. 5 shows an exemplary flowchart of a method 500 for determining, a bioactivity parameter for a ligand, when the ligand is bound to a protein to form a ligand-protein complex, the ligand being bound to the protein in a plurality of docking poses;
- FIG. 6 shows an exemplary schematic illustration of the overall DTIGN 600 for bioactivity prediction based on drug-target interactions;
- FIG. 7 shows the docking pocket selection 700 for an exemplary ligand, CHEMBL202;
- FIG. 8 shows a table 800 of comparison of bioactivity (pIC50 towards CHEMBL202) prediction performance on different data sources, AutoDock Vina and QVina-W, where the ↑ arrows or ↓ arrows denote that larger or smaller values are better;
- FIG. 9 shows a table of the benchmark datasets 900 of bioactivity prediction on ligand-protein complexes;
- FIG. 10 shows a table indicative of the ablation study 1000 on the I1 task (Dihydrofolate Reductase) in the benchmark dataset;

- FIG. 11 shows the attention values 1100 on protein pockets for a test ligand during the training process. The native binding sites depicted above are sourced from the RCSB PDB database. Their details and corresponding pockets are given in the legend;
- FIG. 12 shows a table indicative of the intersection over union 1200 between the "Native Pockets" and "Attention Pockets";
- FIG. 13 shows the box plot 1300 of DTIGN's attention distributions on the ranked docking poses involved in the pIC50 test set (containing 9,975 redocked poses on 1,011 PDB native structures). "Avg RMSD" represents the average RMSD between ligand-protein docking poses and their respective native structures in the PDBBind dataset;
- FIG. 14 shows the box plot 1400 of DTIGN's attention distributions on the ranked docking poses involved in the pKd test set (containing 14,411 re-docked poses on 1,455 PDB native structures). "Avg RMSD" represents the average RMSD between ligand-protein docking poses and their respective native structures in the PDBBind dataset;
- FIG. 15 shows the box plot 1500 of DTIGN's attention distributions on the ranked docking poses involved in the pKi test set (containing 6,196 re-docked poses on 626 PDB native structures). "Avg RMSD" represents the average RMSD between ligand-protein docking poses and their respective native structures in the PDBBind dataset;
- FIG. 16 shows a table of the best cross-validated Pearson's r 1600 on the benchmark datasets;
- FIG. 17 shows a table of the best cross-validated RMSE 1700 on the benchmark datasets;
- FIG. 18 shows a table of the best cross-validated τ_B 1800 on the benchmark datasets;
- FIG. 19 shows the binding poses 1900 between exemplary target protein CHEMBL202 and the test ligand CHEMBL7492 in the: (A) native pose and (B) AutoDock Vina-docked structure in the pocket the present model pays attention to (pocket 2). Non-covalent bonds (polar interactions) between ligand and protein are marked by dashed lines, with binding sites highlighted on the protein sequences above. Common binding sites across native and docking structures are encircled in dashed lines; and
- FIG. 20 shows (A) "CHEMBL7492" in "1u72"; and (B) "CHEMBL22" in "2w3a" indicating the test ligand ID, and its PDB ID with the native structure of the protein, indicative that the DTIGN identifies some native-like docking poses to learn from, which show a RMSD of less than 5Å from the native structure. The correlation between

RMSD and average attention value during model training was evaluated using the Pearson correlation coefficient (r).

DETAILED DESCRIPTION

[0046] The following detailed description refers to the accompanying drawings that show, by way of illustration, specific details and embodiments in which the disclosure may be practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice the disclosure. Other embodiments may be utilized and structural, and logical changes may be made without departing from the scope of the disclosure. The various embodiments are not necessarily mutually exclusive, as some embodiments can be combined with one or more other embodiments to form new embodiments.

[0047] Features that are described in the context of an embodiment may correspondingly be applicable to the same or similar features in the other embodiments. Features that are described in the context of an embodiment may correspondingly be applicable to the other embodiments, even if not explicitly described in these other embodiments. Furthermore, additions and/or combinations and/or alternatives as described for a feature in the context of an embodiment may correspondingly be applicable to the same or similar feature in the other embodiments.

[0048] In the context of various embodiments, the articles “a”, “an” and “the” as used with regard to a feature or element include a reference to one or more of the features or elements. As used herein, the term “and/or” includes any and all combinations of one or more of the associated listed items.

[0049] While terms as "first," "second," etc., may be used to describe various elements, such terms are used only to distinguish one element from another for purposes of clarity, and do not define corresponding elements, for example, an order and/or significance of the elements. Without departing a scope of rights of the specification, a first element may be referred to as a second element, and similarly, the second element may be referred to as the first element.

[0050] Throughout the description, the term “ligand”, as used herein, may refer to a compound known to be potent against a target, e.g. a protein. The compound may be a small molecule, having a molecular weight of less than or equal to 900 Da. In various embodiments, the ligand may refer to a drug, e.g. a pharmaceutical drug. In some embodiments, the ligand may be a ligand antagonist, and may bind to the target protein to block the function of the target protein. In some other embodiments, the ligand may be a ligand agonist, and may bind to the target protein to activate the function of the target protein. Within the context of the disclosure,

term “ligand molecule” may refer to the ligand which may comprise a plurality of ligand atoms, the ligand atoms being bound to each other by ligand covalent bonds.

[0051] Throughout the description, the term “protein”, as used herein, may refer to a target that having a specific function or behavior. The ligand may bind to the protein to form a ligand-protein complex, to alter, for example, inhibit or activate the function or behavior of the protein. In various embodiments, the protein may comprise at least one docking pocket comprising at least one binding site, for binding to the ligand. The binding sites may be categorized into the various docking pockets of the protein based on spatial clustering of the binding sites. Each binding site may comprise at least one receptor, which binds to the ligand. Within the context of the disclosure, the term “protein molecule” may refer to the protein comprise a plurality of protein atoms, the protein atoms being bound to each other by protein covalent bonds. While the term “protein” is used in the disclosure, it is contemplated that the target may also comprise a gene, a ribonucleic acid (RNA), for example, mRNA, ncRNA.

[0052] Throughout the description, the term “ligand-protein complex”, may refer to the complex formed when the ligand is bound to at least one binding site, specifically, a receptor, of the protein. In various embodiments, the ligand may be bound to the at least one binding site of the protein in a plurality of docking poses. Within the context of the disclosure, the term “ligand-protein complex molecule” may refer to the ligand molecule having a ligand atom that may be bound to at least one protein atom of the protein molecule by intermolecular non-covalent bonds.

[0053] Throughout the description, the term “bioactivity”, as used herein, may refer to the ability of the ligand molecule to elicit, induce or modulate a specific biological effect or cellular response in an organism, e.g. a living organism or a tissue. The bioactivity of a ligand molecule may be described with respect to the target protein molecule, and may be induced when the ligand molecule is bound to a receptor comprised in at least one binding site of the protein molecule. In some embodiments, the bioactivity of the ligand molecule may be inhibitory, e.g. via a ligand antagonist. In some other embodiments, the bioactivity of the ligand molecule may activate a cellular response, e.g. via a ligand agonist.

[0054] Throughout the description, the term “adjacent”, as used herein, may refer to the positioning of a ligand or protein atom directly adjacent to another ligand or protein atom, respectively. The term “adjacent”, further refers to the positioning of a ligand atom directly adjacent to a protein atom. That is, there are no intervening atoms between the ligand atom and the adjacent ligand atom, a protein atom and the adjacent protein atom, and a ligand atom and

the adjacent protein atom. Accordingly, the term “intramolecular interaction data” may refer to interaction data between the ligand or protein atom and another ligand or protein atom positioned directly adjacent thereto, respectively. In addition, the term “intermolecular interaction data” may refer to interaction data between the ligand atom and the protein atom positioned directly adjacent thereto.

[0055] Throughout the description, the phrase “a function of”, may include a mathematical function describing a relationship between an input, and an output. In various embodiments, the input may comprise: a ligand distance attribute representative of a physical distance value between two ligand nodes covalently bonded to each other; a protein distance attribute representative of a physical distance value between two protein nodes covalently bonded to each other. Accordingly, the output may comprise: a ligand covalent parameter indicative of the intramolecular interaction data between the two covalently bonded ligand nodes, and a protein covalent parameter indicative of the intramolecular interaction data between two covalently bonded protein nodes, respectively. In various other embodiments, the input may comprise a ligand-protein distance attribute representative of physical distance value of a non-covalent bond binding a ligand node to a respective protein node. The physical distance value of the non-covalent bond may be indicative of force field interactions between the ligand node and the respective protein node. The output may comprise the ligand-protein non-covalent parameter indicative of the intermolecular interaction data, e.g. force field data, between the ligand node and the respective protein node.

[0056] Throughout the description, the term “data” may be understood to include information in any suitable analogue or digital form, for example, provided as a file, a portion of a file, a set of files, a signal or stream, a portion of a signal or stream, a set of signals or streams, waveforms, and the like. The term data, however, is not limited to the aforementioned examples and may take various forms and represent any information as understood in the art.

[0057] According to various embodiments, a circuit may include analog circuits or components, digital circuits or components, or hybrid circuits or components. Any other kind of implementation of the respective functions which will be described in more detail below may also be understood as a "circuit" in accordance with an alternative embodiment. A digital circuit may be understood as any kind of a logic implementing entity, which may be special purpose circuitry or a processor executing software stored in a memory, firmware, or any combination thereof. Thus, in various embodiments, a processor may include one or more circuits. A "circuit" may be a digital circuit, e.g. a hard-wired logic circuit or a programmable

logic circuit such as a programmable processor, e.g. a microprocessor (e.g. a Complex Instruction Set Computer (CISC) processor or a Reduced Instruction Set Computer (RISC) processor). A "circuit" may also include a processor executing software, e.g. any kind of computer program, e.g. a computer program using a virtual machine code such as e.g. Java.

[0058] FIG. 1 shows an exemplary schematic illustration of a system 100 for determining a bioactivity parameter 180 for a ligand 210, when the ligand 210 is bound to a protein 220 to form a ligand-protein complex 250, the ligand 210 being bound to the protein 220 in a plurality of docking poses.

[0059] Referring to FIG. 1, system 100 includes a processor 110 is configured to: obtain, a ligand-protein complex structural data 200 indicative of a structure of the ligand-protein complex 250 as a 3D graph, and a reference bioactivity data of the ligand 102, reference bioactivity data of the protein 104 and reference bioactivity data of the ligand-protein complex 106; determine, for each docking pose of the plurality of docking poses, a ligand-protein complex interaction parameter 130 indicative of intermolecular interaction data of the ligand-protein complex 250, in a respective docking pose; determine, an aggregated ligand-protein complex interaction parameter 170 indicative of an aggregated intermolecular interaction data of the ligand-protein complex 250, in the plurality of docking poses; and determine, the bioactivity parameter 180 based on the aggregated ligand-protein complex interaction parameter 170.

[0060] In various embodiments, the bioactivity parameter 180 may be indicative of a predicted or estimated induced biological effect on an organism. In some embodiments, the ligand 210 may be a ligand agonist and the bioactivity parameter may refer to an EC50 parameter indicative of a half maximal effective concentration of the ligand 210, after a specific exposure time. The EC50 parameter may comprise a value representing a concentration at which the ligand exerts 50% of its maximal biological effect, after a specific exposure time. In some other embodiments, the ligand 210 may be a ligand antagonist and the bioactivity parameter may refer to an IC50 parameter indicative of a half maximal inhibitory concentration of the ligand 210. The IC50 parameter may comprise a measure of the concentration of the ligand 210 required to inhibit a particular biological effect by 50%, after a specific exposure time. Other non-limiting examples of the bioactivity parameter 180 may comprise: an inhibitory constant K_i indicative of a measure of a ligand's ability to inhibit a biological effect; a dissociation constant K_d indicative of a ratio of the ligand's association and disassociation

with a receptor of the protein; an association (binding or affinity) constant K_a indicative of a measure of how the ligand and protein binds with each other; an elimination constant K_e indicative of a measure of the rate at which the ligand is removed from the organism; an inhibition index, reflecting the measure of inhibitory activity; a hit score, representing the success rate of the ligand in a screening assay; a CC50 value, indicating the concentration of a compound causing 50% cytotoxicity; a selectivity index (SI), measuring the ligand's specificity for a target compared to off-target interactions; an activity value, quantifying the ligand's overall efficacy; an area under the curve (AUC), representing the integral of response over a range of concentrations or times; and an AUC ratio, comparing AUC values across different conditions or treatments.

Ligand-protein complex structural data 200

[0061] FIG. 2 shows an exemplary schematic illustration of a ligand-protein complex structural data 200 indicative of a structure of the ligand-protein complex 250, depicted as a 3D graph. FIG. 3 shows a table 300 of exemplary ligand attributes 212a, 214a, 216a, protein attributes 222a, 224a, 226a, 228a, 232a, 234a, ligand covalent bond chemical attributes 215a, 217a, protein covalent bond chemical attributes 221a, 223a, 225a, 227a, 229a, a radial basis function (RBF) $\lambda^{(1)}(d)$ for covalent intramolecular interaction forces between the ligand nodes 212, 214, 216, a radial basis function (RBF) $\lambda^{(1)}(d)$ for covalent intramolecular interaction forces between the protein nodes 222, 224, 226, 228, 232, 234, and two radial basis functions (RBF) $\lambda^{(-2)}(d)$ and $\lambda^{(-6)}(d)$ for non-covalent intermolecular interaction forces between the ligand nodes 212, 214, 216 and the respective protein nodes 222, 224, 226, 228, 232, 234 of the ligand-protein complex 250.

[0062] In various embodiments, the ligand-protein complex structural data 200, and the reference bioactivity data of the ligand 102, the reference bioactivity data protein 104, and the reference bioactivity data ligand-protein complex 106, may be obtained from a memory 120 in data communication with the processor 110. In some embodiments, the processor 110 may be configured to: obtain, data indicative of a structure of the ligand-protein complex 250 from the memory 120, and represent, data indicative of the structure of the ligand-protein complex 250 as a 3D graph, as detailed below.

[0063] The processor 110 may comprise wireless communication means to obtain, the ligand-protein complex structural data 200 and/or the data indicative of a structure of the ligand-protein complex 250 from the memory 120, based on one or more predefined wireless

protocols, Examples of the pre-defined wireless communication protocols include: global system for mobile communication (GSM), enhanced data GSM environment (EDGE), wideband code division multiple access (WCDMA), code division multiple access (CDMA), time division multiple access (TDMA), wireless fidelity (Wi-Fi), voice over Internet protocol (VoIP), worldwide interoperability for microwave access (Wi-MAX), Wi-Fi direct (WFD), an ultra-wideband (UWB), infrared data association (IrDA), Bluetooth, ZigBee, SigFox, LPWan, LoRaWan, GPRS, 3G, 4G, LTE, 5G communication systems and 6G communication systems. Alternatively, the processor 110 may obtain the ligand-protein complex structural data 200 and/or the data indicative of a structure of the ligand-protein complex 250 through wired means (e.g. LAN or Ethernet cable).

[0064] Referring to FIGS. 2 and 3, the structure of the ligand-protein complex 250 as a 3D graph may comprise, the ligand molecule being represented as a ligand 210 comprising a first ligand node 212, a second ligand node 214 and a third ligand node 216, each of the first, second and third ligand nodes 212, 214, 216 corresponding to a ligand atom of the ligand molecule. Each of the first, second and third ligand nodes 212, 214, 216 may be positioned adjacent to each other and may be covalently bonded to each other. Each of the first, second and third ligand nodes 212, 214, 216 may also comprise a respective first, second and third ligand attribute 212a, 214a, 216a, for example, a ligand chemical feature, which may be obtained from the reference bioactivity data of the ligand 102. In various embodiments, the first, second and third ligand attribute 212a, 214a, 216a, may include data indicative of, a respective ligand atom type; a number of covalent bonds of the ligand molecule; an implicit valence of the ligand atom; a hybridization of the ligand atom; whether the ligand atom is part of the aromatic system; a number of connected hydrogens of the ligand atom. Each of the first, second and third ligand nodes 212, 214, 216 may further have a unique 3D coordinate indicative of the position of the first, second and third ligand node 212, 214, 216 on the 3D graph.

[0065] The intramolecular covalent bonds which bind the ligand atoms to each other, may be represented a plurality of ligand edges 215, 217 which binds the first, second and third ligand nodes 212, 214, 216 to each other. For example, a first ligand edge 215 may bind the first ligand node 212 to a second ligand node 214, and a second ligand edge 217 may bind the first ligand node 212 to the third ligand node 216. Each of the first and second ligand edge 215, 217 may comprise, a first and second ligand covalent bond chemical attribute 215a, 217a, respectively, which may be obtained from the reference bioactivity data of the ligand 102. In various embodiments, the first and second ligand covalent bond chemical attribute 215a, 217a, may

include data indicative of, a bond type of the covalent bond; whether the bond is conjugated; whether the bond comprises a ring; a steric nature of the covalent bond. Each of the first and second ligand edge 215, 217 may further comprise, a first and second ligand covalent bond physical attribute 215b, 217b, respectively.

[0066] In various embodiments, the first and second ligand covalent bond physical attribute 215b, 217b may comprise, a ligand distance attribute representative of a physical distance value between a ligand atom and another ligand atom. For example, the first ligand covalent bond physical attribute 215b may comprise a physical distance value between the first ligand node 212 and the second ligand node 214, and the second ligand covalent bond physical attribute 217b may comprise a physical distance value between the first ligand node 212 and the third ligand node 216. In some embodiments, the processor 110 may be configured to determine the first and second ligand covalent bond physical attribute 215b, 217b, by measuring, on the ligand-protein complex structural data 200, the physical distance between the first ligand node 212 and the second ligand node 214, and the first ligand node 212 and the third ligand node 216, respectively.

[0067] Referring to FIGS. 2 and 3, the structure of the ligand-protein complex 250 as a 3D graph may further comprise, the protein molecule being represented as a protein 220 comprising a first protein node 222, a second protein node 224, a third protein node 226, a fourth protein node 228, a fifth protein node 232 and a sixth protein node 234, each of the first to sixth protein node 222, 224, 226, 228, 232, 234, corresponding to a protein atom of the protein molecule. Each of the first to sixth protein node 222, 224, 226, 228, 232, 234 may be positioned adjacent to each other and may be covalently bonded to each other. Each of the first to sixth protein nodes 222, 224, 226, 228, 232, 234, may also comprise a respective of the first to sixth protein attribute 222a, 224a, 226a, 228a, 232a, 234a, for example, a protein chemical feature, which may be obtained from the reference bioactivity data of the protein 104. In various embodiments, the first to sixth protein attribute 222a, 224a, 226a, 228a, 232a, 234a, may include data indicative of, a respective protein atom type; a number of covalent bonds of the protein molecule; an implicit valence of the protein atom; a hybridization of the protein atom; whether the protein atom is part of the aromatic system; a number of connected hydrogens of the protein atom. Each of the first to sixth protein nodes 222, 224, 226, 228, 232, 234, may further have a unique 3D coordinate indicative of the position of the respective one of the first to sixth protein nodes, 222, 224, 226, 228, 232, 234, on the 3D graph.

[0068] In various embodiments, the intramolecular covalent bonds which bind the protein atoms to each other, may be represented as a plurality of protein edges 221, 223, 225, 227, 229 which binds the first to sixth protein nodes 222, 224, 226, 228, 232, 234 to each other. For example, a first protein edge 221 may bind the first protein node 222 to the second protein node 224; a second protein edge 223 may bind the second protein node 224 to a third protein node 226; a third protein edge 225 may bind the third protein node 226 to a fourth protein node 228; a fourth protein edge 227 may bind the fourth protein node 228 to a fifth protein node 232; and a fifth protein edge 229 may bind the fifth protein node 232 to the sixth protein node 234. Each of the first to fifth protein edges 221, 223, 225, 227, 229 may comprise, a first to fifth protein covalent bond chemical attribute 221a, 223a, 225a, 227a, 229a, respectively, which may be obtained from the reference bioactivity data of the protein 104. In various embodiments, the first to fifth protein covalent bond chemical attribute 221a, 223a, 225a, 227a, 229a, may include data indicative of, a bond type of the covalent bond; whether the bond is conjugated; whether the bond comprises a ring; a steric nature of the covalent bond.

[0069] Each of the first to fifth protein edges 221, 223, 225, 227, 229 may further comprise a first to fifth protein bond physical attribute 221b, 223b, 225b, 227b, 229b, respectively. In various embodiments, the first to fifth protein covalent bond physical attribute 221b, 223b, 225b, 227b, 229b may comprise, a protein distance attribute representative of a physical distance value between a protein atom and another protein atom. For example, the first protein covalent bond physical attribute 221b may comprise a physical distance value between the first protein node 222 and the second protein node 224; the second protein covalent bond physical attribute 223b may comprise a physical distance value between the second protein node 224 and the third protein node 226; the third protein covalent bond physical attribute 225b may comprise a physical distance value between the third protein node 226 and the fourth protein node 228; the fourth protein covalent bond physical attribute 227b may comprise a physical distance value between the fourth protein node 228 and the fifth protein node 232; and the sixth protein covalent bond physical attribute 229b may comprise a physical distance value between the fifth protein node 232 and the sixth protein node 234. In some embodiments, the processor 110 may be configured to determine the first to fifth protein covalent bond physical attribute 221b, 223b, 225b, 227b, 229b, by measuring, on the ligand-protein complex structural data 200, the physical distance between the first and second protein nodes 222, 224, the second and third protein nodes 224, 226, the third and fourth protein nodes 226, 228, the fourth and fifth protein nodes 228, 232, the fifth and sixth protein nodes 232, 234, respectively.

[0070] Referring to FIGS. 2 and 3, the structure of the ligand-protein complex 250 as a 3D graph may further comprise, a plurality of intermolecular non-covalent bonds binding a ligand atom to a respective protein atom positioned adjacent to the ligand atom, being represented as a plurality of non-covalent edges 252, 254. For example, the first ligand node 212 may be connected to the first protein node 222 via a first non-covalent edge 252, and may be connected to the sixth protein node 234 via a second non-covalent edge 254. Each of the first and second non-covalent edges 252, 254 may comprise, a first and second non-covalent bond attribute 252b, 254b. In various embodiments, a first non-covalent bond attribute 252b may comprise a ligand-protein distance attribute representative of a physical distance value between the first ligand node 212 and the first protein node 222; and the second non-covalent bond attribute 254b may comprise a physical distance value between the first ligand node 212 and the sixth protein node 234. The physical distance value of the first non-covalent bond attribute 252b may be indicative of a force field interaction between the first ligand node 212 and the first protein node 222; and the second non-covalent bond attribute 254b may be indicative of a force field interaction between the first ligand node 212 and the sixth protein node 234, respectively. The ligand-protein distance attribute may be multiplied with a predetermined ligand-protein vector which accounts for the different type dispersion forces, specifically, the Coulomb and London dispersion forces of the intermolecular non-covalent bonds.

[0071] In various embodiments, the processor 110 may be configured to determine the first non-covalent bond attribute 252b by: measuring, on the ligand-protein complex structural data 200, the physical distance value between the first ligand node 212 and the first protein node 222; comparing, the physical distance value with a predetermined distance threshold value indicative of a presence of the first non-covalent edge 252; and determining, the presence of the first non-covalent edge 252 based on the comparison of the physical distance value and the predetermined distance threshold value. In some embodiments, the first and second non-covalent bond attribute 252b, 254b, may be obtained from the reference bioactivity data of the ligand-protein complex 106.

[0072] In various embodiments, the predetermined distance threshold value may comprise a distance of 5Å, and determining, the presence of the first non-covalent edge 252 may comprise: determining, whether the physical distance value is less than predetermined distance threshold value of 5Å, which is indicative of the presence of the first non-covalent edge 252. In other words, the first non-covalent edge 252 exists if the physical distance value between

the first ligand node 212 and the first protein node 222, is less than the predetermined distance threshold value.

[0073] Referring to FIGS. 1 to 3, to capture the overall ligand-protein complex 250 interaction information, the structure of the ligand-protein complex 250 as a 3D graph may be represented as $\mathcal{G} = (V, \mathcal{E}, \mathcal{R}) = (V_l \cup V_p, \mathcal{E}_l \cup \mathcal{E}_p \cup \mathcal{E}_{lp}, \mathcal{R}_l \cup \mathcal{R}_p)$, where graph nodes, V_l corresponds to ligand nodes, e.g. the first, second and third ligand nodes 212, 214, 216 representative of the ligand atoms; graph nodes V_p correspond to protein nodes, e.g. the first to sixth protein nodes 222, 224, 226, 228, 232, 234 representative of the protein atoms; graph edges $\mathcal{E}_l$ correspond to ligand edges, e.g. first and second ligand edges 215, 217 representative of the intramolecular covalent bonds of the ligand 210; graph edges $\mathcal{E}_p$ corresponds to protein edges, e.g. first to fifth protein edges 221, 223, 225, 227, 229 representative to the intramolecular covalent bonds of the protein 220; and graph edges $\mathcal{E}_{lp}$ corresponds to the non-covalent edges, e.g. first and second non-covalent edges 252, 254 representative of the intermolecular non-covalent bonds binding the ligand 210 to the protein 220. In addition, $\mathcal{R}_l$ represents the 3D coordinates of the ligand nodes, e.g. the first, second and third ligand nodes 212, 214, 216 in the 3D graph, and $\mathcal{R}_p$ represents the 3D coordinates of the protein nodes, e.g. first to sixth protein nodes 222, 224, 226, 228, 232, 234 in the 3D graph.

[0074] Each of the first, second and third ligand attributes 212a, 214a, 216a, and the first to sixth protein attributes 222a, 224a, 226a, 228a, 232a, 234a, may be represented as $\boldsymbol{x}_i^{(a)} \in \mathbb{R}^{n(a)}$, and is associated with its respective 3D coordinates $\boldsymbol{r}_i \in \mathcal{R}$. The first and second ligand edges 215, 217 and first to fifth protein edges 221, 223, 225, 227, 229 may be represented as $e_{ji} \in \mathcal{E}_l \cup \mathcal{E}_p$, representative of the covalent bond which exist between the ligand or protein atom v_i and another respective ligand or protein atom v_j. The first and second ligand covalent bond chemical attribute 215a, 217a of the first and second ligand edges 215, 217, and the first to fifth protein covalent bond chemical attribute 221a, 223a, 225a, 227a, 229a of the first to fifth protein edges 221, 223, 225, 227, 229 may be represented as $\boldsymbol{x}_{ji}^{(b)} \in \mathbb{R}^{n(b)}$.

[0075] The first and second non-covalent edges 252, 254 may be represented as $e_{ki} \in \mathcal{E}_{lp}$, and the first and second non-covalent bond attribute 252b, 254b may be represented as $d_{ik} = |r_i - r_k|$, indicative of a physical distance value between a ligand atom v_i and a respective protein atom v_k non-covalently bonded thereto. The first and second non-covalent bond attribute 252b, 254b may exist when the physical distance value is less than 5Å.

Ligand-protein complex interaction parameter 130

[0076] FIG. 4 shows an exemplary architecture 400 of a ligand-protein, e.g. drug-target, complex 250 graph neural network (DTIGN), for determining the ligand-protein complex parameter 130.

[0077] Referring to FIGS. 1 to 4, the processor 110 is configured to: determine, for each docking pose, a ligand-protein complex interaction parameter 130 indicative of intermolecular interaction data of the ligand-protein complex 250. The ligand-protein complex interaction parameter 130 may be indicative of the overall ligand-protein complex 250 interaction data, which includes by data of the intramolecular bond features which define the intramolecular interactions between the atoms of the ligand molecule, and between the atoms of the protein molecule, and intermolecular bond features which define intermolecular interactions between the atoms of the ligand molecule and the protein molecule in the ligand-protein complex molecule.

[0078] The ligand-protein complex parameter 130, for each docking pose, may comprise, an overall ligand edge parameter 132 indicative of intramolecular interaction data of the ligand covalent bonds which bind the plurality of ligand atoms to each other; an overall protein edge parameter 134 indicative of intramolecular interaction data of the protein covalent bonds which bind the plurality of protein atoms to each other; and an overall non-covalent parameter 136 indicative of intermolecular interaction data between a ligand atom and a respective protein atom of the ligand-protein complex molecule, to capture the overall ligand-protein complex molecule interaction information.

[0079] This may be represented based on a first machine learning algorithm, optionally comprising a deep learning algorithm, such as, but not limited to a neural network-based deep learning algorithm. In various embodiments, the deep learning algorithm may include a graph neural network, and may optionally include a graph convolutional network (GCN). The graph neural network may comprise a message passing (MP) algorithm of the DTIGN which may be formed by stacking two sets of blocks, a first block 410 for the covalent interactions, which includes the overall ligand edge parameter 132 and the overall protein edge parameter 134, and a second block 420 for the non-covalent interactions, which include the overall non-covalent parameter 136, repeated T times, and may be defined in Equation (1),

$$DTIGN(\mathcal{G}) = MLP^{(bio)}\left\{\sum_V\left[MP_T^{(cov)}(V,\mathcal{E},\mathcal{R}) + MP_T^{(ncov)}(V,\mathcal{E},\mathcal{R})\right]\right\} \quad \text{Equation (1),}$$

where a second machine learning algorithm, such as the MLP $MLP^{(bio)}$ which predicts the bioactivity parameter 180 by aggregating representations $\boldsymbol{f}$ from two MP phases. In Equation (1), the covalent and non-covalent messages on all atoms are summed to distinguish molecular sizes and normalized in the batch normalization of the MLP.

[0080] The properties of atoms involved in forming covalent bonds and the distances between them play a crucial role in covalent bond formation. Referring to FIGS. 1 and 4, the first block 410 for the covalent interactions may include the overall ligand edge parameter 132. To determine the overall ligand edge parameter 132, the processor 110 may be configured to: determine, for the first ligand edge 215, a first ligand edge parameter 142 indicative of intramolecular interaction data of the first ligand node 212 and the second ligand node 214 positioned adjacent to the first ligand node 212 and bonded to the first ligand node 212 by the first ligand edge 215, based on the first and second ligand attribute 212a, 214a, the first ligand covalent bond chemical attribute 215a, and a first ligand covalent parameter 144 indicative of the binding stability of the first ligand edge 215.

[0081] The first ligand covalent parameter 144 may be based on a predetermined ligand vector representative of a distance property of the intramolecular covalent bond of the first ligand edge 215, as a function of the first ligand covalent bond physical attribute 215b. The first ligand covalent bond physical attribute 215b may comprise the ligand distance attribute that is representative of the physical distance value between the first ligand node 212 and the second ligand node 214, and the predetermined ligand vector may comprise a predetermined RBF $\lambda^{(l)}(d)$, parameterized by a first power value indicative of the distance property of the intramolecular covalent bond of the first ligand edge 215.

[0082] The predetermined RBF may be a vector from a set of Gaussian RBFs whose inputs is the distance *d* raised to the power of *l*, which may be expressed according to Equations (2) and (3),

$$\lambda^{(l)}(d) = \left[exp\left(-\frac{1}{(\phi_1-\phi_0)^2}\left\|d^l-\phi_0\right\|^2\right),\ldots,exp\left(-\frac{1}{(\phi_{ndim}-\phi_{ndim-1})^2}\left\|d^l-\phi_{ndim}\right\|^2\right)\right] \quad \text{Equation (2)}$$

$$\phi_k = \begin{cases}[d_{min}+k\cdot(d_{max}-d_{min})/(n_{dim}-1)]^l \text{ if } l \geq 0,\\ [d_{max}-k\cdot(d_{max}-d_{min})/(n_{dim}-1)]^l \text{ if } l < 0\end{cases} \quad \text{Equation (3),}$$

where in Equation (2), the mean ϕ_k of the RBFs used in the disclosure is equally sampled between the minimum d_{min} and maximum d_{max} distances, and then raised to the power of *l*; and the variance $(\phi_k-\phi_{k-1})^2$ of the RBFs may be determined by the difference between adjacent means. This configuration allows the RBFs to effectively capture different ranges of

the distance depending on the value of l. For interatomic interactions, d_{min} may be set at 1Å to denote the minimal atomic distance in a stable state, and d_{max}may be set at 6Å to capture significant atomic interactions within this radius. In addition, n_{dim} specifies the number of RBFs, where $n_{dim} = 64$ may be chosen to cover the distance between atoms from 1Å to 6Å.

[0083] Based on Equations (2) and (3) and to capture the distance property of the intramolecular covalent bond of the first ligand edge 215, the first power value may comprise a value equal to 1, i.e. $l = 1$; and the first ligand covalent parameter 144 may be represented as $\lambda^{(1)}(\|r_j - r_i\|)$, where $r_j - r_i$ refers to the ligand distance attribute representative of the physical distance value between the first 212 and second 214 ligand node. Accordingly, the first ligand edge parameter 142 may be defined according to Equation (4),

$$m_{ji}^{(cov)} = W^{(nbr)}\left[h_j^{(t)}, h_{ji}\right] \odot W^{(cov)}\left[\lambda^{(1)}(\|r_j - r_i\|)\right], \forall\, e_{ji} \in \mathcal{E}_l \cup \mathcal{E}_p \quad \text{Equation (4),}$$

where the intramolecular interaction data, i.e. $m_{ji}^{(cov)}$ between two covalently connected nodes may be defined as the element-wise multiplication ($\odot$) of two components: one may be the concatenated and dimensionally reduced representations of adjacent atoms and bond features denoted as $W^{(nbr)}[.,.]$, and the other may be the up sampled distance features between adjacent atom, denoted as $W^{(cov)}\left[\lambda^{(1)}(.)\right]$. The vector $h_j^{(t)}$ may represent the hidden features of the atom v_j at the *t*-th layer, and its initial value is obtained from the atom features $\boldsymbol{x}_j^{(a)}$ through a single-layer neural network, denoted as $h_{ji} = W^{(b)}\boldsymbol{x}_{ji}^{(b)}$.

[0084] The processor 110 may be configured to further determine, for the second ligand edge 217, a second ligand edge parameter 146 indicative of intramolecular interaction data of the first ligand node 212 and the third ligand node 216 positioned adjacent to the first ligand node 212 and bonded to the first ligand node 212 by the second ligand edge 217, based on the first and third ligand attribute 212a, 216a, the second ligand covalent bond chemical attribute 217a, and a second ligand covalent parameter 148 indicative of the binding stability of the second ligand edge 217. The second ligand covalent parameter 148 may be determined in the same manner as the first ligand covalent parameter 144, and the second ligand edge parameter 146 may be determined in the same manner as the first ligand edge parameter 142, as described above.

[0085] Accordingly, the processor 110 may determine, the overall ligand edge parameter 132 based on an aggregation of the ligand edge parameter, for the plurality of ligand edges of the ligand 210. For example, the overall ligand edge parameter 132 may be based on the sum

of the first ligand edge parameter 142 and the second ligand edge parameter 146. The overall ligand edge parameter 132 indicative of the intramolecular interaction data of the ligand covalent bonds which bind the plurality of ligand atoms to each other may also be represented according to Equation (5),

$$m_i^{(cov)} = \sum_{v_j \in N(v_i)} m_{ji}^{(cov)}, \forall\, e_{ji} \in \mathcal{E}_l \cup \mathcal{E}_p \qquad \text{Equation (5).}$$

[0086] Referring to FIGS. 1 and 4, the first block 410 for the covalent interactions may further include the overall protein edge parameter 134. That is, the ligand-protein complex parameter 130, for each docking pose, may further comprise, an overall protein edge parameter 134 indicative of intramolecular interaction data of the protein covalent bonds which bind the plurality of protein atoms to each other. The overall protein edge parameter 134 may be determined in a similar manner to the overall ligand edge parameter 132, as described above.

[0087] Briefly, the processor 110 may be configured to: determine, for the first protein edge 221, a first protein edge parameter 151 indicative of intramolecular interaction data of the first protein node 222 and the second protein node 224 positioned adjacent to the first protein node 222, and bound to the first protein node 222 by the first protein edge 221. The first protein edge parameter 151 may be determined based on, the first protein attribute 222a of the first protein node 222, the second protein attribute 224a of the second protein node 224, the first protein covalent bond chemical attribute 221a of the first protein edge 221, and a first protein covalent parameter 152 indicative of a binding stability of the first protein edge 221. The first protein edge parameter 151 may be defined according to Equation (4).

[0088] In various embodiments, the first protein covalent parameter 152 may be based on a predetermined protein vector representative of a distance property of the intramolecular covalent bond of the first protein edge 221, as a function of the first protein covalent bond physical attribute 221b. The first protein covalent bond physical attribute 221b may comprise, the protein distance attribute that is representative of the physical distance value between the first protein node 222 and the second protein node 224.

[0089] In various embodiments, the predetermined protein vector may comprise the predetermined RBF $\lambda^{(l)}(d)$, parameterized by a second power value indicative of the magnitude of the distance property of the intramolecular covalent bond of the first protein edge 221. The predetermined RBF $\lambda^{(l)}(d)$ may be expressed according to Equations (2) and (3), and the second power value may comprise a value equal to 1, i.e. $l = 1$; and the first protein covalent parameter 152 may be represented as $\lambda^{(1)}(\|r_j - r_i\|)$, where $r_j - r_i$ refers to the

protein distance attribute representative of the physical distance value between the first 222 and second 224 protein nodes.

[0090] The processor 110 may be further configured to: determine, for the second protein edge 223, a second protein edge parameter 153 indicative of intramolecular interaction data of the second protein node 224 and the third protein node 226 positioned adjacent to the second protein node 224, and bound to the second protein node 224 by the second protein edge 223. The second protein edge parameter 153 may be determined based on, the second protein attribute 224a of the second protein node 224, the third protein attribute 226a of the third protein node 226, the second protein covalent bond chemical attribute 223a of the second protein edge 223, and a second protein covalent parameter 154 indicative of a binding stability of the second protein edge 223. The second protein covalent parameter 154 and the second protein edge parameter 153 may be determined in the same manner as the first protein covalent parameter 152 and the first protein edge parameter 151, respectively, as described above.

[0091] The processor 110 may be further configured to: determine, for the third protein edge 225, a third protein edge parameter 155 indicative of intramolecular interaction data of the third protein node 226 and the fourth protein node 228 positioned adjacent to the third protein node 226, and bound to the third protein node 226 by the third protein edge 225. The third protein edge parameter 155 may be determined based on, the third protein attribute 226a of the third protein node 226, the fourth protein attribute 228a of the fourth protein node 228, the third protein covalent bond chemical attribute 225a of the third protein edge 225, and a third protein covalent parameter 156 indicative of a binding stability of the third protein edge 225. The third protein covalent parameter 156 and the third protein edge parameter 155 may be determined in the same manner as the first and second protein covalent parameters 152, 154 and the first protein edge parameters 151, 153, respectively, as described above.

[0092] The processor 110 may be further configured to: determine, for the fourth protein edge 227, a fourth protein edge parameter 157 indicative of intramolecular interaction data of the fourth protein node 228 and the fifth protein node 232 positioned adjacent to the fourth protein node 228, and bound to the fourth protein node 228 by the fourth protein edge 227. The fourth protein edge parameter 157 may be determined based on, the fourth protein attribute 228a of the fourth protein node 228, the fifth protein attribute 232a of the fifth protein node 232, the fourth protein covalent bond chemical attribute 227a of the fourth protein edge 227, and a fourth protein covalent parameter 158 indicative of a binding stability of the fourth protein edge 227. The fourth protein covalent parameter 158 and the fourth protein edge

parameter 157 may be determined in the same manner as the first, second and third protein covalent parameters 152, 154, 156 and the first, second and third protein edge parameters 151, 153, 155 respectively, as described above.

[0093] The processor 110 may be further configured to: determine, for the fifth protein edge 229, a fifth protein edge parameter 159 indicative of intramolecular interaction data of the fifth protein node 232 and the sixth protein node 234, positioned adjacent to the fifth protein node 232, and bound to the fifth protein node 232 by the fifth protein edge 229. The fifth protein edge parameter 159 may be determined based on, the fifth protein attribute 232a of the fifth protein node 232, the sixth protein attribute 234a of the sixth protein node 234, the fifth protein covalent bond chemical attribute 229a of the fifth protein edge 229, and a fifth protein covalent parameter 160 indicative of a binding stability of the fifth protein edge 229. The fifth protein covalent parameter 160 and the fifth protein edge parameter 159 may be determined in the same manner as the first to fourth protein covalent parameters 152, 154, 156, 158 and the first to fourth protein edge parameters 151, 153, 155, 157 respectively, as described above.

[0094] To determine the overall protein edge parameter 134, each of the first to fifth protein edge parameters 151, 153, 155, 157, 159 may be aggregated. In other words, the overall ligand edge parameter 132 may be based on the sum of each protein edge parameter indicative of each protein covalent bond, e.g. first to fifth protein edges 221, 223, 225, 227, 229, of the protein 220. The overall protein edge parameter 134 may be expressed in accordance with Equation (5) as defined above.

[0095] Non-covalent atoms, although not sharing electrons, exert forces on nearby atoms due to the charges carried by the electrons and the resulting electric field. These forces correspond to different types of dispersion forces specific to the non-covalent bond, which include the Coulomb type dispersion force, and the London type dispersion force. Accordingly, to account for these forces, and with reference to FIGS. 1 and 4, the second block 420 for the non-covalent interactions of the ligand-protein complex 250 may include an overall non-covalent edge parameter 136.

[0096] To determine the overall non-covalent edge parameter 136, the processor 110 may be configured to determine, for the first non-covalent edge 252, a first non-covalent parameter 162 indicative of intermolecular interaction data of the first ligand node 212, and the first protein node 222 positioned adjacent to the first ligand node 212, and bound to the first ligand node 212 by the first non-covalent edge 252, based on the first ligand attribute 212a of the first ligand node 212, the first protein attribute 222a of the first protein node 222, and a first ligand-

protein non-covalent parameter 164 indicative of a binding stability of a type of dispersion force of the first non-covalent edge 252.

[0097] In various embodiments, the first ligand-protein non-covalent parameter 164 may be based on a predetermined ligand-protein vector representative of a magnitude of a force field corresponding to the type of dispersion force of the first non-covalent edge 252, as a function of the first non-covalent bond attribute 252b of the first non-covalent edge 252. The first non-covalent bond attribute 252b may comprise, a ligand-protein distance attribute representative of a physical distance value between the first ligand node 212 and the first protein node 222, and in some embodiments, the ligand-protein distance attribute may comprise a distance of less than 5Å. The physical distance value may therefore be indicative of a force field interaction between the first ligand node 212 and the first protein node 222.

[0098] In various embodiments, the predetermined ligand-protein vector may comprise a Coulomb ligand-protein vector comprising the predetermined RBF $\lambda^{(l)}(d)$, parameterized by a Coulomb power value indicative of the magnitude of the force field of the Coulomb type intermolecular dispersion force of the first non-covalent edge 252. The predetermined RBF $\lambda^{(l)}(d)$ may be expressed according to Equations (2) and (3), and the Coulomb power value may comprise a value equal to -2, i.e. $l = -2$. The first ligand-protein non-covalent parameter 164 for a non-covalent bond having a Coulomb type dispersion force may therefore be represented as $\lambda^{(-2)}(\|r_k - r_i\|)$, where $r_k - r_i$ refers to the ligand-protein distance attribute representative of the physical distance value between the first ligand node 212 and the first protein node 222, and may be indicative of the force field interaction between the first ligand node 212 and the first protein node 222.

[0099] In various other embodiments, the predetermined ligand-protein vector may comprise a London ligand-protein vector comprising the predetermined RBF $\lambda^{(l)}(d)$, parameterized by a London power value indicative of the magnitude of the force field of the London type intermolecular dispersion force of the first non-covalent edge 252. The predetermined RBF $\lambda^{(l)}(d)$ may be expressed according to Equations (2) and (3), and the London power value may comprise a value equal to -6, i.e. $l = -6$. The first ligand-protein non-covalent parameter 164 for a non-covalent bond having a London type dispersion force may therefore be represented as $\lambda^{(-6)}(\|r_k - r_i\|)$, where $r_k - r_i$ refers to the ligand-protein distance attribute representative of the physical distance value between the first ligand node

212 and the first protein node 222, and may be indicative of the force field interaction between the first ligand node 212 and the first protein node 222.

[00100] Accordingly, the first non-covalent parameter 162 may be defined according to Equation (6),

$$m_{ki}^{(ncov)} = h_k^{(t)} \odot W^{(ncov)}\left[\lambda^{(-2)}(\|r_k - r_i\|), \lambda^{(-6)}(\|r_k - r_i\|)\right], \forall\, e_{ki} \in \mathcal{E}_{lp} \quad \text{Equation (6),}$$

where the intermolecular interaction data, i.e. $m_{ki}^{(ncov)}$ between a non-covalent ligand atom v_i and a nearby non-covalent protein atom v_k within a distance of d_{max}, e.g. 5Å as the element-wise multiplication ($\odot$) of the hidden layer features of the neighboring atoms, denoted as $h_k^{(t)}$, with the concatenated up sampled distances to the power value of -2 (Coulomb type dispersion force) and power value of -6 (London type dispersion force), which may be denoted as $W^{(ncov)}[.,.]$.

[00101] The processor 110 may be further configured to determine, for the second non-covalent edge 254, a second non-covalent parameter 166 indicative of intermolecular interaction data of the first ligand node 212, and the sixth protein node 234 positioned adjacent to the first ligand node 212, and bound to the first ligand node 212 by the second non-covalent edge 254, based on the first ligand attribute 212a of the first ligand node 212, the sixth protein attribute 234a of the sixth protein node 234, and a second ligand-protein non-covalent parameter 168 indicative of a binding stability of a type of dispersion force of the second non-covalent edge 254. The second ligand-protein non-covalent parameter 168 and the second non-covalent parameter 166 may be determined in the same manner as the first ligand-protein non-covalent parameter 164 and the first non-covalent parameter 162, respectively, as described above. For example, the second ligand-protein non-covalent parameter 168 may account for the type of dispersion force, e.g. Coulomb type dispersion force, or London type dispersion force, of the second non-covalent edge 254, based on the predetermined RBF $\lambda^{(l)}(d)$ being parametrized by the Coulomb power value, i.e. $l = -2$, or the London power value, i.e. $l = -6$.

[00102] Accordingly, the processor 110 may determine, the overall non-covalent edge parameter 136 based on an aggregation of the non-covalent parameters, for the plurality of non-covalent edges of the ligand-protein complex 250. For example, the overall non-covalent edge parameter 136 may be based on the sum of the first non-covalent parameter 162 and the second non-covalent parameter 166. The overall non-covalent edge parameter 136 may be indicative

of the intermolecular interaction data non-covalent bonds which bind the ligand atom to a respective protein atom, and may be represented according to Equation (7),

$$m_i^{(ncov)} = \sum_{v_k \in N(v_i)} m_{ki}^{(ncov)}, \forall\, e_{ki} \in\ \mathcal{E}_p \qquad \text{Equation (7).}$$

[00103] In various embodiments, the processor 110 may further comprise the second machine learning algorithm, e.g. the MLP (see also Equation (1)), for determining the ligand-protein complex interaction parameter 130. The second machine learning algorithm may be used to update the ligand atom and protein atoms hidden features, based on a combination of the intramolecular interaction data of the first and second ligand edges 215, 217; the first to fifth protein edges 221, 223, 225, 227, 229 (see Equation (5)), and the intermolecular interaction of the first and second non-covalent edges 252, 254 (see Equation (7)) parts along with the atom hidden features, and may be expressed according to Equation (8),

$$h_i^{t+1} = MLP^{(cov)}\left(h_i^{(t)} + m_i^{(cov)}\right) + MLP^{(ncov)}\left(h_i^{(t)} + m_i^{(ncov)}\right), t = 1,2,\dots,T \qquad \text{Equation (8)}$$

[00104] Accordingly, the representation $\boldsymbol{f}$ of the ligand-protein complex 250 in a docking pocket (including at least one binding site) of the plurality of docking pockets, may be obtained by summing up the hidden features of all atoms after undergoing updates for *T* layers, i.e., $\boldsymbol{f} = \sum_i h_i^{(T)}$, and this representation $\boldsymbol{f}$, may be utilized in the third machine learning algorithm for self-attention learning, and the first machine learning algorithm for semi-supervised learning. In other words, referring to Equation (1), $\sum_V \left[MP_T^{(cov)}(V, \mathcal{E}, \mathcal{R}) + MP_T^{(ncov)}(V, \mathcal{E}, \mathcal{R})\right]$ may be represented as $\boldsymbol{f}$. As such, the objective function for training the bioactivity parameter 180 determination prediction capability of the DTIGN may be expressed according to Equation (9),

$$L_{bio} = \frac{1}{N_s}\sum_{s=1}^{N_s}(MLP^{(bio)}(f_s) - y_s)^2 \qquad \text{Equation (9),}$$

where s and N_s stands for the index and total number of training samples, respectively.

<u>Aggregated ligand-protein complex interaction parameter 170</u>

[00105] To determine the ligand-protein complex interaction parameter 130 discussed above, candidate pockets comprising at least one verified binding sites of the protein was selected to gain insights into the ligand-protein complex 250 interactions. While these candidate pockets represent the possible area of ligand-protein complex 250 binding, many ligands 210 may interact with target proteins 220 without clear identification of their binding pockets due to the lack of native structures.

[00106] To address these issues, and with reference to FIG. 1, the processor 110 may comprise the first machine learning algorithm, e.g. a deep learning algorithm, such as, but not limited to a graph neural network for the identification of native-like binding pockets and docking poses from the pocket-ligand complexes docked onto candidate pockets of the target protein 220. In various embodiments, the processor 110 may further comprise a third machine learning algorithm which may comprise a multi-head self-attention algorithm (also known as the one or more self-attention heads mechanism), for determining an aggregated ligand-protein complex interaction parameter 170.

[00107] During model training, the docking poses data of the same ligand 210 onto multiple candidate pockets may be packaged to facilitate training of the DTIGN using the third machine learning algorithm. In various embodiments, the self-attention layer of the third machine learning algorithm, combined with a residual module, may consolidate pocket-ligand representations, for determining the aggregated ligand-protein complex interaction parameter 170 indicative of an aggregated intermolecular interaction data of the ligand-protein complex 250, in the plurality of docking poses, thereby capturing the involvement of ligands 210 in relevant target protein 220 interactions. In some embodiments, the residual module may comprise, a dynamic residual module for modelling dynamic residual attention, or a grouped residual layer to replace the query, key value linear layer in the self-attention algorithm. In other words, the residual module may be used to add the pocket-ligand representation to the output of the multi-head self-attention algorithm, before they are aggregated together (see Equations (10) and (15) below).

[00108] To determine the aggregated ligand-protein complex interaction parameter 170, the processor 110 may be configured to: determine, for each docking pose of the plurality of docking poses, a ligand-protein complex representation parameter 172 indicative of intermolecular interaction data of an updated ligand-protein complex, the updated ligand-protein complex formed when the ligand is bound to at least one identified binding site comprised in at least one identified docking pocket, of the protein to form the updated ligand-protein complex, in a respective docking pose, wherein determining the ligand-protein complex representation parameter 172 comprises, processing, the ligand-protein complex interaction parameter 130 with the third machine learning algorithm; and determine, the aggregated ligand-protein complex interaction parameter 170 based on an aggregation of the ligand-protein complex representation parameter 172, for each docking pose of the updated ligand-protein complex. This may be expressed in accordance with Equations (10) to (16) below.

[00109] In the present disclosure, the graph $\mathcal{G}_n$ may comprise the same ligand 210 which may be docked into N_p candidate pockets on the target protein with poses $n = 1, 2, \dots, N_p$. For each pocket-ligand pose *n*, the corresponding graph feature $\boldsymbol{f}_n$ using the DTIGN as described in the above may be obtained. Then, the queries q_n, keys k_n, and values v_n for each pose in each attention head of the third machine learning algorithm, are obtained from different single-layer neural networks based on $\boldsymbol{f}_n$, as expressed by Equation (10),

$$q_n = W^Q f_n,\ k_n = W^K f_n,\ v_n = W^V f_n, n = 1,2, \dots, N_p \quad \text{Equation (10),}$$

where $W^Q,\ W^K,\ W^v$ refers to weight matrices which reduce the input feature dimensions to enable efficient self-attention in each attention head.

[00110] Subsequently, for each pose *n*, the query q_n and the keys k_n for all poses, $m = 1, 2, \dots, N_p$, may be used to calculate the self-attention weights of the third machine learning algorithm, and may be expressed by Equation (11),

$$\alpha_{n,m} = \frac{q_n^T \cdot k_m}{\sqrt{d_k}}, \forall\, m, n = 1,2, \dots, N_p \quad \text{Equation (11),}$$

where d_k refers to the dimension of the keys.

[00111] Then, a normalized exponential function, such as a SoftMax function, may be used to normalized the self-attention weights $\alpha_{n,m}$ of the third machine learning algorithm, between poses *n* and *m*, which may be expressed according to Equation (12),

$$\hat{\alpha}_{n,m} = \frac{\exp(\alpha_{n,m})}{\sum_m exp(\alpha_{n,m})}, \forall\, m, n = 1,2, \dots, N_p \quad \text{Equation (12).}$$

[00112] The ligand-protein complex representation parameter 172 may comprise the output representation of each attention head of the third machine learning algorithm, which may be obtained by weighting all the values $\boldsymbol{v}_m$ for poses $m = 1, 2, \dots, N_p$ according to the normalized self-attention weights $\hat{\alpha}_{n,m}$ and summing then, in accordance with Equation (13),

$$\bar{f}_n^{(head)} = \sum_{m=}^{M} \hat{\alpha}_{n,m} \cdot \boldsymbol{v}_m, \forall\, n = 1,2, \dots, N_p \quad \text{Equation (13).}$$

[00113] The representations from different attention heads of the third machine learning algorithm may then be concatenated (denoted as $\oplus_{head}$) to obtain the output representation of the third machine learning algorithm, e.g. multi-head self-attention mechanism, as expressed by Equation (14),

$$\bar{f}_n = \oplus_{head}\, \bar{f}_n^{(head)} \quad \text{Equation (14).}$$

[00114] The representations of all pocket-ligand complexes *n* for the same ligand 210 may then be added with corresponding graph feature $\boldsymbol{f}_n$ and summed to obtain the aggregated

ligand-protein complex interaction parameter 170, i.e. multi-pose aggregated representation $\bar{\boldsymbol{f}}$, in accordance with Equation (15),

$$\bar{f} = \sum_n (\bar{f_n} + f_n) \qquad \text{Equation (15).}$$

[00115] In various embodiments, the multi-pose aggregated representation $\bar{\boldsymbol{f}}$ may replace the single-pose representation $\boldsymbol{f}$ in Equation (9) for determining the bioactivity parameter 180, and the bioactivity loss may be optimized in accordance with Equation (16),

$$\mathcal{L}_{bio}^{(attn)} = \frac{1}{N_s} \sum_{s=1}^{N_s} [MLP^{(bio)}(\bar{f_s}) - y_s]^2 \qquad \text{Equation (16),}$$

where the superscript (attn) may be used to distinguish the loss function $\mathcal{L}_{bio}^{(attn)}$ obtained from features $\boldsymbol{f}$ aggregated by the third machine learning algorithm, i.e. multi-head self-attention mechanism.

[00116] Accordingly, in various embodiments, the bioactivity parameter 180 may be determined based on the aggregated ligand-protein complex interaction parameter 170, i.e. multi-pose aggregated representation $\bar{\boldsymbol{f}}$.

Semi-supervised learning

[00117] Within the collected dataset of the reference bioactivity data for the ligand 102, the protein 104, and the ligand-protein complex 106, and the ligand-protein complex structural data 200, there are typically only a small fraction of bioactive ligands molecules that may have publicly available native structures data (typically obtained through X-ray or Nuclear Magnetic Resonance) with the target protein molecules.

[00118] Leveraging the native structures data of the ligand 210, the protein 220 and/or the ligand-protein complex 250, the processor 110 may comprise the first machine learning algorithm, i.e. the deep learning algorithm, such as, but not limited to a graph neural network algorithm, that may provide a semi-supervised learning strategy to guide the learning of the third machine learning algorithm, e.g. self-attention mechanism. In various embodiments, the first machine algorithm may be the same machine algorithm used for determining the ligand-protein complex parameter 130 and/or for the identification of native-like binding pockets and docking poses from the pocket-ligand complexes docked onto candidate pockets of the target protein 220. The native structures data of the ligand 210, the protein 220 and/or the ligand-protein complex 250 may be stored in the memory 120, or may be stored in another processor or server.

[00119] In various embodiments, based on the native structures data of the ligand 210, the protein 220 and/or the ligand-protein complex 250, protein atoms within residues may be extracted from the protein 220, to form at least one local pocket-ligand complex where at least one protein atom lies within 5Å of any ligand atoms. Data on the at least one local pocket-ligand complexes may be input into the DTIGN to determine a native ligand-protein complex interaction parameter 182 and a native aggregated ligand-protein complex interaction parameter 184, in a similar manner for determining, the ligand-protein complex interaction parameter 130, and for determining, the aggregated ligand-protein complex interaction parameter 170, respectively, details of which have been omitted for conciseness. This may be represented as native representations $\boldsymbol{f}_{native}$.

[00120] In various embodiments, for ligand-protein pairs with native structures, the processor 110 may be further configured to, determine, a similarity, for example, a cosine similarity between the aggregated ligand-protein complex interaction parameter 170, i.e. multi-pose aggregated representation $\bar{\boldsymbol{f}}$, and the native aggregated ligand-protein complex interaction parameter 184, i.e. native representation $\boldsymbol{f}_{native}$ to ensure feature alignment. The loss function may therefore be defined according to Equation (17),

$$\mathcal{L}_{sim} = \frac{1}{N_{native}} \sum_{f_{native}} \left(1 - \frac{f_{native} \cdot \bar{f}}{\|f_{native}\|_2 \|\bar{f}\|_2}\right) \quad \text{Equation (17),}$$

where N_{native} refers to the count of ligand-protein pairs with native structure data in the reference bioactivity data for the ligand 102, the reference bioactivity data protein 104, and the reference bioactivity data ligand-protein complex 106.

[00121] In various embodiments, the native aggregated ligand-protein complex interaction parameter 184, i.e. native representation $\boldsymbol{f}_{native}$, may further be input into the DTIGN, and processed by the second machine learning algorithm, e.g. MLP, for optimizing the determination of the bioactivity parameter 180, to ensure alignment with the actual bioactivity y_{native} of the ligand 210. In this respect, the total loss function for training the DTIGN may be defined according to Equation (18),

$$\mathcal{L}_{total} = \mathcal{L}_{bio}^{(attn)} + \mathcal{L}_{bio}^{(native)} + \gamma \mathcal{L}_{sim} = \frac{1}{N_s} \sum_{s=1}^{N_s} \left\{ \left(MLP^{(bio)}\left(\bar{f}_s\right) - y_s\right)^2 + \mathbb{I}[\exists f_{native} \text{ for } s] \cdot \frac{1}{N_{native}} \sum_{s=1}^{N_{native}} [\left(MLP^{(bio)}(f_{native}) - y_{native}\right)^2 + \gamma\left(1 - \frac{f_{native} \cdot \bar{f}}{\|f_{native}\|_2 \|\bar{f}\|_2}\right)] \right\} \quad \text{Equation (18)}$$

where $\mathbb{I}[\exists f_{native} \text{ for } s]$ equals 1 if the training ligand s possesses at least one native structure with the target protein 220; otherwise, it equals 0. The coefficient γ balances the contributions

of bioactivity parameter 180 prediction losses and the cosine similarity loss. The notation N and N_{native} represent the number of training ligands s and their native structures with target proteins 220, respectively. Notably, when a ligand 210 interacts with a target protein 220 through multiple native structures, i.e., $N_{native} > 1$, their representations may be summed to obtain the native aggregated ligand-protein complex interaction parameter 184, i.e. native representation $\boldsymbol{f}_{native}$.

[00122] According to another aspect of the disclosure, there is provided a drug discovery platform comprising system 100 described with reference to FIGS. 1 to 4.

[00123] FIG. 5 shows an exemplary flowchart of a method 500 for determining, a bioactivity parameter for a ligand, when the ligand is bound to a protein to form a ligand-protein complex, the ligand being bound to the protein in a plurality of docking poses. The bioactivity parameter, the ligand, the protein and the ligand-protein complex may refer to bioactivity parameter 180, ligand 210, protein 220 and the ligand-protein complex 250 of system 100 and discussed with reference to FIGS. 1 to 4.

[00124] Method 500 may comprise providing a processor for executing the steps of: obtaining, a ligand-protein complex structural data indicative of a structure of the ligand-protein complex and a reference bioactivity data of the ligand, the protein and the ligand-protein complex; the ligand-protein complex structural data comprising, a plurality of ligand nodes of the ligand, each ligand node having a ligand attribute; a plurality of protein nodes, each protein node having a protein attribute; a plurality of non-covalent edges connecting a ligand node to a respective protein node, each non-covalent edge having a non-covalent bond attribute (step 502); determining, for each docking pose of the plurality of docking poses, a ligand-protein complex interaction parameter indicative of intermolecular interaction data of the ligand-protein complex, in a respective docking pose, wherein determining the ligand-protein complex interaction parameter comprises, determining, for each non-covalent edge, a non-covalent parameter indicative of intermolecular interaction data of a first ligand node, and a first protein node positioned adjacent thereto and bound to the first ligand node by a first non-covalent edge, based on, a first ligand attribute of the first ligand node, a first protein attribute of the first protein node, and a first ligand-protein non-covalent parameter indicative of a binding stability of a type of dispersion force of the first non-covalent edge; and determining, the ligand-protein complex interaction parameter based on an aggregation of the non-covalent parameters, for the plurality of non-covalent edges, of the ligand-protein complex (step 504);

determining, an aggregated ligand-protein complex interaction parameter indicative of an aggregated intermolecular interaction data of the ligand-protein complex, in the plurality of docking poses (step 506); and determining, the bioactivity parameter based on the aggregated ligand-protein complex interaction parameter (step 508).

[00125] In various embodiments, the first ligand-protein non-covalent parameter may be based on a predetermined ligand-protein vector representative of a magnitude of a force field corresponding to the type of dispersion force of the first non-covalent edge, as a function of the non-covalent bond attribute of the first non-covalent edge.

[00126] In various embodiments, the predetermined ligand-protein vector may comprise a Coulomb ligand-protein vector comprising a predetermined RBF parametrized by a Coulomb power value indicative of the magnitude of the force field of a Coulomb dispersion force of the first non-covalent edge. In various other embodiments, the predetermined ligand-protein vector may comprise, a London ligand-protein vector comprising a predetermined RBF parametrized by a predetermined London power value indicative of the magnitude of the force field a London dispersion force of the first non-covalent edge.

[00127] In various embodiments, the non-covalent bond attribute of the first non-covalent edge may comprise, a ligand-protein distance attribute representative of a physical distance value between the first ligand node and the first protein node, of the ligand-protein complex. The physical distance value may be indicative of a force field interaction between the first ligand node and the first protein node. In some embodiments, to determine the ligand-protein distance attribute, method 500 may further comprise: measuring, on the ligand-protein complex structural data, the physical distance value between the first ligand node and the first protein node; comparing, the physical distance value with a predetermined distance threshold value indicative of a presence of the first non-covalent edge; and determining, the presence of the first non-covalent edge based on the comparison of the physical distance value and the predetermined distance threshold value.

[00128] In various embodiments, the ligand-protein complex structural data may further comprise, a plurality of ligand edges binding the plurality of ligand nodes to each other, in the ligand, each ligand edge having a ligand covalent bond chemical attribute, and a ligand covalent bond physical attribute, wherein the method 500 may further comprise: determining, for each ligand edge of the plurality of ligand edges, a ligand edge parameter indicative of intramolecular interaction data of the first ligand node and a second ligand node positioned adjacent thereto, and bound to the first ligand node by a first ligand edge, the ligand edge

parameter determined based on, the first ligand attribute of the first ligand node, a second ligand attribute of the second ligand node, the ligand covalent bond chemical attribute of the first ligand edge, and a first ligand covalent parameter indicative of a binding stability of the first ligand edge, and determining, the ligand-protein complex interaction parameter further based on, an aggregation of the ligand edge parameters, for the plurality of ligand edges.

[00129] In various embodiments, the first ligand covalent parameter may be based on a predetermined ligand vector representative of a distance property of the intramolecular covalent bond of the first ligand edge, as a function of the ligand covalent bond physical attribute of the first ligand edge

[00130] In various embodiments, the predetermined ligand vector may comprise a predetermined RBF parametrized by a first power value indicative of the distance property of the intramolecular covalent bond, of the first ligand edge.

[00131] In various embodiments, the ligand covalent bond physical attribute of the first ligand edge may comprise, a ligand distance attribute representative of a physical distance value between the first ligand node and the second ligand node, of the ligand.

[00132] In various embodiments, the ligand-protein complex structural data may further comprise, a plurality of protein edges binding the plurality of protein nodes to each other, in the protein, each protein edge having a protein covalent bond chemical attribute, and a protein covalent bond physical attribute, wherein the method 500 may further comprise: determining, for each protein edge of the plurality of protein edges, a protein edge parameter indicative of intramolecular interaction data of the first protein node and a second protein node positioned adjacent thereto, and bound to the first protein node by a first protein edge of the plurality of protein edges, the protein edge parameter determined based on, the first protein attribute of the first protein node, a second protein attribute of the second protein node, the protein covalent bond chemical attribute of the first protein edge, and a first protein covalent parameter indicative of a binding stability of the first protein edge, and determining, the ligand-protein complex interaction parameter further based on, an aggregation of the protein edge parameters, for the plurality of protein edges.

[00133] In various embodiments, the first protein covalent parameter may be based on a predetermined protein vector representative of a distance property of the intramolecular covalent bond of the first protein edge, as a function of the protein covalent bond physical attribute of the first protein edge.

[00134] In various embodiments, the predetermined protein vector may comprise a predetermined RBF parametrized by a second power value indicative of the distance property of the intramolecular covalent bond, of the first protein edge.

[00135] In various embodiments, the protein covalent bond physical attribute of the first protein edge may comprise, a protein distance attribute representative of a physical distance value between the first protein node and the second protein node, of the protein.

[00136] In various embodiments, method 500 may further comprise, a first machine learning algorithm, the first machine learning algorithm optionally including a deep learning algorithm such as, but not limited to, a neural network-based deep learning algorithm.

[00137] In various embodiments, method 500 may further comprise, a second machine learning algorithm, for determining the ligand-protein complex interaction parameter, wherein the second machine learning algorithm optionally comprises, an MLP.

[00138] In various embodiments, method 500 may further comprise, a third machine learning algorithm, for determining the aggregated ligand-protein complex interaction parameter, wherein the third machine learning algorithm optionally comprises, a multi-head self-attention algorithm.

[00139] In various embodiments, the bioactivity parameter may comprise, at least one of: an EC50 parameter indicative of a half maximal effective concentration of the ligand; and/or an IC50 parameter indicative of a half maximal inhibitory concentration of the ligand.

[00140] According to yet another aspect of the disclosure, there is provided a computer readable medium comprising instructions, which when executed by the processor, causes the processor to perform method 500.

[00141] While the determination of the ligand-protein complex interaction parameter 130, the aggregated ligand-protein complex interaction parameter 170 and bioactivity parameter 180 have been discussed with respect to the exemplary ligand-protein complex 250 shown in FIG. 2, embodiments of the disclosure are not limited thereto, and the system 100 and method 500 of the present disclosure may be applicable to other ligand-protein, e.g. drug-target complexes having a different number of ligand nodes, ligand edges, protein nodes, protein edges, and/or non-covalent edges.

EXAMPLES

[00142] The system 100 and method 500 for determining a bioactivity parameter 180 for a ligand 210, herein disclosed are further illustrated in the following examples, which are provided by way of illustration and are not intended to be limiting the scope of the present disclosure.

[00143] In the examples below, the term “ligand” may refer to the ligand 210 of the present disclosure; the term “protein” may refer to the protein 220 of the present disclosure; the term “Drug-Target” may refer to the ligand-protein complex 250 of the present disclosure; the term “bioactivity” may refer to the bioactivity parameter 180 of the present disclosure; the term “multi-head self-attention; self-attention mechanism” may refer to the third machine learning algorithm of the present disclosure; and the term “semi-supervised component” may refer to the first machine learning algorithm, e.g. deep learning algorithm, such as, but not limited to the graph neural network of the present disclosure.

[00144] FIG. 6 show an exemplary schematic illustration of the overall DTIGN 600 for bioactivity prediction based on drug-target interactions. With reference to FIG. 6, the contributions of the present disclosure may be summarized as follows:

- The bioactivity prediction in drug discovery is extended beyond traditional single-molecule QSAR models by utilizing molecular interaction information generated through molecular docking methods. Current bioactivity prediction methods largely neglect this information. In the present disclosure, this information was validated through a newly designed network, DTIGN, and a set of comprehensive experiments.
- Based on interatomic interactions, DTIGN is devised. Employing multi-head self-attention within DTIGN facilitates the identification of the true binding pockets and poses within docking results. The added semi-supervised component, e.g. first machine learning algorithm comprising a deep learning algorithm, such as, but not limited to a graph neural network, empowers DTIGN to leverage scarce ligand-protein native structures for refining model attention.
- The experimental results provided here confirm that DTIGN outperforms 9 leading non-interaction-based bioactivity prediction methods with an average improvement of 27.03%. In addition, native-like docking poses receive high attention in DTIGN, indicating that its accurate self-attention to native-like poses is the key to accurate bioactivity predictions.

I. Data Preparation

A. Bioactivity Data Collection and Preprocessing

[00145] **Bioactivity Data Source**: The present disclosure utilized bioactivity data from publicly available datasets, primarily sourced from ChEMBL, one of the major repositories for bioactive molecules with drug-like properties. ChEMBL comprises over 2 million molecules, encompassing 1.5 million assays across 15,000 protein targets, spanning various organisms. The present disclosure focused on Single Proteins (distinct from Protein Compounds or Families) with more than 1000 bioactive small molecules. Out of the 6,778 human proteins in ChEMBL, only 900 met their criteria. The dataset includes diverse bioactivity measurements (e.g., IC50, EC50, Kd, Ki, % inhibition, potency, etc.), and molecular representations (e.g., SMILES and ChEMBL Identifier), totaling 6.75 million records.

[00146] **Bioactivity Data Preprocessing**: The raw data underwent several preprocessing steps. Initially, molecules with "Null" assay values were removed. Assay types with a sample size of less than 6 were excluded due to insufficient data for robust model training, validation, and testing. Some assays with low bioactivity do not have precise values, but are marked as greater than a certain threshold. For these non-critical samples, their thresholds may be taken as their assay values (e.g., set the assay "> 10000" to the assay "10000"). The present disclosure standardized values to the appropriate scale (e.g., from "IC50 = 1 nM" to "pIC50 = $-\log 10^{-9}$ M = 9"). In cases where the same assay type had distinct units that could not be converted, the assay type were separated them into different datasets (e.g., "Activity in percentage" and "Activity in nM"). Duplicate assay values with identical units were averaged for data consistency.

[00147] **Bioactivity Datasets Construction**: Each distinct assay category is treated as an individual dataset for each specific target, as different assay types hold varying biological implications. Within each dataset, data is partitioned into a 5-to-1 ratio, with 5 parts serving as training data and the remaining part as test data. The model is trained on every four parts of the training data and its performance is verified on the remaining part of the training data. The model with the best verification performance is tested on the test set. During this process, molecules are divided based on their Bemis-Murcko scaffolds and grouped with identical scaffolds within the same dataset whenever possible. This approach is employed to amplify the discrepancy among the training, validation, and testing data, thus facilitating the assessment of the models' generalization capabilities.

B. Drug-Target Interaction Data Generation

[00148] Ligands with similar structures can exhibit significantly different bioactivities, which can be distinguished by their interactions with proteins. This fact motivates the inclusion of drug-target interactions in the process of predicting bioactivity in this present disclosure.

[00149] FIG. 7 shows the docking pocket selection 700 for an exemplary ligand, CHEMBL202. FIG. 8 shows a table 800 of comparison of bioactivity (pIC50 towards CHEMBL202) prediction performance on different data sources, AutoDock Vina and QVina-W, where the ↑ arrows or ↓ arrows denote that larger or smaller values are better. FIG. 9 shows a table of the benchmark datasets 900 of bioactivity prediction on ligand-protein complexes.

[00150] **Binding Sites and Docking Pockets**: The present disclosure first used the identifiers in the ChEMBL bioactivity datasets (as constructed in Section I-A) to retrieve the functional binding sites of target proteins from the UniProt database. Next, the present disclosure selected corresponding crystal structures of the protein from the RCSB Protein Data Bank (PDB) that feature high resolution and ligands bound in the majority of binding sites. Subsequently, the present disclosure annotated binding sites on the crystal structure of each protein, partitioning pockets based on the spatial clustering of binding sites. Finally, the present disclosure included a large pocket comprising all identified binding sites. Example pockets selected for the exemplary target protein CHEMBL202 are illustrated in FIG. 7.

[00151] **Docking Method**: The present disclosure compared the prediction performance of data generated from two molecular docking software, AutoDock Vina and QVina-W. Taking CHEMBL202 as an example, the present disclosure docked 957 ligands with known bioactivity onto 7 pockets, resulting in 6,446 molecules. They were divided into training (5,312) and testing (1,134) sets using a 5-fold cross-validation strategy. The geometric interaction graph neural network (GIGN) was the backbone network, optimizing Mean Squared Error (MSE). Despite noise due to multiple pockets, the present disclosure assessed docking methods generating training data. Graph Convolutional Network (GCN) and Universal 3D Molecular Representation Learning Framework (Uni-Mol) were used as controls. Metrics included Pearson correlation (r), Root Mean Square Error (RMSE), and Kendall Tau-b Coefficient τ_B. The performance of models trained on docking results using AutoDock-Vina and QVina-W was compared. The results in FIG. 8 highlighted their superior performance over ligand-based methods, particularly Vina-Docked with lower RMSE but higher r and τ_B. Therefore, AutoDock Vina was selected for ligand-protein complex generation in the present disclosure.

[00152] It is worth noting that, with the development of computing power and the emergence of technologies like cloud computing, molecular docking between large-scale molecules and proteins is no longer a challenge. For instance, on a single NVIDIA GeForce RTX 3060 GPU card, an average of 16,222 molecules can be docked on common proteins per day using a GPU version of AutoDock Vina.

[00153] **Benchmark Datasets:** The present disclosure selected target proteins with abundant pIC50 or pEC50 bioactivity measurements in the CHEMBL bioactivity datasets to create diverse benchmark datasets, as outlined in FIG. 9. These datasets vary across characteristics, involving the count of unique ligand molecules (from 957 to 9,573), binding sites (from 2 to 24), chosen pockets (from 2 to 7), and top *n* poses per pocket (from 1 to 4). The dataset ID includes "E" for pEC50 bioactivity and "I" for pIC50 bioactivity. A higher numerical value in the ID signifies more unique ligand molecules (approximately doubled for each subsequent number).

II. Experiments

A. Experimental preparation

[00154] **Data Introduction**: The bioactivity data used to evaluate the proposed method was sourced from ChEMBL. Details on data collection and preprocessing, are outlined in Section I above. The data of pocket-ligand complex comes from the docking results of 3D ligand molecule conformations (generated using RDKit based on molecular SMILES strings) with the target protein X-ray structure (obtained from UniProt) using AutoDock Vina. The selection of docking pockets was based on binding sites recorded on UniProt (see Section I-B above and FIG. 7 for more details on the pocket selection).

[00155] **Comparison Methods**: The prediction of ligand bioactivity based on deep learning primarily employs graph neural networks (GNNs) on ligand structures. The present disclosure consequently compares two major categories of methods: 2D and 3D GNNs built on ligand structures. The 2D GNNs include GCNs, graph attention networks (GATs), graph isomorphism networks (GINs) pre-trained using supervised learning and context prediction, message passing neural networks (MPNNs), Weave, neural fingerprint (Neural FP), and Attentive FP. As for the 3D GNNs, the present disclosure considered the typical spatial graph convolutional networks (SGCN) and the state-of-the-art Uni-Mol pre-trained model. As an additional baseline, the present disclosure also employed a simple MLP to predict bioactivity based on the AutoDock-Vina scores of the ligand-protein poses involved.

[00156] **Evaluation Metrics**: To comprehensively evaluate the model performance, the present disclosure employs two commonly used metrics for bioactivity prediction: the Pearson correlation coefficient (r) and RMSE, along with a ranking metric, the τ_B.

[00157] The r metric assesses the linear correlation between two variables and may be defined according to Equation (19) as follows:

$$r = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{y})}{\sqrt{\sum i=1^{n}(y_i-\bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}} \quad \text{Equation (19),}$$

where, y_i and $\hat{y}_i$ represent the true and predicted bioactivity values, respectively, and $\bar{y}$ and $\bar{\hat{y}_i}$ are the means of the true and predicted bioactivity values, respectively. n stands for the total number of test samples. A larger r value indicates higher overall accuracy in the model predictions.

[00158] RMSE, a widely adopted metric for assessing regression model performance, is defined as the square root of the average of squared differences between predicted and actual values according to Equation (20):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \text{Equation (20),}$$

where y_i and $\hat{y}_i$ represent the true and predicted bioactivity values, respectively, and n denotes the total number of test samples. Lower RMSE values indicate better performance of the regression model.

[00159] τ_B is a non-parametric statistic utilized to measure the agreement between two rankings, commonly used in fields like statistics and computational biology. It is calculated by counting concordant and discordant pairs in the rankings according to Equation (21):

$$\tau_B = \frac{C-D}{C+D} \quad \text{Equation (21),}$$

where C refers to the number of concordant pairs, and D represents the number of discordant pairs. A concordant pair implies items ranked similarly in both rankings, while a discordant pair refers to differing rankings. τ_B values range from -1 to 1, where 1 signifies complete agreement, 0 signifies no agreement, and -1 signifies complete disagreement between the rankings.

[00160] As for the 3D GNNs, they include the SGCN and the Uni-Mol. The source codes and usage instructions for these two 3D GNNs can be accessed on GitHub. Notably, GINs and Uni-Mol offer pre-trained models on large-scale graph data, enabling the present disclosure to fine-tune these models for enhanced performance.

[00161] **Implementation Details:** The training data is divided into 5 subsets to perform 5-fold cross-validation, where the RMSE is employed to assess the model generalization on the validation set. Training is stopped when the validation RMSE does not decrease over 100 epochs, and the model with the lowest validation RMSE will be evaluated for performance on the test set. The model selected with the lowest validation RMSE among the stopping points of 5 folds is considered the final result. The validation RMSE during the initial 40 epochs of training is not considered for the stopping criterion, as it is considered a warm-up phase for the model. The balance coefficient is set to be 1 in all experiments.

[00162] During batch training, all poses of a single ligand docking with protein pockets are treated as a single entity to train the self-attention mechanism fully. Due to varying numbers of pockets and docking poses for different proteins, the present disclosure set a uniform batch size of 128, which is divided by the number of pockets and poses to determine the number of unique ligands sampled per batch. For optimization, the present disclosure employed the Adam optimizer with an initial learning rate of 10^{-4} and a weight decay of $10^{-4.3}$. The learning rate is decayed by 95% every 10 training epochs. The graph convolutional module has 3 layers (i.e., $T = 3$). The hidden feature dimension of the model is consistently set to 256. The multi-head self-attention layer employs 8 heads, with a dropout rate 0. The MLP consists of three linear layers with LeakyReLU activation and BatchNorm1d.

[00163] Data containing both bioactivity records and native ligand-protein complexes are highly limited. Consequently, apart from the ablation study, these rare data instances are exclusively reserved for validating the accuracy of the self-attention mechanism, without further utilization for model training.

[00164] **Experimental Settings**: To ensure a fair comparison, consistency with the recommended model settings of the baseline models was maintained as closely as possible while configuring DTIGN.

B. Ablation Study

[00165] FIG. 10 shows a table indicative of the ablation study 1000 on the I1 task (Dihydrofolate Reductase) in the benchmark dataset.

[00166] To validate the efficacy of the self-attention mechanism, and semi-supervised learning, ablation experiments were conducted on the I1 task (protein name: dihydrofolate reductase) in the benchmark dataset (as detailed in FIG. 9). FIG. 10 shows the comparison of the present method with or without edge features (denoted as “w/o edge feature”) and a multi-

head self-attention mechanism (denoted as "w/o self-attention"). Default training utilizes the MSE loss for bioactivity prediction, which is $\mathcal{L}_{bio}$ in Equation (9). However, with the introduction of the multi-head self-attention mechanism, the present disclosure switch to using $\mathcal{L}_{bio}^{(attn)}$ in Equation (18) for optimizing DTIGN based on attention-aggregated features. If semi-supervised learning is enabled, the complete loss as defined in Equation (18) may be employed.

[00167] **DTIGN Outperforms Leading GNN Methods in Predicting the Molecular Bioactivity against Dihydrofolate Reductase**: The results in FIG. 10 demonstrate that DTIGN and its variants, which utilize pocket-ligand complex data onto the dihydrofolate reductase protein, generally outperform advanced methods that solely rely on ligand data. Specifically, compared with the best-performing methods on each metric, DTIGN exhibits improvements of 44.83% in bioactivity prediction correlation (measured by the Pearson correlation coefficient: r), 11.02% in accuracy (measured by the RMSE), and 59.82% in ranking (measured by the τ_B).

[00168] **Ablation Study on the on the Edge Feature and Self-Attention Mechanism of DTIGN:** The first two rows of the DTIGN results in FIG. 10 indicate that incorporating edge features enhances bioactivity value prediction accuracy by 20%, while slightly diminishing ranking consistency by 6.4%. Furthermore, as shown in the second and the third rows of the DTIGN results in FIG. 10, integration of the multi-head self-attention mechanism improves bioactivity prediction correlation and ranking consistency by 16.8% and 17.5%, respectively, with a minor 1.4% precision reduction. Notably, none of these models were trained with native ligand-protein complex structures.

[00169] **Ablation Study on the Semi-Supervised Learning of DTIGN**: As shown in the last two rows of the DTIGN results in FIG. 10, introducing a limited number (accounting for about 3% of the training data) of X-ray ligand-protein complexes for additional semi-supervised learning resulted in improvements on all metrics. Specifically, bioactivity prediction correlation, precision, and ranking consistency each increased by 3.4%, 7.4%, and 8.2%, respectively. In summary, the results highlight the benefits of the design of DTIGN, including its edge feature definition, self-attention module design, and the integration of semi-supervised learning, in achieving accurate bioactivity prediction.

C. Validation of Self-Attention Accuracy

[00170] FIG. 11 shows the attention values 1100 on protein pockets for a test ligand during the training process. The native binding sites depicted above are sourced from the RCSB PDB

database. Their details and corresponding pockets are given in the legend. FIG. 12 shows a table indicative of the intersection over union 1200 between the "Native Pockets" and "Attention Pockets". FIG. 13 shows the box plot 1300 of DTIGN's attention distributions on the ranked docking poses involved in the pIC50 test set (containing 9,975 redocked poses on 1,011 PDB native structures). "Avg RMSD" represents the average RMSD between ligand-protein docking poses and their respective native structures in the PDBBind dataset. FIG. 14 shows the box plot 1400 of DTIGN's attention distributions on the ranked docking poses involved in the pKd test set (containing 14,411 re-docked poses on 1,455 PDB native structures). "Avg RMSD" represents the average RMSD between ligand-protein docking poses and their respective native structures in the PDBBind dataset. FIG. 15 shows the box plot 1500 of DTIGN's attention distributions on the ranked docking poses involved in the pKi test set (containing 6,196 re-docked poses on 626 PDB native structures). "Avg RMSD" represents the average RMSD between ligand-protein docking poses and their respective native structures in the PDBBind dataset.

[00171] **The ligand-protein poses that are docked in the actual binding pockets tend to attract the attention of DTIGN during training**: In this section, the present disclosure demonstrates the binding pocket and pose identification capabilities of the self-attention mechanism through unsupervised training on the dihydrofolate reductase, a protein with a high number of pockets (as shown in FIG. 9). The present disclosure investigated the alignment between self-attention-highlighted pockets and native pockets in FIG. 11. Firstly, the present disclosure queried the RCSB PDB database to identify the known binding sites on the protein (CHEMBL202) and the binding site where the test ligand (CHEMBL7492) binds, as indicated in the annotations above FIG. 11. Subsequently, the present disclosure plotted the model attention values obtained during training for the ligand-protein docking poses across various pockets in FIG. 11. The present disclosure observed that ligand-protein poses docked on pockets 1 and 2 consistently received high proportions of attention throughout the training process, aligning with the documented binding sites (within pockets 1 and 2) of the ligand to the protein in the RCSB PDB database.

[00172] **Quantifying the accuracy of self-attention in DTIGN:** The pockets of model attention (denoted as "Attention Pockets") are sorted based on the average attention values during model training, selecting the same number of pockets as those actually binding, from high to low. Then, the Intersection over Union (IoU) between the "Attention Pockets" and the pockets of native structures (denoted as "Native Pockets") is calculated in FIG. 12. The results

reveal that out of the 17 training ligands with native structures related to the target, 13 achieve an IoU score exceeding 0.5, indicating successful binding pocket detection. It is worth noting that the native structures associated with the training ligands were not utilized during unsupervised training; the training ligands were only trained on multiple docked pocket-ligand complexes and their bioactivities.

[00173] **Generalization of DTIGN's model attention on unseen test data:** The model also demonstrates its ability to identify correct binding pockets for unseen test ligands. It achieves an IoU score above 0.5 in 6 out of 9 cases where native structures are available (as shown in FIG. 12). This further underscores the effectiveness of our self-attention mechanism in identifying binding pockets. Moreover, both the training and test ligands consistently maintain a mean IoU score above 0.5 between "Attention Pockets" and "Native Pockets", indicating the generalization ability of the present approach in pocket identification.

[00174] **Generalization of DTIGN's model attention on PDBBind re-docked data**. To validate the model's generalization ability on the PDBBind dataset, the present disclosure utilized QuickVina 2.1 to re-dock proteins and ligands from the PDBBind v2020 dataset (http://www.pdbbind.org.cn/download.php), generating 10 ligand-protein docking poses for each PDB code (one code represents one unique native structure) if feasible. These were generated using the box center set to the average coordinates of the native ligand, and docked with a box size of (15Å). The generated 179,878 docking poses involve 18,197 PDB codes and three types of bioactivity assays (pIC50, pKd, and pKi). For each type of bioactivity assay, the present disclosure grouped identical scaffolds within the same subset (i.e., the training, validation, and test sets) based on Bemis-Murcko scaffolds whenever possible to evaluate the generalization ability of DTIGN. The three types of bioactivity assays (pIC50, pKd, and pKi) used for training DTIGN contain 56,725, 53,289, and 39,282 poses re-docked on 5,756, 5,390, and 3,959 PDB codes, respectively. Those three types of bioactivity assays used for testing DTIGN contain 9,975, 14,411, and 6,196 poses re-docked on 1,011, 1,455, and 626 PDB codes, respectively.

[00175] Following that, DTIGN was trained and tested using these re-docked poses along with each type of bioactivity assay from the PDBBind v2020 dataset. To provide a more realistic assessment of DTIGN's attention generalization ability, the present disclosure refrained from using native structures during training to refine the model's attention, considering the scarcity of bioactivity-assayed native structures in practice. Subsequently, the present disclosure analyzed the attention distribution of the re-docked ligand-protein poses

across various RMSD rankings. As illustrated in FIG. 13, external test poses with high RMSD rankings, which indicate native-likeness, generally received more attention from DTIGN. These test samples' native structures were not involved in DTIGN's training. From the additional experiments shown in FIGS. 14 and 15, the present disclosure observed consistent results across different bioactivity assays, demonstrating a robust generalization ability of DTIGN's attention. Simultaneously, the present disclosure noticed that docking poses with an average RMSD greater than 9Å diverted some attention from the DTIGN model. Therefore, the present disclosure does not recommend using docking software that generates docking poses with an average RMSD greater than 9Å for DTIGN learning.

D. Benchmark Tests

[00176] FIG. 16 shows a table of the best cross-validated Pearson's *r* 1600 on the benchmark datasets. FIG. 17 shows a table of the best cross-validated RMSE 1700 on the benchmark datasets. FIG. 18 shows a table of the best cross-validated τ_B 1800 on the benchmark datasets.

[00177] In this section, the present disclosure validates the general effectiveness of their method on all tasks in the benchmark dataset (as detailed in FIG. 9). Following the experimental setup outlined in FIG. 9, the testing performances on "DTIGN Unsupervised" are presented in FIGS. 16 to 18. The benchmark test results demonstrate that DTIGN outperforms the compared methods on the majority of the datasets and achieves the best overall performance across all metrics. On average of all datasets, there is a 32.12% increase in r, a 5.52% decrease in RMSE, and a 43.44% increase in τ compared with the best-performed baseline methods. Overall, the average improvement is calculated to be 27.03% by averaging the improvements across all metrics mentioned above.

E. Review of Docking Poses Based on Attention Allocation

[00178] FIG. 19 shows the binding poses 1900 between exemplary target protein CHEMBL202 and the test ligand CHEMBL7492 in the: (A) native pose and (B) AutoDock Vina-docked structure in the pocket the present model pays attention to (pocket 2). Non-covalent bonds (polar interactions) between ligand and protein are marked by dashed lines, with binding sites highlighted on the protein sequences above. Common binding sites across native and docking structures are encircled in dashed lines. FIG. 20 shows (A) "CHEMBL7492" in "1u72"; and (B) "CHEMBL22" in "2w3a" indicating the test ligand ID, and its PDB ID with the native structure of the protein, indicative that the DTIGN identifies

some native-like docking poses to learn from, which show a RMSD of less than 5Å from the native structure. The correlation between RMSD and average attention value during model training was evaluated using the Pearson correlation coefficient (*r*).

[00179] **The docking poses that the model tends to focus on have some similarities to their native poses:** The precision of self-attention was validated in Section II-C above. Here, the present disclosure retraces molecular docking results using the pocket of model attention. Taking CHEMBL202 and native-structured CHEMBL7492 as examples, the present disclosure compares docking results used for training. AutoDock Vina-docked poses in attention-concentrated pockets 1 and 2 closely resemble each other (one of them is analyzed in FIG. 19B). These poses are positioned between the two pockets and have four out of the five binding sites between the ligand and the protein in common with their native structure (as shown in FIG. 19A), making their overall conformation similar. Focusing on these accurate docking poses can improve DTIGN's bioactivity predictions.

[00180] **DTIGN identifies some native-like docking poses to learn from**: By calculating the RMSD between the native pose and the docking poses within each pocket, the investigated the docking accuracy in the pockets of model attention. To eliminate coordinate biases, the present disclosure aligned docking results with native ligand-protein complexes before RMSD calculation. Results show that the highly attended docking poses rank among the top RMSD results (as shown in FIG. 20), contributing to extracting more useful information regarding ligand-protein interactions.

III. Summary

[00181] This present disclosure pushes the boundaries of bioactivity prediction through deep learning, leveraging molecular docking to create ligand-protein complex data and extract valuable insights from molecular interactions. The present disclosure introduces the DTIGN, which incorporates intermolecular interactions, self-attention, and semi-supervised learning to enhance the accuracy of bioactivity prediction. DTIGN captures force field information by defining intermolecular interactions with high-order negative powers of distance in Coulomb and London forces. Its self-attention mechanism identifies some native-like docking poses for DTIGN to learn. Additionally, DTIGN is capable of enhancing its predictive performance by limited native structure data through semi-supervised learning. Overall, DTIGN demonstrates its superior performance compared to state-of-the-art bioactivity prediction methods, as shown in the present disclosure's benchmark tests.

[00182] While the disclosure has been particularly shown and described with reference to specific embodiments, it should be understood by those skilled in the art that various changes in form and detail may be made therein without departing from the spirit and scope of the disclosure as defined by the appended claims. The scope of the disclosure is thus indicated by the appended claims and all changes which come within the meaning and range of equivalency of the claims are therefore intended to be embraced.